SURVEY METHODOLOGY

25 1975-2000

Catalogue No. 12-001-XPB

A JOURNAL PUBLISHED BY STATISTICS CANADA

JUNE 2000

•

VOLUME 26

•

NUMBER 1

Statistics Canada Statistique Canada

Canada

SURVEY METHODOLOGY

A JOURNAL PUBLISHED BY STATISTICS CANADA

June 2000 • VOLUME 26 • NUMBER 1

Published by authority of the Minister
responsible for Statistics Canada

July 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 CDN $47 per year. The price excludes Canadian sales taxes. Additional shipping charges apply for delivery outside Canada: United States, CDN $12 ($6 × 2 issues); Other Countries, CDN $20 ($10 × 2 issues). Subscription order should be sent to Statistics Canada, Dissemination Division, Circulation Management, 120 Parkdale Avenue, Ottawa, Ontario, Canada K1A 0T6 or by dialling 1 800 700-1033, by fax 1 800 889-9734 or by E-mail: 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, the Statistical Society of Canada and l'Association des statisticiennes et staticiens du Quebéc.

SURVEY METHODOLOGY

A Journal Published by Statistics Canada

Volume 26, Number 1, June 2000

CONTENTS

Survey Methodology, June 2000
Vol. 26, No. 1, pp. 1-2
Statistics Canada

In This Issue

This issue of *Survey Methodology* continues the celebration of 25 successful years that was marked by the publication of the December 1999 issue. The first seven papers of this issue, by prominent statisticians working in survey methods, were invited to help mark this occasion but were not included in the December issue due to space limitations. I would like to extend a special word of thanks to all of the authors who helped to make these two celebration issues so special and memorable.

To start off this special issue Kalton reviews developments in survey research over the past 25 years since *Survey Methodology* first started publishing. He first describes developments in survey taking as a profession, specifically the rise of specialist journals and professional associations for survey methodologists, as well as the international and multidisciplinary aspects of the profession. He then reviews developments in survey methods including questionnaire design, data collection, non-sampling errors, sampling methods and estimation. Finally he discusses the rise in importance of panel surveys and international surveys, administrative data sources and analysis of survey data.

Bellhouse traces the parallel developments in survey taking and computing over the twentieth century. He first describes the interaction between census taking and the early development of computing machines and digital computers. Later, developments in scientific computation lead to the use of more sophisticated statistical methods and models. He concludes his story with discussions of the development of statistical software for surveys and of model-related methods.

Bailar discusses the role of statistics in census taking, with particular emphasis on errors in census counts due to census errors of various sorts and adjustment of census counts using sample based estimates of net undercount. The various sources of errors in censuses are described. Use of statistical methods for census evaluation, quality control in census processing, and imputation is also discussed. Using a model for census bias and variance, the potential efficacy of census adjustment procedures is illustrated.

Isaki, Tsay and Fuller consider estimation of census adjustment factors using data from the 1990 post enumeration survey. Their estimators are based on a components of variance model with a fixed linear predictor and a random effect describing the unknown true adjustment factor for each of 336 post-strata. They consider alternatives based on using an estimate of the full variance-covariance matrix of the direct survey errors of the post-stratum adjustment factors versus using only the diagonal elements. Use of the diagonal elements only can reduce the effects of instability in the estimate of the full variance-covariance matrix. In an empirical comparison they find that a compromise between these two extremes works best. They also restrict the model based adjustment factors so that the estimate of total population matches that obtained from the direct survey estimates of these adjustment factors.

Lachapelle and Kerr present an innovative use of a coverage study to examine the demographic estimates of population. Their approach decomposes the results from Statistics Canada's Reverse Record Check (RRC) to provide an additional source of data that can be compared to the more traditional administrative record based estimates of the components of growth. The objective of this comparison is to identify major sources of error in either the administrative record based or the RRC estimates. They also show how the error of closure can be decomposed into two parts: differences between the RRC and the Census estimates of enumerated population and differences between the RRC and administrative record based estimates of growth.

In their paper Feder, Nathan and Pfeffermann consider repeated sampling from a hierarchical population. At each fixed time point the population can be described by a two level model; first and second level random effects are then allowed to evolve stochastically over time. In particular, the case where second level units remain in the sample for only a few occasions, as for example in many labour force surveys, is considered. A two step estimation procedure is proposed. In the first step the two-level model is fit to each time point independently to obtain estimates of the fixed effects. Time series parameters are estimated in the second step. Sampling weights can be incorporated into both steps to account for possibly informative sampling.

Rivest and Belmonte propose measurement of the mean square error of small area estimators conditionally on the realized smoothing model. They propose a natural estimator for this MSE; however, the estimator can be quite unstable when there is a lot of smoothing. They also propose a correction for bias in the case that the distributions of the direct estimators are skewed. Finally

they investigate the properties of their estimator in an Empirical Bayesian context and illustrate their method using undercoverage data from the 1991 Canadian Census.

Shao addresses an important topic - the evaluation of cold deck imputation methods. Since computer technology continues to make it easier to store and access data from previous and related surveys, imputation methods that make use of this auxiliary data will become increasingly important. As a result, Shao takes the first steps in evaluating how various cold deck imputation methods will perform relative to other imputation methods.

Thompson and Frank discuss model based estimation for link-tracing designs. In link-tracing designs, links are followed from one respondent to another. Network sampling and snowball sampling are just two examples. After a general introduction to the area, they present several link-tracing designs. They then present a graphical model for the linked population. Finally they develop likelihood based inference procedures for such populations using data from link-tracing designs.

Théberge attempts to solve the problem of extreme weights due to the calibration estimator by relaxing somewhat the calibration equation requirements. In fact, the problem is one of minimization similar to that encountered in ridge regression. He also reviews other means of restricting weights. He discusses the asymptotic properties of calibrated weights, and provides necessary and sufficient conditions for the existence of restricted weights satisfying the calibration equation. He also outlines a way of formulating the estimation problem by controlling the significance given to the calibration equation, and describes various means of restricting weights that do not rely on the use of a specific distance. Finally, he suggests an estimator having restricted weights that is useful for small domains, and deals with outliers by developing a method similar to that used to handle extreme weights.

Two short notes conclude this issue. Losinger, Garber, Wagner and Hill present a case study in the care that must be taken when adjusting for non response in different waves of a survey. Finally, Shaffer looks at the estimation of regression coefficients using survey data when the assumption of fixed auxiliary variables is relaxed.

You may recall that the December issue of *Survey Methodology* was made available, on an experimental basis, in an electronic format on the Statistics Canada web site. There was also a web based survey to guage your reactions and preferences with respect to an electronic version of the journal. Although there was quite a bit of interest in an electronic version, it seems that the time is not yet ripe for publishing electronically on a regular basis. We will certainly be reconsidering this option in the near future, and your responses to the survey will help to improve any future electronic version. In the meantime, we will continue to publish a print version of the journal for the foreseeable future.

M.P. Singh

Survey Methodology, June 2000
Vol. 26, No. 1, pp. 3-10
Statistics Canada

Developments in Survey Research in the Past 25 Years

GRAHAM KALTON[1]

ABSTRACT

In recognition of *Survey Methodology*'s silver anniversary, this paper reviews the major advances in survey research that have taken place in the past 25 years. It provides a general overview of developments in: the survey research profession; survey methodology – questionnaire design, data collection methods, handling missing data, survey sampling, and total survey error; and survey applications – panel surveys, international surveys, and secondary analysis. It also attempts to forecast some future developments in these areas.

KEY WORDS: Survey research profession; Survey methodology; Survey applications; Questionnaire design; International surveys.

1. INTRODUCTION

Survey Methodology is celebrating its silver anniversary this year. In recognition of this milestone, this paper aims to review the major developments in survey research over the past 25 years. I should note, however, that for several reasons I shall be somewhat lax in my dating of events. First, there was, of course, no watershed in survey research in 1975. Rather, many of the major developments over the past quarter century built on the foundations laid by earlier work. Second, it takes time for many advances in methodology to be fully accepted and adopted. Third, I am using as my benchmark a text on survey methodology that Sir Claus Moser and I published in the United Kingdom in 1971 (the second edition of *Survey Methods in Social Investigation*, hereafter referred to as *Survey Methods*), so that my time frame actually extends over 30 years or so.

The paper reviews the developments in survey methodology, including questionnaire design, survey sampling, data collection methods, data processing, and survey analysis. Computers will feature prominently in the discussion since they have had a major impact on many, but not all, methodological developments. The paper also reviews the effects of these methodological developments on the practice of survey research, including the growth in panel surveys, international surveys, and secondary analysis. The main emphasis is on population surveys, but some references are also made to establishment surveys. Also, in reflecting my experience, the paper will no doubt have a slant toward work done in the United States. Before turning to developments in survey methods and practice, I will first describe the great expansion that has taken place in the number of surveys being conducted and the emergence of a clearly identified survey research profession.

2. THE SURVEY RESEARCH PROFESSION

Most of the history of survey research is contained in the twentieth century. The field began to take off in the 1930's, grew considerably during the Second World War, and has grown at a substantial rate ever since. By 1975, surveys of both households and establishments were well established as the means to meet the needs for statistical data of policymakers and researchers on a wide range of subjects, such as manufacturing and trade, agriculture, employment and unemployment, family expenditure, nutrition, health, education, travel, aging, and crime. In addition, surveys conducted by academic and other researchers in sociology, economics, political science, psychology, education, social work and public health, public opinion and election polls, and market research have flourished. The field has continued to expand at a rapid rate in the past 25 years, particularly as more policymakers have learned to appreciate the value of survey data and as advances in survey methods have enhanced the ability of survey researchers to respond to the demands for statistical data. The continuing demand of policymakers for more and more sophisticated data has prompted advances in survey methodology and has also led to the solidification of a broadly based survey research profession.

The rapid growth in survey research has come about in part because of an expansion in the range of topics that are considered suitable for study using survey methods. Adventuresome researchers have constantly and successfully challenged the conventional wisdom of their times about the subject matters that surveys could cover. These challenges have continued during the past 25 years so that there are now very few subjects that are ruled out for study in surveys based on valid probability samples. Some of the new subjects of study are sensitive ones, such as sexual behavior and illicit drug use, for which the application of survey methods has required the development of special data collection techniques. Other new subjects have required the incorporation of additional data collection methods, such as medical examinations for sampled individuals, videotaping of teacher-student interactions in classrooms, and placing environmental monitoring equipment in sampled households. Tackling more difficult subject matters has been a constant stimulus to methodological research.

[1] Graham Kalton, Westat, 1650 Research Boulevard, Rockville, Maryland, USA 20850. E-mail: KaltonG1@westat.com.

Prior to 1975 there were no widely distributed specialist journals in survey methodology. Refereed papers on survey methodology were published in a variety of journals. Statistical journals published, and continue to publish, papers mainly on the more statistical aspects of survey research, particularly survey sampling. Journals like *Public Opinion Quarterly* published, and continue to publish, papers on survey methodology. Market research journals publish papers on survey methodology relevant to market research. Journals in various subject-matter disciplines in the social sciences, public health, *etc.*, sometimes publish papers on survey methods relevant to their disciplines. This situation was not ideal since there was no natural outlet for some good papers on survey research methods and because the literature was widely scattered. The introduction of *Survey Methodology* in 1975 and the *Journal of Official Statistics* in 1985, both now well-established journals, has remedied this situation.

Another notable development has been the establishment of professional associations for survey methodologists. For example, the International Association of Survey Statisticians (IASS) was founded in 1975 as a section of the International Statistical Institute; the Section on Survey Research Methods of the American Statistical Association was established in 1978, after being a subsection of the Social Statistics Section from 1974 to 1977; and the Social Statistics Section of the Royal Statistical Society was formed in 1976, initially as the Social Statistics and Survey Methodology Study Group.

In recent years, several of these associations, sometimes together with other associations (particularly the American Association for Public Opinion Research), have collaborated to run international conferences on specific topics in survey methodology. A special feature of these conferences is that many of them have been structured to cover their chosen topics in a comprehensive manner so that they could generate well-rounded texts. This feature was introduced to address the shortage of literature on survey methodology that resulted from the fact that survey methodologists are practitioners with little time to publish. The result has been the production of edited volumes on such topics as panel surveys, telephone surveys, business surveys, measurement errors in surveys, survey quality, and computer-assisted survey information collection.

Many other conferences on survey methodology have also been held in recent years. Some have been organized by government agencies, such as Statistics Canada, the U.S. Census Bureau and the U.S. Federal Committee on Statistical Methodology (also founded in 1975). Others have been organized by professional associations, such as the IASS and the Association for Survey Computing. The proceedings from these conferences, and those of the Section on Survey Research Methods of the American Statistical Association, contribute greatly to the growth in the literature on survey methodology.

Two other aspects of the development of the survey profession deserve comment. One is its internationalism. The international conferences described above have led to publications with authors from many different countries. Although there are cultural differences between countries that need to be taken into account in data collection, research on survey methodology shares a good deal in common across countries. In addition, international surveys are becoming more prevalent, with the need to standardize procedures across countries (see the discussion below). In general, international cooperation in survey research is progressing well, but there is one area where much more could be done. Like the developed countries, the developing and transition countries need statistical data from surveys. However, they often lack the necessary expertise. The IASS, international agencies like the U.N. Statistical Office, a number of government statistical agencies, and a number of other bodies make valuable contributions to training survey researchers from developing and transition countries, but the level of support currently available for this training falls far short of what is needed.

Another noteworthy aspect of the development of the survey research profession is its multidisciplinary nature. As survey research has become established as a profession, it has developed a number of subdisciplines. Thirty or so years ago, a survey methodologist might expect to cover all aspects of the subject, but that is no longer possible at the highest technical level. The statistical level of the techniques of survey sampling and survey analysis used by survey statisticians has advanced greatly, survey methodologists are increasingly using theories and techniques from sociology, psychology, and anthropology, and computer specialists now need to use much more sophisticated methods for data capture and processing than in the past. This inevitable segmentation of survey methodology as the field progresses puts at risk a unified professional identification, particularly since the subdisciplines are each also associated with their own different fields. Given the importance of interdisciplinary collaboration in survey research, mechanisms to foster that collaboration may be needed in the future (see also section 5).

As with the developing and transition countries, the developed countries face a shortage of well-trained survey statisticians and methodologists. There is the need both to attract more people into the profession and to provide more training opportunities for them. There are a few graduate programs at universities and some faculty who specialize in the field, but the numbers are inadequate given the needs. The multidisciplinary collaboration involved in constructing and conducting a survey implies that the training should have a multidisciplinary component, so that the various specialists can communicate effectively with one another. Moreover, the instructors should include persons with practical survey experience. These specifications make it even more difficult for a graduate program in survey methodology to be mounted in most universities. An

alternative approach is that adopted by the Joint Program in Survey Methodology (JPSM) at the University of Maryland, a program set up with U.S. government funds to address the shortage of trained survey researchers in the federal government. The JPSM is built on a collaboration of two universities (the University of Maryland and the University of Michigan) and a private survey research organization (Westat), with important contributions from experts in survey methodology in the government, other organizations, and other universities to support its various graduate programs. In a related approach, the Department of Social Statistics at the University of Southampton and the U.K. Office for National Statistics have recently jointly developed a master's degree program in official statistics, with significant teaching contributions in both survey methodology and other aspects of official statistics being made by government statisticians. The Department is also collaborating with an independent survey research organization (the National Centre for Social Research) in the Centre for Applied Social Surveys, one activity of which is to run short courses in survey methodology.

3. DEVELOPMENTS IN SURVEY METHODS

The computer revolution that began to have a significant impact on survey analysis in the 1960's has been the dominating force behind the advancement of survey methodology over the past 25 to 30 years. The ability to process and analyze survey data much more readily than in the past has supported the use of more advanced statistical methods. It has also contributed greatly to more sophisticated demands from survey data users, stimulating the development of improved methodology for all aspects of the survey process.

The chapter on processing survey data in *Survey Methods* contains a description of punch cards that were widely used 30 years ago for the analysis of survey data, together with a description of unit record equipment (counter-sorters and tabulators) and computers. At that time computers were well on the way to replacing unit record equipment, but they were not routinely available to survey researchers. The computers of the day were large mainframe machines and punch cards were the usual input medium for survey data. Programs for survey analysis were limited in number and in scope. Today, the situation is, of course, totally different, and the impact of this change on survey research is hard to overstate.

It is against this backdrop of the computing explosion that the advances in other aspects of survey methodology should be assessed. The rest of this section briefly outlines what I view to be the significant advances that have been made in the past quarter century in the areas of questionnaire design, data collection, missing data, survey sampling, and total survey error.

Questionnaire design. The critical role of questionnaire design in achieving high-quality survey data has been well recognized from the early days. While some first-rate research was being conducted on improving questionnaire design in the 1960's and 1970's, the number of researchers involved in tackling this extremely challenging area was very limited. This situation has improved subsequently in large part due to what has become known as the Cognitive Aspects of Survey Methodology (CASM) movement. The CASM movement aims to attract researchers from the cognitive and social sciences to address the difficult problems of formulating survey questions that produce appropriate responses. The attention generated by this movement has created renewed interest in this field.

The CASM movement has not identified ready-made solutions to the problems of response errors in surveys. It would have been unrealistic to expect that all that was needed was the importation of existing theories from cognitive psychology and other disciplines into questionnaire design. What the movement has achieved is greater efforts to tackle the subject from a theoretical perspective. Also, the CASM movement has contributed greatly to more rigorous pretesting of survey questionnaires. Some of the pretesting techniques that have been developed in the past 25 years occurred independently of the CASM movement, but the sustained attention that pretesting now receives owes a great deal to that movement. A direct effect of the CASM movement has been the creation of the so-called "cognitive laboratories" that are now widely used for pretesting questionnaires, using such techniques as "think alouds" and extensive probing. Focus groups – which have a long history in questionnaire design, particularly in market research – are also much more widely used than in the past. In addition, behavior coding is now used widely in pretesting.

An associated development in the past few years has been a more theoretical approach to the design of forms that are to be completed by survey respondents. This research takes account of theories that indicate how individuals approach documents and how they most naturally work their way through them. This important subject received little attention for many years. The current research holds considerable promise for making survey forms much more user friendly, with the hope that this may improve both the quality of the data collected and response rates.

Data collection. *Survey Methods* contains two main chapters on data collection methods, one on mail questionnaires and one on face-to-face interviewing (there is also a chapter on documents and observation). There are only a few minor references to telephone interviewing, in part because of the low level of telephone penetration in the United Kingdom at that time. However, even in the United States where telephone penetration was much higher, back in 1975 many survey researchers had serious doubts about the collection of data for household surveys by telephone, at least for government surveys with major policy implications. That situation has changed considerably. Today, many U.S. government surveys are conducted by telephone.

One concern about telephone surveys is the noncoverage of households without telephones. With telephone coverage in the U.S. currently around 95 percent, the noncoverage of nontelephone households may be considered acceptable for surveys of the general population. However, a sizable number of surveys focus on subpopulations with lower telephone coverage rates, such as the poor; for such surveys telephone noncoverage is a serious concern. Another concern is nonresponse. Nonresponse rates for telephone surveys are appreciably higher than for comparable face-to-face interview surveys, and the gap appears to be widening. In making a choice between telephone and face-to-face modes of data collection, the large cost savings that accrue from the use of telephone interviewing often override the higher response rates achievable with face-to-face interviewing. Nevertheless, the risk of appreciable bias that is associated with high levels of nonresponse in telephone surveys (frequently as high as 40 percent or more, even with determined follow-up efforts) is a serious and often underrated concern. The likelihood of increasing nonresponse rates to telephone surveys raises questions about the role of telephone data collection in the future.

An important advance in data collection methods in recent years has been the introduction of computer-assisted methods, such as computer-assisted personal interviewing (CAPI) and computer-assisted telephone interviewing (CATI). These methods facilitate more complex skip patterns, prevent interviewers from deviating from the specified question sequence, provide for easy insertion of responses from earlier questions (*e.g.*, if a son's name is recorded as "Peter" in answer to one question, "Peter" can be inserted in the wording of a subsequent question), and enable edit checks to be carried out as the interview progresses and corrections made as necessary. By entering the data directly into a computer file, they also permit more timely processing. The development of general purpose programs for CAPI or CATI data collections, including sampling and scheduling, is a complex operation. Several programs are now available for this purpose. Future developments should see more flexible programs and authoring systems that are simpler to apply.

In the past few years, another form of computer-assisted survey information collection has emerged. This is computer-assisted self-interviewing (CASI), of which there are several variants: video-CASI, in which the respondent reads the questions on the computer screen and enters the answers on the keyboard; audio-CASI, in which the respondent listens to questions on headphones connected to a laptop computer and enters the answers on a keyboard; and telephone audio-CASI in which the audio-CASI interview is conducted by telephone, either with the respondent calling into the computer or with the respondent being transferred to the computer interview once the call has been established by a telephone interviewer. All these versions of CASI avoid the respondent-interviewer interactions that apply with other interviewing methods, and may therefore be particularly useful for collecting data on sensitive issues. They can also be developed in different languages if necessary. The audio variants avoid the requirement that the respondent is literate. These methods have appeared only recently and their use may be expected to expand appreciably in the future.

Some business surveys are now conducted using audio-CASI methods. An advantage to respondents is that they can call in to a toll-free number at a time convenient to them. They then listen to voice-digitized survey questions and enter responses on the keypad of a touchtone telephone. A variant of this methodology is for the respondents to answer verbally, with the responses interpreted using voice recognition techniques. The use of this methodology may increase in the future as voice recognition methods improve.

Another recent development has been the collection of survey data over the Internet. This methodology is particularly attractive for some types of establishment surveys and for surveys of populations of individuals who have access to the Internet and experience in using it. One approach is to send the questionnaire by email, which may be suitable for individuals who have known email addresses (*e.g.*, the employees of a firm with its own network). Another approach is to post the questionnaire at a web site, with respondents using a password to gain access to it. At this time, the Internet is not appropriate for use in surveys of the general population because of the high proportion of persons without ready access to it, the lack of a sampling frame, and likely low response rates. The temptation to collect a large sample of Internet responses to a survey questionnaire in an uncontrolled fashion should be avoided. This approach would simply replicate the errors made with the infamous 1936 Literary Digest Poll.

Missing data. Missing data occur in surveys through total nonresponse, item nonresponse, and noncoverage. During the past 25 years and even earlier, there has been increasing concern that total nonresponse rates have been rising. This trend is hard to document and indeed analyses of trend data from different surveys have led to different conclusions about the existence of a trend. Yet there is common agreement among survey practitioners that it has become more difficult over time to obtain cooperation. Various reasons have been suggested, such as less novelty in participating in a survey, more working people with less leisure time, fear of crime in face-to-face surveys, and the negative effects of telemarketing in telephone surveys, but there are no definitive explanations. Whatever the reasons, greater efforts now need to be made to achieve a high response rate than was the case in earlier times. These efforts include increased numbers of calls to contact respondents, greater efforts in refusal conversion, and the greater use of incentives. In the past decade, a sizeable number of experimental studies have been conducted in face-to-face and telephone interview surveys to test the effects on response rates of

various monetary and nonmonetary incentives and the level of monetary incentives, thus replicating in an interview setting the kinds of studies that were conducted with mail questionnaires in earlier decades.

Noncoverage is a recognized concern in telephone surveys, but it has received less attention in face-to-face interview surveys, and certainly less attention than the problem of nonresponse. Yet the level of noncoverage in face-to-face interview surveys among certain segments of the population (*e.g.*, young black males in the United States) can be high. Moreover, little is known about those not covered, except that they can be expected to be different in many ways from those covered. It is a source of survey error that would benefit from greater attention in the future. Noncoverage is often especially severe when a survey of a rare population (*e.g.*, teenagers) is conducted with sample members being identified through a large-scale screening survey. Given the increasing interest in surveying rare populations, this type of noncoverage warrants particular attention.

Twenty-five years ago, item nonresponse was generally handled by simply dropping the cases from the analysis in question, for example computing percentage distributions for the subset of cases with acceptable responses. In essence, the implicit assumption being made was that the item nonresponses were missing completely at random (MCAR). Although that practice is still applied in many surveys, increasingly some form of imputation is being used to assign values for the missing responses in a manner that takes account of responses to other survey questions. This process replaces the often untenable MCAR assumption by a missing at random (MAR) assumption, that is that the item nonresponses are missing at random conditional on the auxiliary variables used in the imputation. Although imputation methods were occasionally used 25 years ago, most of the substantial literature on the subject has appeared since 1975. Current methods rely heavily on the computer power that is now available. Imputation remains an area of active research with two main foci: the development of imputation methods that maintain the covariance structure of the survey data set, taking into account that nearly all of the survey variables may be subject to item nonresponse; and the computation of variance estimates for survey estimates that are based on data some of which are imputed (see the discussion below).

Data editing is closely related to imputation. It has also experienced significant advances in recent years, taking advantage of increased computing power to develop more complex editing procedures than could have been employed in the past. Like imputation, editing is the subject of much current research interest and further developments can be expected.

The growth in computing power is also a major factor in the development and widespread use of weighting adjustments for nonresponse and noncoverage. Weighting class adjustments for nonresponse and noncoverage (poststratification) were applied when unit record equipment was used for survey analysis, but the methods were necessarily relatively simple. Now, more complex weighting class methods and calibration methods incorporating numerous auxiliary variables are widely used, often after exploratory analyses have been conducted to identify appropriate auxiliary variables.

Survey sampling. The main methods of sample design (*e.g.*, stratification, multistage sampling, sampling with unequal probabilities) were developed in the early years and were described in textbooks that appeared in the 1950's. The developments in the past quarter century have been refinements and extensions of these methods, for example to random digit dialing (RDD) sampling for telephone surveys. Here again, the ability of the computer to process large volumes of data in census files and other large sampling frames has enabled survey statisticians to construct more efficient sample designs than in the past.

One area of research in recent years has been on methods for sampling rare populations, either in a special survey or by oversampling in a general survey. This interest is part of the extension of survey demands to provide results for many different domains, including small domains such as racial and ethnic minorities, children in poverty, age/sex groups, and geographical subdivisions (see also the reference to small area estimation below). The aim of the research is to develop efficient sample designs and data collection methods for sampling such domains in situations where special frames for those domains are unavailable. Since the demands for domain results continue to grow, ways to survey rare populations in a cost-effective manner will continue to be sought.

In the 1970's, the design-based mode of inference that is generally adopted with sample surveys was strongly challenged by those who argued that it should be replaced by the model-dependent methods used in the rest of statistics. That debate has waned, and the design-based framework remains in place (see the further discussion below). In this context, the terminology should be clarified: from early on, the design-based mode of inference incorporated the use of models in improving the precision of survey estimates (*e.g.*, regression estimates), but the estimates remained consistent under that mode of inference irrespective of the validity of the model. Thus, the procedures are model-assisted as distinct from model-dependent. The suitability of model-dependent estimates depends on the validity of the model (or the robustness of the estimates to model failure). The computing developments of recent years have facilitated the greater use of models, and of more complex models, within the design-based model-assisted framework of inference.

These remarks should not be interpreted to imply that model-dependent methods have no place in survey research. On the contrary, the methods for handling missing data described above are necessarily model-dependent. Model-dependent methods are also used increasingly in producing estimates for small domains (generally small geographic

areas). Such methods are needed when the sample sizes in the domains are too small (they may often be zero) to produce design-based estimates of adequate precision. In this situation, small area estimates may be produced by borrowing strength from survey data for other areas or time periods through a statistical model that relates the survey data to other, generally administrative, data. The rapid growth in social programs that distribute funds to small geographic entities has led to a substantial demand for up-to-date small area estimates. As a result, small area estimation has become a major area of research activity in recent years, and is likely to remain so in the years to come.

Variance estimation for estimates from complex sample designs has been another major area of development in the past quarter century. Methods based on Taylor's series approximations and replication methods were being used in the 1960's, but they were not routinely applied and were largely confined to research studies. This situation has changed dramatically as a result of the increases in computing power and the development of a number of computer packages for the computation of sampling errors for estimates from complex (typically stratified multistage) sample designs. It is now fairly common practice to compute sampling errors routinely in analyzing survey data.

A notable development in recent years has occurred in the area of the application of analytic models to survey data. This area is one where there remains a debate about the choice between a design-based and model-dependent mode of inference. Within the design-based framework, there have been both theoretical advances in the application of regression models, categorical models, survival models, multilevel models, *etc.*, with survey data and in software for computing variances for these models. At present, survey analysts often conduct their exploratory analyses using the greater flexibility of standard statistical packages, and compute the design-based variances using survey sampling variance estimation software only at the final stages of their analyses. In the future, survey sampling variance estimation procedures should become more fully integrated into standard packages.

An area of much current research activity is the computation of variance estimates for survey estimates that are based on responses some of which are imputed. One approach is the application of multiple imputation procedures to complex sample designs, an application that makes strong use of current computing power. Other methods are being developed under the standard design-based mode of inference (necessarily with model assumptions). The future may see the incorporation of these methods into the survey sampling variance estimation programs so that they can be readily applied.

Total survey error. The preceding discussion has treated the various components of the survey process individually. A well-designed survey, however, is the blending together of the components into an effective package taking cost considerations into account. The last 25 years have seen a firmer recognition of the issue, with heightened attention to the concepts of total survey error and total survey design. With constrained resources, a survey design reflects trade-offs between, for example, sample size, the extent of non-response conversion undertaken, questionnaire length, and the quality of data obtained by different modes of data collection. In analyzing survey data, the quality of the estimates should properly be assessed in terms of the total survey error from all sources, not just sampling error. For both design and analysis, detailed information is needed on the various sources of error and their effects on the survey estimates. Moreover, since surveys are multipurpose studies, with many different analytic goals, the information requirements are extensive. The rapidly growing literature on survey errors from different sources is helpful for addressing total survey error and total survey design within cost constraints, but more studies are still needed.

The total survey error and total survey design concepts are most readily applied to repeated surveys. Information on error sources can be accumulated from one round to the next and can then be used to determine priorities for where improvements in the survey methods are most needed. One use of the quality profiles that provide integrated accounts of what is known about the error sources in a survey (see the discussion below) is to guide the choice of priorities for methodological improvements.

4. OTHER DEVELOPMENTS

This section reviews a number of areas of survey research in which important developments have occurred in the past 25 years, other than the strictly methodological areas discussed in section 3. The set is not intended to be an exhaustive one. It includes only areas that I consider to have undergone major change.

Panel surveys. The benefits of longitudinal data obtained from panel surveys have long been recognized, and panel surveys were being conducted in the 1940's and 1950's. At that time, however, the complexities of creating longitudinal data sets, combining the data collected in different waves, were severe. Panel surveys were often mostly analyzed only cross-sectionally, and this was a major source of criticism of the method. Today, the advances in computing and also in techniques for longitudinal analysis have changed the situation dramatically. Nevertheless, the complexities of longitudinal data, especially the problem of missing data, remain. Longitudinal methods of analysis are now widely used, although many panel surveys are still analyzed mostly cross-sectionally, with too little attention to the wide range of issues that their longitudinal data could illuminate.

There has been an enormous growth in panel surveys in the past 20 years, covering a wide range of subjects, including education, labor force transitions, health, and voting behavior. Panel surveys of household economics, modeled on the University of Michigan's Panel Study of

Income Dynamics that began in 1968, have become popular and are now being conducted in a sizeable number of countries. There are also panels like Statistics Canada's Survey of Labour and Income Dynamics and the U.S. Census Bureau's Survey of Income and Program Participation that use similar approaches.

It seems likely that the use of panel designs will increase even more in the future. The challenge is to make full use of the longitudinal data produced, since the analytic potential of a panel survey increases exponentially with the number of waves of data it collects. In addition, the significant advances in techniques for longitudinal analysis being made by biostatisticians and others provide the tools for more sophisticated analyses than in the past. Many skilled analysts are needed if the data collected in a panel survey are to be fully analyzed. The growth of secondary analysis (see the discussion below) holds promise for fuller use of panel survey data in the future.

International surveys. The last 25 years have seen the emergence of international surveys of various kinds, ranging from surveys promoted by international agencies to the coordination of independent country surveys to provide cross-national comparisons. A major breakthrough in this area came with the World Fertility Survey (WFS), which conducted surveys in 42 developing countries and 20 developed countries during the period 1974-1982. The WFS not only collected valuable data on fertility, but in many countries it also provided technical assistance in survey research that helped to develop an infrastructure of survey taking. The ongoing Demographic and Health Survey began shortly after the end of the WFS and to date has conducted surveys in more than 50 countries.

Education has been the subject of a number of international surveys including, for example, the Third International Mathematics and Science Study (41 countries in 1995) and its replication (40 countries in 1999); the Programme for International Student Assessment (about 30 countries in 2000); the Second Civics in Education Study (about 20 countries in 1999); the IEA Reading Literacy Study (about 30 countries in 1991). The ongoing International Adult Reading Literacy Survey is collecting comparable information about literacy levels of adults in a number of countries around the world. Two examples of other internationally organized survey designs are the Multiple Indicator Cluster Survey from UNICEF and the Social Dimensions of Adjustment Integrated Survey from the World Bank. A related activity is the coordination of surveys in the European Union by Eurostat. An example of cross-national collaboration on surveys is provided by the International Social Survey Programme, a continuing annual survey program on social science topics that now has 33 member countries.

The development of international survey programs has occurred for two separate reasons. One is the growing interest in the comparison of survey results across countries. The other is to assist countries, particularly developing and transition countries with limited survey experience, in the conduct of surveys that will provide important data for planning purposes. Considerable expansion in international survey activity can be expected in the future for both of these reasons.

Linkages to administrative data. The increases in computing power and the resultant ability to conduct more sophisticated analyses have led to a demand for more data on the sampled units. Analysts want to answer more complex questions than was the case in the past and some of the data they need may not be readily collectable in a survey, at least with the required level of quality. Even if the data were collectable, the collection could create excessive respondent burden. This situation has led to the search for alternative sources for the data, with data taken from those sources then being linked to the survey responses. Thus, for example, tax records might provide valuable earnings histories for sampled individuals over a timespan for which the respondents could not provide the data, or medical records might provide the amounts of medical expenses paid directly by insurers that are unknown to the respondents. These kinds of linkages have been made much more feasible by the significant expansion in the number of administrative record systems now available in electronic form.

There has been considerable interest in linking administrative record data to social survey data in recent years and a number of surveys have made such linkages. However, there are generally significant problems to overcome in gaining access to administrative data and serious concerns about protecting the survey respondents' privacy. These issues have severely limited the use of administrative record linkages in household surveys to date. Despite the substantial potential benefits of such linkages, it is not clear to what extent these barriers can be overcome.

In contrast, administrative data have become a key element in the conduct of economic surveys and censuses and, in a number of cases, they have replaced the data that used to be collected from respondents. The result has been a substantial decrease in respondent burden, improved data quality, more timely reporting, and reduced costs.

Secondary analysis. The increases in computing power, the increasing numbers of surveys being conducted, and the increased sophistication of the data collected in surveys have all stimulated a major growth in the secondary analysis of survey data. Public-use files are now more routinely made available, sometimes through survey data archives, to enable secondary analysts to conduct their own analyses, thus permitting survey data to be more thoroughly analyzed. Associated with this activity, increased attention has been needed to protect the survey respondents' confidentiality and to ensure that data files released to secondary analysts are not used to breach confidentiality. With secondary analysis undoubtedly continuing to expand in the future,

continued attention will need to be given to ways to release survey data in a manner that protects respondents but does not seriously curtail the range of analyses that can be conducted.

Survey quality. Increasing attention is being given to different aspects of survey quality. In the past few years, a number of survey organizations have become interested in survey process quality, applying the ideas of total quality management to survey processes. Greater attention than in the past is being given to quality taken in the broad sense to include the accuracy of the estimates produced, relevance, timeliness, accessibility and cost-efficiency and in the narrower sense of accuracy alone. Users of survey estimates and secondary analysts of survey data need to be informed about the overall quality of the survey data, including sampling errors, nonresponse and noncoverage, response errors and processing errors. While this need has long been recognized, current practice in reporting survey quality is often seriously deficient. There are signs that more attention is now being given to this area. The introduction of quality profiles that provide full and integrated reports on the quality of the data in ongoing surveys is an important contribution.

5. CONCLUDING REMARKS

This section attempts to predict some major considerations for survey research in the next 10 to 20 years. The computer revolution that has transformed the nature of survey research over the past 25 years is still in progress, and further developments can be expected in many aspects of collecting, processing, and analyzing survey data. The telecommunications industry is also in a state of rapid innovation, and the changes are likely to affect the ways that survey data are collected. It seems likely that greater use will be made in the future of mixed-mode designs, taking advantage of new modes for respondents with access to them (*e.g.*, the Internet) and using conventional modes for other respondents. Thus the effect of mode on survey responses will continue to be an important concern.

In general, it seems probable that the demand for survey data will continue to grow rapidly as more policy analysts learn to take advantage of survey data. Increasingly, survey estimates will be needed for small domains, especially small geographic domains, as policymakers target their programs to special population subgroups. Currently, most of the demand for survey data comes from central governments; in the future the demand from provincial and local governments may expand. The difficulty here is that surveys cost almost as much for small populations as for large ones. Local governments may therefore often be unable to afford the cost of a survey unless inexpensive methods can be found.

The major concern for the future of survey research is that respondents' willingness to participate in surveys may continue to decline, and that increased efforts in data collection will not fully counteract this effect. Thus, response rates will fall. This comment is of particular salience for telephone surveys, where nonresponse rates are already high. A significant increase in telephone nonresponse rates could even lead to the demise of telephone data collection for household surveys.

Finally, the next decade or so may well see the emergence of a new and different professional society for survey researchers that more broadly represents the interests of all members of the profession. Since survey sampling was at the forefront of the developments of survey research in the early years, survey research has strong ties with statistical societies. However, those ties tend to concentrate on survey statistics. There are also ties with societies for public opinion research, market research, and various subject matter disciplines, such as sociology and psychology, primarily for survey researchers who deal with the nonsampling aspects of survey research. Similarly, there are ties with computing societies for those working on survey computing. As yet, however, there is no society that aims to bring survey researchers of all disciplines together. The years to come may see the creation of such a society to promote exchanges across the different disciplines and thereby help to advance the field. Were such a society to be formed, it would not affect the need for the current ties that survey researchers have with statistical and other societies. Survey researchers need to keep in touch both with the developments taking place in survey research broadly and also with the developments in their own disciplines.

ACKNOWLEDGEMENTS

I am grateful to Joe Waksberg and Dan Levine for helpful suggestions in the preparation of this paper.

Survey Methodology, June 2000
Vol. 26, No. 1, pp. 11-20
Statistics Canada

Survey Sampling Theory Over the Twentieth Century and its Relation to Computing Technology

D.R. BELLHOUSE[1]

ABSTRACT

Computation is an integral part of statistical analysis in general and survey sampling in particular. What kinds of analyses can be carried out will depend upon what kind of computational power is available. The general development of sampling theory is traced in connection with technological developments in computation. What is possible in theory is only practicable with the proper computing technology. At the same time new developments in technology can motivate new areas of theory to investigate. One hundred years ago, it was the requirements of statisticians that spurred on technological development. Although theoretical developments in sampling theory have often run ahead of computational capabilities, it is now the case that survey statisticians are now followers of computing technology that has been motivated by others instead of acting as the catalyst that leads to technological change.

KEY WORDS: Analysis of survey data; Digital computers; Punch cards; Scientific programming; Statistical software; Survey data analysis; Survey estimation.

1. INTRODUCTION

There are several ways to approach the history of survey sampling. Two are very tempting, but will not be followed here. The first is to examine sampling in the context of the history of ideas – who formulated them and then how and why they are formulated, promoted, defended and discarded or supplanted. With respect to the personalities, it is not necessarily the one who espouses the idea first who is given prominence but the one who promotes it the best or the one who can best put the idea into practice. The approach of the history of ideas has been followed to a certain extent by Kruskal and Mosteller (1980) and Bellhouse (1988) who examined the progression of ideas beginning with the espousal of the representative method by Kaier (1897) over censuses combined with the use of randomization in surveys by Bowley (1906). The whole story of the debates over the foundations of sampling falls directly under this approach. From this debate, which was initiated by Godambe (1955), has emerged the continuing question of when to use models in sampling design and estimation. A second way to approach the history of sampling is to look at sampling theory as a branch of mathematics and then to fit this development into the general pattern of how research in mathematics evolves. Complicating this is that there are several approaches to how mathematics evolves, as discussed in Gillies (1992). One approach is to note that periodically there are results which seem to open up new areas of research while other areas become seemingly complete or "fished out" for new research ideas. Emerging areas of research often attract several talented researchers to work on these new problems and away from other potential research problems. This has its parallels in sampling. Hansen and Hurwitz (1943) obtained results on sampling with probability proportional to size and with replacement. Then Horvitz and Thompson (1952) extended this idea to sampling without replacement. The basic problem in unequal probability sampling without replacement is to find a sampling design that yields the desired inclusion probabilities. This resulted in several papers on the subject culminating in the review monograph by Brewer and Hanif (1983). Lately, very few papers are written to promote new without replacement sampling designs that result in inclusion probabilities proportional to a size variable. However, statistics and survey sampling cannot be equated to pure mathematics. Much of statistical research is motivated by practical problems in data interpretation and analysis not by abstract ideas.

In view of the explosion of technology over the 20th century, I chose another approach. This is to view the history of sampling over the 20th century as the history of the interplay between ideas that have been put into practice and computing technology that has defined the limits of practice or that has encouraged ideas for new developments in practice. The development of sampling methods may be categorized by the intersection of two strands: the use of surveys for descriptive and analytic purposes, and whether or not hypothetical models should be used.

2. BEGINNINGS: THE FIRST HALF OF THE TWENTIETH CENTURY

The first two major breakthroughs for survey sampling, one in the formulation of a statistical concept and the other in the development of technology, occurred at the end of the nineteenth century. Both breakthroughs faced some initial opposition or apathy, the idea more so than the technology,

[1] D.R. Bellhouse, Department of Statistical and Actuarial Sciences, University of Western Ontario, London, Ontario, N6A 5B7, Canada.

but both prevailed and were developed further. These breakthroughs were: (1) Kaier's (1895/6, 1897, 1905) espousal of sampling through a "representative method" over attempts at complete enumeration for social surveys, and (2) the development of punch card machines for data processing by Hollerith (1894). Both breakthroughs were directly related to survey or census work. This was the first and last time that survey or census issues inspired major technological innovation. From then on, survey sampling has adapted itself to the available technology.

Kaier's idea was to get a sample that was an approximate miniature of the population. Through sampling, more detailed information could be obtained and more specialized studies could be carried out, all at a fraction of the cost of a census. The idea initially met with opposition and it took upwards of a decade for his ideas to be accepted.

The development of machinery by Herman Hollerith for data processing came directly out of the needs of the U.S. Bureau of the Census and the encouragement of the Bureau's Director of Vital Statistics, John Shaw Billings. The events that led to this development are described by Willcox (1926):

> "While the returns of the Tenth Census [1880] were being tabulated at Washington, Billings was walking with a companion through the office in which hundreds of clerks were engaged in laboriously transferring items of information from schedules to the record sheets by the slow and heartbreaking method of hand tallying. As they were watching the clerks he said to his companion, 'There ought to be some mechanical way of doing this job, something on the principle of the Jacquard loom, perhaps, whereby holes on a card regulate the pattern to be woven.' The seed fell on good ground. His companion was a young talented engineer in the office who first convinced himself that the idea was practicable and then that Billings had no desire to claim or use it. Thereafter he devoted the bulk of his life with great ultimate profit for himself and the world to ripening the invention and securing its adoption. I have no need to describe or eulogize Hollerith machines."

A full description of the development and use of these machines for surveys is given in Mandeville (1946). Hollerith's machine was applied to processing the 1890 U.S. census. While the 1880 census took over seven years to complete, the 1890 census was finished by early 1895. The Bureau used 180 tons of cards that were processed at a speed of 6,900 cards per 6½-hour day. Not only did the machine save time, it also significantly reduced tabulation errors. The punched card machine was used to process the 1891 Census of Canada, but it did not see early use in the censuses of the United Kingdom and the rest of the British Empire. It was felt that the level of detail required in these censuses did not justify the use of a Hollerith machine since the time saved by the machine would be balanced by the time taken to punch the card (Hooker 1894). In a paper on census taking Baines (1900) expressed a preference for manual over machine tabulation, especially when labour was cheap. Despite these initial misgivings, improvements to the machine continued and the use of the Hollerith machine for statistics became highly developed by mid-century. Hartley (1946) demonstrated the most sophisticated use of these punched card machines for statistical analysis. This included the calculation of moving averages and serial correlations as well as the solution of simultaneous equations on Hollerith machines.

After these near simultaneous and unrelated innovations in ideas and technology, theory ran ahead of practice for the next 50 or 60 years. Theoretical developments in sampling continued through the first half of the century. Out of discussions over the path to follow in the "representative method," Bowley (1926) put together a monograph describing all the known theoretical results in sampling under random selection and under purposive selection. In addition, he developed the theory for stratified sampling under proportional allocation. The triumph of randomization over purposive selection was due to Neyman (1934) who showed why randomization gave a more reasonable solution to sampling problems than purposive selection. Although not the first to do so, he also developed optimal allocation strategies for stratified sampling. Prior to the middle of the century the last major development, in terms of sampling design with accompanying estimates and variance estimates, was the concept of unequal probability sampling introduced by Hansen and Hurwitz (1943).

The practical implementation of these theoretical results was limited to relatively small-scale surveys. The analyses for most surveys used calculators, either electric ones such as those manufactured by Friden, Marchant or Monroe, or hand calculators operated by turning a crank such as the Brunsviga used by Pearson and the Millionaire used by Fisher. Since the labour in the analysis increased significantly with the sample size, standard errors were seldom calculated, and when calculated the correct formulas were seldom applied. Bowley (1936) describes a typical situation showing the infrequency of standard error calculations:

> "Tabulation is usually a dull and tedious job, but there is a certain interest in watching the entries accumulating in a cross table and seeing the gradual growth of continuity out of randomness. When the results take the form of a frequency curve, and especially if we have reason to expect a normal curve and find it, we have good reason to suppose that we have measured satisfactorily a real entity. Thus the distribution of price changes or their logarithms on a normal scale gives a great deal of support to the validity of an index number. In such cases the computation of standard error is reasonable."

Box and Thomas (1944) describe a survey of approximately 4,500 respondents stratified by the industry in which they worked. The standard errors, when presented, were calculated using the formula for simple random sampling. A decade later Deming (1956) noted:

> "Although the possibility of showing a valid standard error is by definition a feature of any probability sample, it is a fact that results of probability samples have too often appeared in the past without standard errors because of the sheer labor of computation."

It is within this context that Mahalanobis (1946) suggested the technique of interpenetrating subsamples. This technique, which Mahalanobis developed at the Indian Statistical Institute in the 1930's (Murthy 1967 and Deming 1956), is very simple: two or more independent subsamples are chosen according to the same sampling design. Then the variation between the subsample estimates of the population total provides an unbiased estimate of the variance of the final estimator of the total. Computationally, the method has distinct advantages in the punch card environment where sums are easier to obtain than variances. With interpenetrating subsamples the main computational effort is in finding the subsample estimates that are based on sums only. The Indian Statistical Institute obtained its first Hollerith machine in 1944. Prior to that time, tabulations and other calculations were done by hand. The Institute's Annual Report for 1945-46 published in *Sankhyā* shows the initial unease that always greets technological change and the eventual positive benefits to change. With respect to the introduction of these machines, the report states:

> "Contrary to apprehensions among certain sections of workers that the Hollerith machine would to a large extent eliminate manual computations, it was found that new and detailed studies which could not be formerly undertaken could now be handled without difficulty so that the demand for trained computers in the later stages was on the increase. In addition to routine projects undertaken from time to time, special studies such as mechanical solution of determinants, construction of tables, fitting of orthogonal polynomials, *etc.* were conducted."

In the United States, Deming (1956), for example, picked up on the general idea and put forward methods of replicated sampling. The U.S. Bureau of the Census used this method for variance estimation. At the Bureau this idea evolved into pseudo-replication, or eventually balanced repeated replication, for variance estimation (McCarthy 1969).

3. THE ADVENT OF THE DIGITAL COMPUTER

The initial development of the digital computer was for military purposes during the Second World War (Ceruzzi 1998). For some years after the war the military continued to play a central role in the advancement of computing. By the 1950's commercial uses were developed for the computer, and this is where sampling practice begins to catch up with sampling theory. The first generation of commercial computers included the UNIVAC followed by the IBM 700 series. These computers contained thousands of vacuum tubes as internal memory. The tubes for the IBM machine were about three inches in diameter and held 1,024 bits of information. The UNIVAC ran at 2.25 MHz and could carry out 465 multiplications per second. For both machines, data were input via punched cards and stored data was on magnetic tape rather than continued use of the punched cards. The 1961 census in the United Kingdom underscores the continuing central role of the military in computing at this point in time. The census was processed on an IBM 705 computer (Benjamin 1961). The computer belonged to the War Office and was used by the Royal Army Pay Corps. The census workers were able to use the computer when not in use by the army. Information was input via cards punched in one location and then taken to the computer in another location.

Although it was not at the forefront of the development of the computer as it had been with the Hollerith equipment, the U.S. Bureau of the Census was central in the initial commercial development of the digital computer. Not only did the Bureau receive the first UNIVAC that was produced, but also some of its employees participated in design decisions for its construction (Ceruzzi 1998 and Hansen 1987). The computer was delivered in March of 1951 and was used for processing the 1950 census. It ran 24 hours a day all week until the task was completed. Once the census work was completed, the computer was used for other censuses and surveys including the Current Population Survey. Technology was now catching up to theory; the computer was now used for better calculation of variance estimates. It also opened up new possibilities, in particular imputation of missing values. With respect to variance estimates Hansen, Hurwitz, Nisselson and Sternberg (1955) comment:

> "Until the acquisition of a high-speed electronic computer, the UNIVAC, extensive approximations were introduced into the estimates of variances to avoid computations that would be exceedingly time consuming with the available equipment. The availability of the UNIVAC makes it possible to avoid

> most of these approximations. Even with the electronic computer, however, the work of making variance computations would be extremely heavy if variances were computed for all items directly. Approximate methods will continue to be used in the future, but they will be evaluated by more exact computations than have been feasible in the past."

Other statistical organizations followed but at a slower pace. The slow pace in Canada was perhaps due in part to the American experience. A 1956 report to the Dominion Statistician at the Dominion Bureau of Statistics in Canada on the subject of computing at the Bureau of the Census (reported and quoted by Worton 1998) states:

> "Subject-matter people ... are not entirely convinced that the UNIVAC system has given them the results which might be expected from a computer system. Undoubtedly UNIVAC has given a great deal of trouble – much of it probably not the fault of UNIVAC at all. Factors such as poor programming, inadequate analysis of the job, inexperienced operating staff, maintenance problems, and even friction between the three operating groups, *i.e.*, the subject matter staffs, the Central Operations Group, and the Central Electronics Unit are reflected in the performance of the UNIVAC system."

The Dominion Bureau of Statistics, now Statistics Canada, obtained its first computer in 1960, an IBM 705. The computer was used to process the 1961 census. As noted already, the British used an army-owned computer to process their 1961 census. In the late 1940's, Mahalanobis was on a list showing interest in obtaining one of the first UNIVACs (Ceruzzi 1998). However, the annual reports of the Indian Statistical Institute published in *Sankhyā* show that the Institute did not obtain a computer until 1956 at which time it received an HEC-2M.

Variance estimation for survey estimates of means, totals and proportions was now feasible for large-scale surveys. Widespread use of this technology now depended on two things – access to a computer, which was an expensive item to buy, and appropriate software to carry out the calculations.

4. SCIENTIFIC PROGRAMMING

Certain kinds of research, and the application of these research results, are possible only with computing. These possibilities expand not only with the expansion in computing power, but also with easier access to the computer's power through programming languages or packaged programs. For several years the most popular scientific programming language was FORTRAN (FORmula TRANslation). This was introduced in 1957 by IBM for its 704 computer. Part of what popularized FORTRAN was the development of the WATFOR (WATerloo FORtran) compiler at the University of Waterloo in 1965. This popular compiler, which was used for teaching purposes, combined with the dominance of IBM in the marketplace made FORTRAN accessible to many students and subsequently to researchers (Ceruzzi 1998). In reporting on the development of his own computer programs for survey research, Yates (1973) shows how pervasive FORTRAN had become over the 1960's. Yates's programs for the computer at Rothamsted Experimental Station were originally written in the late 1950's with code specific to the computer they had. In the mid-1960's the code was written in Extended Mercury Autocode. By the end of the 1960's this code had to be translated into FORTRAN using a machine translator; otherwise it was not usable at any other computer location. The earliest use of FORTRAN in sampling that I can find is in Fan, Muller and Rezucha (1962). These three individuals, all of who worked at IBM, developed algorithms and accompanying FORTRAN code to select simple random samples by computer.

There were two different paths that were followed in the application of FORTRAN programming to survey sampling. One was among statistical agencies or survey research centres and the other was among individual academic researchers. The kind of work followed along each path is strongly correlated with the evolving power of the computer and the dominance of IBM (and hence FORTRAN) in the market. By the end of the 1960's, many institutions had new and more powerful mainframe computers, often one of the IBM 360 series that was originally announced in 1964. Moreover, the software (FORTRAN in particular) remained compatible with machine changes and upgrades, especially for machines in the IBM 360 series (Ceruzzi 1998). The Dominion Bureau of Statistics obtained its first IBM 360 in 1969, while for example the Universities of Manitoba, Toronto and Waterloo obtained their first machines in the years 1966-67 (Day 1971). At the agencies and research centres, various formulae and procedures necessary to survey design and analysis were computerized. For example, Fellegi, Gray and Platek (1967) report that when the Canadian Labour Force Survey was redesigned over 1964-65, sample selection by Fellegi's (1963) method of unequal probability sampling was coded into a FORTRAN routine. From the University of Michigan Survey Research Center, Kish and Frankel (1970) report that they had FORTRAN code for obtaining variance estimates for a variety of statistics including regression coefficients using balanced repeated replication. By the mid-1960's academic researchers began to use the computer via FORTRAN programming to study, numerically or empirically, the sampling theory that they or others had derived. One of the first was Sedransk (1965) who carried out some efficiency comparisons in FORTRAN on an IBM 7074 (marketed by IBM in 1964) for a double sampling scheme. In particular, efficiency comparisons were made between optimal values

for the first and second phase sample sizes and an approximation to the optimal values. The computations involved taking expected values over a trinomial distribution in which several conditions had been imposed. The use of the computer here was to obtain a numerical comparison between exact methods and approximate ones. By the end of the decade a new kind of computer-based research process emerged. Rao and Bayless (1969) and Bayless and Rao (1970) compared several unequal probability sampling schemes by generating all possible samples and calculating the exact finite population mean square error for several real and constructed populations. It then became the norm to carry out extensive empirical studies on any newly proposed estimator or design.

The past 30 years have seen remarkable changes in computing technology. Modern computers are much faster, physically smaller and have much greater storage capacity. The steady increase in computing power and the availability of standard programming languages has allowed survey researchers to expand as well into survey data analysis. This technological change is reflected in developments in sampling theory for variance estimation. From the 1960's to the 1980's there were three basic computerized approaches to variance estimation of complex survey statistics: Taylor linearization (see Woodruff 1971, for early references to its usage), jackknife (first proposed in sampling by Durbin 1959) and balanced repeated replication (McCarthy 1969). The rise of computing power saw a new technique, Efron's (1982) bootstrap, for variance estimation. This new statistical technique, which was contemporaneous with the development of networked RISC (Reduced Instruction Set Computing) workstations running under a UNIX operating system, is highly computer intensive. Over the 1980's RISC workstations gradually replaced most mainframes in research organizations. Near the end of this transition away from mainframes, Rao and Wu (1987) extended bootstrap methodology to variance estimation for smooth statistics under stratified multistage designs.

The most recent software to have an effect on statistical research is the development of computer algebra packages. Although computer algebra has been in existence for some time, it is only in the last decade that it has progressed to the point that it is accessible to many researchers. With computer algebra many complex manipulations can be done automatically and much quicker than by hand and without risk of error. Similar to several other areas of statistics, many of the algebraic manipulations in sampling theory are related to algorithms that generate partitions. Based on the computer algorithms developed by Andrews and Stafford (1993) and Stafford and Andrews (1993), Stafford and Bellhouse (1997) have extended computer algebra techniques to survey sampling theory. Using their methodology, most of the results of so-called classical sampling theory, either existing in the literature or yet to be obtained, can be derived automatically.

5. ANALYSIS OF SURVEY DATA

While steady and substantial progress had been made in research on problems of survey estimation or enumerative surveys over the 20th century, by 1970 little had been accomplished on the analytical aspects of surveys. The terms "enumerative" and "analytical" surveys were coined by Deming in 1950 (Deming 1953). In the same article he also gives a succinct definition:

> "Briefly, the enumerative question is how many? The analytic question is why? is there a difference between two classes, and if so, how big are the differences?"

There is an implication in this quotation that the purpose of analytical surveys was for comparisons of domain means. Certainly, throughout the 1960's the understanding of what constituted an analytical survey was often limited to this. Cochran (1963) states:

> "In an analytical survey, comparisons are made between different subgroups of a population, in order to discover whether differences exist among them that may enable us to form or to verify hypotheses about the forces at work in the population."

Yates (1960) also focused mainly on domain comparisons in his discussion of analytic surveys. He did, however, discuss regression analysis and the problem of attenuation, but not the problem of general survey weights. Skinner, Holt and Smith (1989) attribute the pioneering work in analytical surveys to social scientists, Paul Lazarsfeld in particular. I will use the theoretical development of regression analysis in complex surveys to illustrate these connections to social science, in this case economics.

One of the earliest studies to take into account the survey weights in regression analysis was by Klein and Morgan (1951). At the time both were at the University of Michigan; Morgan was in the Survey Research Center. At the outset of their paper they state:

> "The sample design, the methods of collecting the data, and underlying economic behavior will all contribute to the formulation of the model. The study of data collected in consumer surveys has convinced us that one cannot proceed simply by the application of conventional statistical methods in the estimation of economic relationships because of the existence of some basic difficulties which we classify as follows: (1) weighting of observations, (2) heteroscedasticity, (3) nonlinearities, (4) the choice of alternative economic concepts, (5) errors of observation."

They addressed the first four "basic difficulties" but not the fifth. In their analysis of the approximately 2,300 responses

to the Survey of Consumer Finances, which was a multi-stage sample, Klein and Morgan used the survey weights through weighted least squares estimation of the regression parameters but ignored the clustering effect when it came to variance estimation. They noted that in many cases the use of the survey weights had little effect on the estimates of the regression coefficient estimates but noted that there was a reduction in the estimated variance for the model error. Though Klein went elsewhere, Morgan remained at the Michigan Survey Research Center. Twenty years later, he and another (Lansing and Morgan 1971) gave an overview of the state of the art for the analysis of economic survey data. Not much had changed in terms of the incorporation of the survey design into the analysis. The same is true for other areas of social research; in many cases not even the survey weights were used. In the economics literature debate continued for at least twenty years over whether to use the survey weights in regression analysis; Porter (1973) has several references to this debate.

It was out of this milieu that Kish, who also worked at the Michigan Survey Research Center, initially put forward the concept of the design effect (Kish 1957), which is the measure of increase or decrease in variance over simple random sampling experienced in a survey with a design other than simple random sampling. Design effects have become central to many aspects of the analysis of complex survey data. With respect to regression analysis, Kish and Frankel (1970) studied the design effects in the estimation of regression coefficients. They used balanced repeated replication to obtain their variance estimates. It is not entirely clear in their presentation exactly what regression coefficients they were estimating. Later, the parameters were explicitly spelled out in Kish and Frankel (1974). Specifically, the finite population parameters are what would be obtained in least squares estimation of superpopulation regression parameters when the entire finite population is available. Estimation of these parameters has become one of the standard approaches to regression analysis from complex surveys. Fuller (1975), using Taylor approximations to the variances, put the whole inference process on a solid theoretical foundation by providing limit theorems for the estimates. In addition, he addressed the one problem that Klein and Morgan (1951) ignored: errors in the variables or measurement errors in the independent variables.

Konijn (1962) took a different approach to regression analysis. Under a cluster sampling design, he assumed different simple linear regression models within each cluster. The parameters of interest were weighted averages of regression parameters with the weights given by the cluster sizes. This approach is model-based in the sense that it is the model parameters that are of interest, not a finite population parameter. Konijn's approach was not followed for several years. However, there is now a substantial literature that has grown out of this model-based approach; Pfeffermann (1993) contains several references.

With regard to the social science origins of survey analysis, there were similar experiences in categorical data analysis. The sociological literature from the 1960's and on contains many examples of categorical data analysis ignoring the sampling design. After Rao and Scott (1981, 1984) developed contingency table and goodness of fit analyses for complex surveys, Rao and Thomas (1988) tried to promote this methodology among sociologists using a review article. A search through citation indexes shows that, although this work has had great impact in the statistical and medical literature, it has had little impact in the sociological literature. The reason for this may be due, in part, to lack of computer software. The most popular software among sociologists, which is SPSS, does not at the moment contain any routines for the analysis of complex survey data. This points to a wider problem: regression, categorical data analysis and other techniques that have been proposed for complex surveys are not widely practicable without the appropriate computer software. Fuller himself tried to respond to this need by developing a packaged program for survey data analysis (Hidiroglou, Fuller and Hickman 1980).

6. STATISTICAL SOFTWARE FOR SURVEY RESEARCH

Frank Yates at Rothamsted Experimental Station was the first statistician to develop software for survey research. His work began in the late 1950's (Yates and Simpson 1960). Originally, programs were written that were specific to each survey. This evolved into a general-purpose program by the early 1960's (Simpson 1961). Although it was the first in the field and was available for many years, it never achieved widespread popularity. There are at least four reasons for its general lack of success, reasons that point to the success of other software developers.

(1) The package was not user friendly. In his obituary of Yates, Dyke (1995) made allusion to this fact. He says:

> "Yates believed that the analyst should understand the relevant theory, and so be ready to specify in exact detail what he wanted. Perhaps for this reason the program was not excessively easy to use! But its power and flexibility, and uncluttered clarity of its output were, and are, outstanding."

(2) It was too expensive for what it did and could not compete with cheaper competitors. Wolter (1985) lists a number of packages that were available in the mid-1980's. At the time the package was twice as expensive as SUDAAN but could do only tabulations, whereas SUDAAN had the additional capability of regression analysis and ratio estimation.

(3) Marketing is an important factor in the success of a product. Yates appeared to be more interested in tinkering with his product to improve it rather than investing time in marketing it.

(4) Other than a manual, by 1985 there was no technical support for the package.

Yates was not alone in having software that did not catch on. I had the same experience when I developed variance estimation software based on tree traversal algorithms (Bellhouse 1985). Other than the expense factor (mine was free), my package was a living example for the other three reasons why some software does not fly.

By the early 1970's there were over 40 packaged programs and routines, written mainly in FORTRAN, that would do statistical analyses (Schucany, Minton and Shannon 1972). Of these original packages only two have remained popular in the marketplace, SAS first released in 1970 and SPSS released in the late 1960's.

The survey software that has maintained predominance in the market for several years is SUDAAN developed by B.V. Shah of the Research Triangle Institute (Shah 1978 and 1984). It is marketed well and fully supported by its developer. It was originally accessed as a SAS procedure and has now become a stand-alone package. The tie with SAS was probably one of the reasons for its initial success. Those who were familiar with SAS could easily familiarize themselves with this new procedure, or equivalently the package, so that in a sense it was user friendly. Further, the package has continued to keep pace with survey research. The original program contained routines to calculate standard errors for survey estimates including means, totals, proportions and ratios. This was expanded to include regression analysis in the late 1970's when research on regression in complex surveys was under way. The program now contains routines for regression analysis, logistic regression, categorical data analysis and survival analysis. It has also kept pace with developments in computing machinery. Originally developed on a mainframe computer, the package is now available for use on a PC. It still maintains its links to SAS, although SAS currently has its own survey analysis procedures under development.

Currently, there are several other programs for survey analysis. The most popular among these programs, in addition to SUDAAN, are STATA and WesVarPC. While SUDAAN has been linked to SAS, the future development of WesVarPC, which was originally developed by the research corporation Westat, has been turned over to SPSS. Further, the survey routines in STATA are part of a larger statistical analysis package. As with mergers in the general business world, along with product and service integration, the future trend for survey data analysis packages is to become part of an omnibus statistical package. The development and maintenance of statistical packages, for survey research or for a wider context, is a time-consuming enterprise requiring a substantial capital investment. This can only be done by a well-financed organization.

SUDAAN, STATA and WesVarPC, along with the software packages GES from Statistics Canada and another named CLAN, have been recently reviewed and evaluated in Bergdahl, Black, Bowater, Chambers, Davies, Draper, Elvers, Full, Holmes, Lundqvist, Lundström, Nordberg, Perry, Pont, Prestwood, Richardson, Skinner, Smith, Underwood and Williams (1999). SUDAAN and STATA have also been evaluated by Cohen (1997). Among three of the packages reviewed (STATA, SUDAAN and WesVarPC), SUDAAN appears to have the most options. For example, Bergdahl *et al.* (1999) note that SUDAAN carries out variance estimation for complex statistics using any one of Taylor linearization, jackknife and balanced repeated replication. WesVarPC covers jackknife and balanced repeated replication, while STATA relies solely on Taylor linearization. So far, none of the packages does variance estimation using the bootstrap. It may just be a matter of time before this technology is incorporated into these packages. For some of its public use sample files, Statistics Canada provides bootstrap variance estimation procedures in SAS code. These procedures, however, are specific to the surveys in question.

7. MODELS IN SAMPLING

Models have come in and out of favour among sampling practitioners. Due to Neyman's (1934) pioneering work, the paradigm of randomization and the randomization distribution was paramount until the 1960's. However, the use of models did not disappear during the intervening years. Cochran (1946), for example, used models to study certain sampling designs and was able to conclude that systematic sampling was a good design to use under certain population structures. The 1960's debate over models arose out of the questioning of the foundations of sampling initiated by Godambe (1955). Since then the use of models has not only crept back in to sampling theory but has flourished substantially.

Since the 1960's the use of models in sampling has gone in several directions. At the same time, the practical and general use of models in survey estimation and analysis is only feasible with high speed computing and the appropriate software. In keeping with the theme I have been following here, I will take a very narrow approach to models by tying their usage to computing technology.

Several model-related methodologies have been computerized, either through the provision of numerical examples to illustrate the use of the methodology or through simulation studies to examine how the methodology works. At the present time there is only one model-related approach that has matured to the point where a general package program is available. This is the model-assisted approach that C.-E. Särndal has taken over several years resulting in

generalized regression estimation or GREG. The bulk of the work is summarized in Särndal, Swensson and Wretman (1992). The work was initially motivated by the debates over the foundations of sampling. Under a model, a best, in some sense, estimator of a finite population parameter can be derived. Those on the side promoting randomization inference pointed out that when the model fails the associated estimator can perform very poorly. The solution propounded by Särndal was to obtain the estimate under the model and then to adapt it in such a way that it would remain consistent and perform adequately under the randomization distribution. It is an attempt to obtain the best of both worlds. Generalized regression estimation, as well as several other estimators, have been programmed into GES, a generalized estimation system developed at Statistics Canada. This SAS-based software is aimed at the descriptive side of surveys rather than the analytic and is described in Estevao, Hidiroglou and Särndal (1995). It is a package that could easily catch on under the right conditions.

8. CONCLUSIONS

Developments in sampling research are inextricably tied to computing and computational methods. Where research is headed will be guided, in part, by computer developments. What the immediate future holds for computing is greater speed and greater storage capacity so that packages can become bigger and more comprehensive. Generally acceptable practices in survey estimation and the analysis of survey data will be determined by the contents of generally available computer packages for survey sampling. On the research methodology side, new methodology will continue to be increasingly computer intensive. One other foreseeable development is the explosion of the internet. As a result of this explosion, several complete survey datasets are now easily available via the web. The extensive testing of new methodology on a variety of real surveys prior to publication of the methodology may soon become the norm.

ACKNOWLEDGEMENT

This work was supported by a grant from the Natural Sciences and Engineering Research Council of Canada.

REFERENCES

ANDREWS, D.F., and STAFFORD, J.E. (1993). Tools for symbolic computation of asymptotic expansions. *Journal of the Royal Statistical Society, B*, 55, 613-628.

BAINES, J.A. (1900). On census-taking and its limitations. *Journal of the Royal Statistical Society*, 63, 41-71.

BAYLESS, D.L., and RAO, J.N.K. (1970). An empirical study of stabilities of estimators and variance estimators in unequal probability sampling (n = 3 or 4). *Journal of the American Statistical Association*, 65, 1645-1667.

BELLHOUSE, D.R. (1985). Computing methods for variance estimation in complex surveys. *Journal of Official Statistics*, 1, 323-329.

BELLHOUSE, D.R. (1988). A brief history of random sampling. *Handbook of Statistics*. (Eds. C.R. Rao and K.R. Krishnaiah) 6, 1-14. Amsterdam: North-Holland.

BENJAMIN, B. (1961). The 1961 census of population. *Incorporated Statistician*, 11, 130-143.

BERGDAHL, M., BLACK, O., BOWATER, R. CHAMBERS, R., DAVIES, P., DRAPER, D., ELVERS, E., FULL, S., HOLMES, D., LUNDQVIST, P., LUNDSTRÖM, S., NORDBERG, L., PERRY, J., PONT, M., PRESTWOOD, M., RICHARDSON, I., SKINNER, C., SMITH, P., UNDERWOOD, C., and WILLIAMS, M. (1999). *Model Quality Report in Business Statistics Volume II: Comparison of Variance Estimation Software and Methods*. London: Office of National Statistics.

BOWLEY, A.L. (1906). Address to the Economic and Statistics Section of the British Association for the Advancement of Science, York. *Journal of the Royal Statistical Society*, 69, 540-558.

BOWLEY, A.L. (1926). Measurement of the precision attained in sampling. *Bulletin of the International Statistical Institute* 22 (1), 1-62.

BOWLEY, A.L. (1936). The application of sampling to economic and sociological problems. *Journal of the American Statistical Association*, 31, 464-480.

BOX, K., and THOMAS, G. (1944). The Wartime Social Survey. *Journal of the Royal Statistical Society*, 107, 151-189.

BREWER, K.R.W., and HANIF, M. (1983). *Sampling with Unequal Probabilities*, (Lecture Notes in Statistics, Volume 15). New York: Springer-Verlag.

CERUZZI, P.E. (1998). *A History of Modern Computing*. Cambridge, Massachusetts: MIT Press.

COCHRAN, W.G. (1946). Relative accuracy of systematic and stratified random samples for a certain class of populations. *Annals of Mathematical Statistics*, 17, 164-177.

COCHRAN, W.G. (1963). *Sampling Techniques*, (2nd Edition). New York: Wiley.

COHEN, S.B. (1997). An evaluation of alternative PC-based software packages developed for the analysis of complex survey data. *American Statistician*, 51, 285-292.

DAY, N. (1971). *Canadian Computer Census* 1971. Toronto: Canadian Information Processing Society.

DEMING, W.E. (1953). On the distinction between enumerative and analytic surveys. *Journal of the American Statistical Association*, 48, 244-255.

DEMING, W.E. (1956). On simplifications of sampling designs through replication with equal probabilities and without stages. *Journal of the American Statistical Association*, 51, 24-53.

DURBIN, J. (1959). A note on the application of Quenouille's method of bias reduction to the estimation of ratios. *Biometrika*, 46, 477-480.

DYKE, G. (1995). Obituary: Frank Yates. *Journal of the Royal Statistical Society, A*, 158, 333-338.

EFRON, B. (1982). *The Jackknife, the Bootstrap and Other Resampling Plans*. Philadelphia: Society for Industrial and Applied Mathematics.

ESTEVAO, V., HIDIROGLOU, M.A., and SÄRNDAL, C.-E. (1995). Methodological principles for a generalized estimation system at Statistics Canada. *Journal of Official Statistics*, 11, 181-204.

FAN, C.T., MULLER, M.E., and REZUCHA, I. (1962). Development of sampling plans by using sequential (item by item) selection techniques and digital computers. *Journal of the American Statistical Association*, 57, 387-402.

FELLEGI, I.P. (1963). Sampling with varying probabilities and without replacement: rotating and non-rotating samples. *Journal of the American Statistical Association*, 58, 183-201.

FELLEGI, I.P., GRAY, G.B., and PLATEK, R. (1967). The new design of the Canadian Labour Force Survey. *Journal of the American Statistical Association*, 62, 421-453.

FULLER, W.A. (1975). Regression analysis for sample survey. *Sankhyā, C*, 37, 117-132.

GILLIES, D. (1992). *Revolutions in Mathematics*. Oxford: Clarendon Press.

GODAMBE, V.P. (1955). A unified theory of sampling from finite populations. *Journal of the Royal Statistical Society, B*, 17, 269-278.

HANSEN, M.H. (1987). Some history and reminiscences on survey sampling. *Statistical Science*, 2, 180-190.

HANSEN, M.H., and HURWITZ, W.N. (1943). On the theory of sampling from finite populations. *Annals of Mathematical Statistics*, 14, 333-362.

HANSEN, M.H., HURWITZ, W.N., NISSELSON, H., and STERNBERG, J. (1955). The redesign of the Current Population Survey. *Journal of the American Statistical Association*, 50, 701-719.

HARTLEY, H.O. (1946). The application of some commercial calculating machines to certain statistical calculations. Supplement to *Journal of the Royal Statistical Society*, 8, 154-183.

HIDIROGLOU, M.A., FULLER, W.A., and HICKMAN, R.D. (1980). *SUPER CARP*. Ames: Iowa State U.P.

HOLLERITH, H. (1894). The electrical tabulating machine. *Journal of the Royal Statistical Society*, 57, 678-689.

HOOKER, R.H. (1894). Modes of census-taking in the British Dominions. *Journal of the Royal Statistical Society*, 57, 289-368.

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.

KAIER, A.N. (1895/6). Observations et experiences concernant des dénombrements representatives. *Bulletin of the International Statistical Institute*, 9, 176-183.

KAIER, A.N. (1897). *The Representative Method of Statistical Surveys* (1976, English translation of the original Norwegian). Oslo: Central Bureau of Statistics of Norway.

KAIER, A.N. (1905). Untitled speech with discussion. *Bulletin of the International Statistical Institute*, 14, 119-134.

KISH, L. (1957). Confidence intervals for clustered samples. *American Sociological Review*, 22, 154-165.

KISH, L., and FRANKEL, M.R. (1970). Balance repeated replication for standard errors. *Journal of the American Statistical Association*, 65, 1071-1094.

KISH, L., and FRANKEL, M.R. (1974). Inference from complex samples. *Journal of the Royal Statistical Society B*, 36, 1-37.

KLEIN, L.R., and MORGAN, J.N. (1951). Results of alternative statistical treatments of sample survey data. *Journal of the American Statistical Association*, 46, 442-460.

KONIJN, H.S. (1962). Regression analysis in sample surveys. *Journal of the American Statistical Association*, 57, 590-606.

KRUSKAL, W., and MOSTELLER, F. (1980). Representative sampling, IV: the history of the concept in statistics 1895 - 1939. *International Statistical Review*, 48, 169-195.

LANSING, J.B., and MORGAN, J.N. (1971). *Economic Survey Methods*. Ann Arbor: Survey Research Center.

MAHALANOBIS, P.C. (1946). Recent experiments in statistical sampling in the Indian Statistical Institute. *Journal of the Royal Statistical Society*, 109, 325-378.

MANDEVILLE, J.P. (1946). Improvements in methods of census taking and survey analysis. *Journal of the Royal Statistical Society*, 109, 111-129.

MCCARTHY, P.J. (1969). Pseudo-replication: half samples. *Review of the International Statistical Institute*, 37, 239-264.

MURTHY, M.N. (1967). *Sampling Theory and Methods*. Calcutta: Statistical Publishing Society.

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

PFEFFERMANN, D. (1993). The role of sampling weights when modeling survey data. *International Statistical Review*, 61, 317-337.

PORTER, R.D. (1973). On the use of survey sample weights in the linear model. *Annals of Economic and Social Measurement*, 2, 141-158.

RAO, J.N.K., and BAYLESS, D.L. (1969). An empirical study of stabilities of estimators and variance estimators in unequal probability sampling of two units per stratum. *Journal of the American Statistical Association*, 64, 540-559.

RAO, J.N.K., and SCOTT, A.J. (1981). The analysis of categorical data from complex sample surveys: chi-squared tests for goodness of fit and independence in two-way tables. *Journal of the American Statistical Association*, 76, 221-230.

RAO, J.N.K., and SCOTT, A.J. (1984). On chi-squared tests for multiway contingency tables with cell proportions estimated from survey data. *Annals of Statistics*, 12, 46-60.

RAO, J.N.K., and THOMAS, D.R. (1988). The analysis of cross-classification data from complex sample surveys. *Sociology Methodology*, 18, 213-269.

RAO, J.N.K., and WU, C.F.J. (1987). Resampling inference with complex survey data. *Journal of the American Statistical Association*, 83, 321-241.

SÄRNDAL, C.-E., SWENSSON, B., and WRETMAN, J.H. (1992). *Model Assisted Survey Sampling*. New York: Springer-Verlag.

SCHUCANY, W.R., MINTON, P.D., and SHANNON, B.S. (1972). A survey of statistical packages. *Computing Surveys*, 4, 65-79.

SEDRANSK, J. (1965). A double sampling scheme for analytical surveys. *Journal of the American Statistical Association*, 60, 985-1004.

SHAH, B.V. (1978). SUDAAN: Survey data analysis software. *Proceedings of the Statistical Computing Section, American Statistical Association.*

SHAH, B.V. (1984). Software for survey data analysis. *American Statistician*, 38, 68-69.

SIMPSON, H.R. (1961). The analysis of survey data on an electronic computer. *Journal of the Royal Statistical Society, A*, 124, 219-226.

SKINNER, C.J., HOLT, D., and SMITH, T.M.F. (1989). *Analysis of Complex Surveys*. New York: Wiley.

STAFFORD, J.E., and ANDREWS, D.F. (1993). A symbolic algorithm for studying adjustments to the profile likelihood. *Biometrika*, 80, 715-730.

STAFFORD, J.E., and BELLHOUSE, D.R. (1997). A computer algebra for sample survey theory. *Survey Methodology*, 23, 3-10.

WILLCOX, W.F. (1926). The past and future developments of vital statistics in the United States I: John Shaw Billings and federal vital statistics. *Journal of the American Statistical Association*, 21, 257-266.

WOLTER, K.M. (1985). *Introduction to Variance Estimation*. New York: Springer-Verlag.

WOODRUFF, R.S. (1971). A simple method for approximating the variance of a complicated estimate. *Journal of the American Statistical Association*, 66, 411-414.

WORTON, D.A. (1998). *The Dominion Bureau of Statistics: A History of Canada's Central Statistical Office and Its Antecedents*, 1841-1972. Montreal and Kingston: McGill-Queen's University Press.

YATES, F. (1960). *Sampling Methods for Censuses and Survey*, (3rd edition). London: Griffin.

YATES, F. (1973). The analysis of surveys on computers – features of the Rothamsted Survey Program. *Applied Statistics*, 22, 161-171.

YATES, F., and SIMPSON, H.R. (1960). A general program for the analysis of surveys. *Computer Journal*, 3, 136-140.

Survey Methodology, June 2000
Vol. 26, No. 1, pp. 21-30
Statistics Canada

The Past is Prologue

BARBARA A. BAILAR[1]

ABSTRACT

Mahalanobis provided an example of how to use statistics to enlighten and inform government policy makers. His pioneering work was used by the US Bureau of the Census to learn more about measurement errors in censuses and surveys. People have many misconceptions about censuses, among them who is to be counted and where. Errors in the census do occur, among them errors in coverage. Over the years, the US Bureau of the Census has developed statistical techniques, including sampling in the census, to increase accuracy and reduce response burden. A root-mean-square-error model was developed to estimate the joint effects of variance and bias in the census. The model is used in this paper to look at the joint effects of response variance, adjustment of the bias caused by the undercount, and the use of sampling for follow-up.

KEY WORDS: Censuses; Mahalanobis; Root-mean-square-error model; Sampling in the census.

1. INTRODUCTION

Perhaps it has always been so – that statistics, as a body of information, does not always support the actions that politicians want to take. In some countries, data from censuses are not made public, because knowledge is power. However, in our society, the power of statistics is used to inform us about needs for action, or how well we're doing as a country, or as the basis of comparison among groups. We are used to seeing and trusting statistics on an everyday basis, though most of us give little attention to how they are produced, by whom, and at what cost.

Over the last few decades, there have been many issues where statistics and politics have been in conflict. Employment and unemployment data are often used by politicians, especially in an election year. If the unemployment figures are low, the incumbents cite that figure and take the credit. If the employment figures show that many new jobs are being created, that number is cited. Either political party can use these data to make whatever political points seem salient. An attempt by the Nixon Administration to restrict access to these data led to new protections, such that the employment and unemployment data are released on the first Friday of every month by the Commissioner of the Bureau of Labor Statistics at a meeting of the Joint Economic Committee on Capitol Hill.

The definition of poverty is currently under discussion. When the poverty measure was invented by Molly Orshansky, there were not the large transfer payment systems that exist today. Because of income received or benefits paid, poverty today does not mean what poverty did 30 years ago. However, each political administration watches the poverty numbers very closely. These numbers were used by critics of the Reagan Administration to illustrate the growing burden of the poor in an administration that was alleged to be more interested in serving the rich. That Administration argued that by including medical benefits and other transfer payments, the poor were better off than before.

Probability samples of the U.S. population are now used to study sexual behavior. Much of our information on sexual behavior goes back to Kinsey. The National Opinion Research Center (NORC) at the University of Chicago has conducted two large surveys of sexual behavior in the U.S. One of these, *Sex in America*, (Michael, Gagnon, Laumann and Kolata 1994) reported on a national sample of persons aged 18-59, and was not funded by the government. The second researched the sexual behavior of adolescents and, in both cases, federal funding for these studies was questioned because powerful constituencies did not want the subject matter to be examined. The second study was finally funded by the government.

Privacy issues abound. For example, there is broad concern about the confidentiality of individual medical records and the need for researchers to access them. Privacy issues for groups are less widely recognized. Certain groups may not want to report fully in a decennial census or survey because they do not want to attract attention. Though people who are in the country illegally are supposed to be included in the census, many of them fear that government authorities looking at block statistics could use the information to raid certain blocks.

My last example here of issues in which politics and statistics are having a disagreement, is the decennial census. For decades, an undercount in the census and its differential impact on minority populations has been well-documented. The Census Bureau has studied this issue for years and now has the statistical tools and methods to represent the uncounted individuals in the census totals. Yet this "adjustment" is opposed by many politicians because of an anticipated effect on the drawing of election district boundaries. However, the uses of the census extend far beyond apportionment and redistricting. The battle before the 2000 Census has been unusually intense.

[1] Barbara A. Bailar, National Opinion Research Center, 1155 East 60th Street, Chicago, Illinois, U.S.A.

Given these instances in which politics and statistics are confronting each other, it is useful to step back in time to review the contributions of Mahalanobis to the government of India. His methods were used successfully by the U.S. Census Bureau to learn much of what we know about errors in the census. I will review Mahalanobis' contributions, then return to a discussion of the census, the statistical tools currently used in the census, additional tools that could be used, and then conclude with a plea for Congress and the Census Bureau to follow the tradition of continuous improvement in the census through the use of statistical tools.

2. THE MAHALANOBIS LEGACY

Mahalanobis played an important role in the methodology we take for granted today. He was trained to teach physics, but became increasingly interested in statistical problems and then in building the Indian Statistical Institute. His work on the utilization of interpenetrated subsamples of the population was innovative, and gave great impetus to research on the effects of interviewers on survey and census statistics. He paid great attention to the need for pilot studies to test the implementation of survey techniques. As time went along, he enlarged his interests from sampling and surveys, in which he provided much needed information to the government, to planning and economic development. He was appointed Honorary Statistical Advisor to the Cabinet in January, 1949 and placed in charge of the Central Statistical Unit in the same year. The central role of statistics in government planning was, no doubt, due to the force of the man himself as well as his research findings. He saw the role of statistics as a system to serve the cause of planned development and envisioned a feedback arrangement between statistics and planning (Rudra 1996).

The particular contributions I wish to stress today are his major roles in sample surveys and in measuring error of all kinds – errors of observation, errors of measurement, sampling errors, copying errors, printing errors. Much of his early work on showing the variability in statistics caused by interviewers was in crop statistics (Mahalanobis 1950). He was one of the first to say, and then show, that the overall error in survey statistics was not just sampling variance but also the variance arising from the human element. One way to study such errors was by the use of interpenetrated subsamples. In the words of Mahalanobis,

> "When two (or more) samples are drawn from the same population and covered according to the same survey design, the results based on the different samples are equally valid, even though they are derived by different operational units; and divergences between the different sets of estimates supply directly some idea of the margin of uncertainty." (Mahalanobis and Lahiri 1961)

Mahalanobis demonstrated that statistics based on samples were at least comparable to, and often more accurate than statistics based on a census, in the 1940's, when sampling was still not fully accepted. He believed, as many of us now do, that samples can be better controlled than can a census. He stated (Mahalanobis and Lahiri 1961) that the magnitude of discrepancies found in a census of jute production made it appear that a census may not provide accurate estimates for small areas. The random component of the non-sampling error may add enough error that results for a large area may be no different from those obtained by a sample survey. What holds for a large area does not naturally follow for small areas.

The U.S. Census Bureau used Mahalanobis' techniques to learn more about the underlying variability of census numbers.

3. WHAT DO PEOPLE THINK A CENSUS IS

To most people, taking a census means that enumerators go out and count everyone. There are three things that people seem to think about censuses. One is that everyone is counted. A second is that an enumerator sees everyone. A third is that the census is without error. Let's look at these one by one.

Often, everyone is not supposed to be counted in a national census, and who should be counted varies from country to country, and over time within a country. For example, military personnel and their families located outside of the country could be counted or not. Civilian aliens temporarily in the country as seasonal workers could be counted or not. From these illustrations one can see that a primary necessity in census-taking is defining the scope of the census.

So, by definition, certain groups of people are not to be counted in the census. This is by design of the Census Bureau. Other people make individual or family decisions not to be counted in the census. In earlier times, some families did not report children who suffered from some diseases or retardation. Some people who have had unfortunate episodes with the legal system may decide not to be counted. These may be people who are in the country illegally, those who are hiding from law enforcement, and those who fear, for whatever reason, the consequences of being counted. In 1990, there were people who said they would not be counted because they thought the census was too intrusive.

Finally, there are people missed, not by design but by accident. Perhaps they lived in buildings that were missed, perhaps they lived on the street and were missed. Perhaps they were away during the census period. During 1998 there were many reports of how much harder it was to survey people who live in gated communities. It may be that some of these people are missed because of the overzealousness of the community guards. In some communities good

maps are unavailable or not updated, so groups of people may be missed.

In any case, not everyone is counted in a census and never was.

The second myth to be refuted is that an enumerator sees everyone and knows who should be in the census or not. This never happened, even in the early censuses in the U.S., when U.S. Marshals took the census and the country was much smaller. In fact, early censuses were of households, not of individuals. This means that there were no questions asked of individuals but instead there was interest in how many people were in the household, how many were men and how many were women, how many were in different age groups, and so forth. The totals were posted in public places. Starting in 1880 the canvasser method of taking a census, where enumerators went from door to door, came into being. It is this kind of census that made some believe that an enumerator saw everyone. However, a single household member usually responded for the whole family. The enumerator did not see those who were sick, those at work, those who were away temporarily, or those who were, for some reason or another, not in the room when the enumerator visited.

Though the enumerator-type census was an improvement over one taken by the marshals, research using interpenetrated subsamples showed that census enumerators still added a considerable amount of error to the census statistics. The enumerators were influenced by their own expectations and by responses of others in their enumeration district. Also, some did not understand the instructions and reported things incorrectly. An experiment in the 1950 census showed that enumerators added considerable variability to the census statistics (Hanson and Marks 1958). Indeed, the statistics gathered from a census had the same level of variability, due to enumerators, as a 25-percent sample. This is the main reason the Census Bureau turned to the use of self enumeration in the 1960 census and progressively expanded it in later censuses. Now, if a household receives the census form by mail, fills it out, and sends it in, and no errors require resolution, no enumerator will call at the household.

The third myth is that census taking occurs without error. No one who now works on censuses or surveys believes that, but other people do. The Census Bureau encourages that belief by publishing data down to the last digit. For example, the population of the United States in 1990 was reported and published as 248,718,301 in the *Statistical Abstract*.

Even some of those who have worked closely with a census cannot see it as a statistical process that carries with it a certain amount of error. Because the error is not routinely quantified and published along with the census numbers, some cannot believe the error exists. Some persons working in the Population Division of the U.S. Census Bureau in the 1940's and 50's believed that the census was the best way to learn about any subject, and that sample surveys were inferior. Repeated demonstrations of accuracy in survey results and of bias in census data did not change their minds.

Anyone who comes into regular contact with the census now knows that there is error in the data. First, though sampling variance cannot occur for items collected on a 100-percent basis, there may still be substantial response variance introduced by effects of enumerators, respondents, and coders on census data. Second, bias affects responses to many census questions even when a person is correctly counted. Bias also affects counts when enumerators do not count everyone. The Census Bureau conducts an evaluation program as part of every census, documents the amount of error, and uses those data to attempt to improve the next census.

Large groups of people are affected by census error. The undercounting bias affects minority populations and children at a much higher rate than other populations (Edmonston and Schultze 1993). Thus, communities that are largely African-American, Hispanic, or American Indian are underrepresented in distributions of potential power and money, while those statistics that are based on children under 10 are subject to a large error.

Over the years, the Census Bureau has reported numerous studies looking at the balance between cost and accuracy. One mentioned before is the use of self-enumeration. At smaller levels of population, the effect of response variance, primarily caused by interviewers, was very high. Just as with sampling error, as the size of the area increased, and the number of enumerators who collected the data increased, the effect lessened. When the mail return rate was close to 80 percent, the response variance decreased to about one-quarter of that of a 25-percent sample (Bailar 1969).

Thus, commonly held images of the census are not always true. Also, the census is not always the same. The Census Bureau has made many changes in census taking since the first census in 1790. The number of questions, the kinds of questions, who is counted and where, who does the counting, how people are assigned to a geographic domain, how missing characteristics are handled, and the gradual increase of asking most questions of a sample have changed over the years. The next section shows how the use of statistical tools has changed the census in this century.

4. DEVELOPMENT OF STATISTICAL TOOLS IN A CENSUS

Two elements have changed the methods of the U.S. Decennial Census considerably since 1940: the use of computers; and the use of statistical techniques. At times, the two elements have complemented each other, for example in the fast processing for imputation of missing data using a "hot deck" procedure. While computers have profoundly affected the census, the remainder of this discussion will focus on the statistical methodology.

One of the major advances starting in 1940 has been the use of sampling in the census. In 1940, as documented by Waksberg and Hanson (1965), there were three major uses of sampling. One was for the collection of data deemed supplementary to the main census questions. Questions such as mother tongue, veteran status, and fertility were asked of a 5-percent sample. A second use was for certain analytic studies requiring clerical transcription and coding. To avoid a long timespan for the transcription and coding to take place, a sample of census questionnaires was selected and the transcription and coding occurred only for them. A third use was for the verification of large-scale clerical operations such as editing, coding, key-punching, and so forth. Prior to 1940, all verification was on a 100-percent basis.

To describe the next leap forward, Waksberg and Hanson said:

> "A major step forward in the use of sampling in census work took place in the 1950 Census of Population and Housing. This grew out of a profound change in attitude regarding the role of sampling. Whereas in 1940 sampling had been considered applicable only for items of supplementary and secondary interest, in 1950 the entire range of census activities was examined to determine, on a logical basis, where complete counts were necessary and where samples could provide adequate information."

The increased use of sampling for population characteristics, for sample tabulations, and for verification was successful and evaluations showed that, even with the addition of sampling error, overall error was less than if earlier techniques had been used with no sampling. This was a reinforcement of the lesson learned earlier by Mahalanobis.

During the 1950 Census, the Bureau did a great deal of research to learn the effect of response biases and response variances on census data. Waksberg and Hanson declared that it was misleading to assume that the census, without sampling, was without error. In 1950, an experiment was conducted to estimate the effect of census enumerators on census data. By using the method of interpenetrated subsamples introduced by Mahalanobis, pairs of adjacent census areas were merged and assignments to the enumerators were randomized. Since the assignments were over the same area, differences between enumerators did not reflect differences in the type of area. The main finding of the study was that a full census in which enumerators went door to door to collect the census information had response variability that made the census the equivalent of a 25-percent sample (Hanson and Marks 1958). Using that result, as well as studies of biases in various census items, Waksberg and Hanson formulated a model in which census results were subject to a relative response bias of 6 percent and a response variance equal to the sampling variance of a 25-percent household sample. They used this model to generate Table 1 which shows the magnitude of total error in census data with and without sampling.

The authors point out that for a characteristic describing 500 individuals in an area of 2,500 people, the increase in the total root mean square error arising from sampling variability is only about 25%. For larger areas and larger cells, the additional error due to sampling is even smaller.

These data were studied carefully before the decision to increase the use of sampling in the 1960 Census. In practice, sampling made even greater gains than those anticipated by the model. The authors state "Thus for a great many published statistics, the reliability was better with the use of sampling than would have been possible otherwise." (Waksberg and Hanson 1965.)

Table 1

Expected Root Mean Square Error (RMSE) of Estimated Cell Frequencies for Individual Items Based on a Complete Census and On a 25-Percent Sample of Households

Area of 2,500 Population having RMSE based on			Area of 10,000 Population having RMSE based on			Area of 50,000 Population having RMSE based on		
Cell Frequency	Complete Census	25-percent Sample	Cell Frequency	Complete Census	25-percent Sample	Cell Frequency	Complete Census	25-percent Sample
12	7	10	50	1	20	250	34	46
50	14	19	200	30	40	1,000	85	105
125	22	31	500	52	67	2,500	180	200
500	49	62	2,000	140	160	10,000	620	650
1,250	89	102	5,000	320	330	25,000	1,520	1,530

Note 1: Computations assume a relative response bias of 6 percent and response variance equal to the sampling variance for a 25-percent sample.

Note 2: The accuracy of the results (cell frequencies) is measured by a certain kind of average of the actual errors that would occur, the root mean square error (RMSE). A useful working rule would be to assume that approximately two-thirds of all results of a census or a sample would differ from their true cell frequencies by no more than their RMSE's.

A large-scale evaluation and research program of the Decennial Census program began in the 1950's and is now an integral part of the Census. Part of the program tests new methods for possible use in the following census and part of it focuses on the evaluation of the current census. It was as part of this program that the Bureau started measuring the undercount in the census. It was also this program in which the response variance due to enumerators was measured before and after the advent of self-enumeration. (In 1960, after self-enumeration was introduced, the response variance decreased to 1/4 of the 1950 level.) Since mail-back rates have decreased substantially since 1980, that variance may have increased again, perhaps substantially.

Other studies included research on alternative ways to measure the undercount, record checks to measure the accuracy of census data, and a study of using the Post Office not only to deliver census questionnaires but to notify the Census about missed addresses and duplicate forms.

Sampling is now used extensively to control the quality of the large-scale clerical tasks associated with the census. In past censuses, verification was usually dependent, in which the verifier reviewed the coder's work and determined whether the correct codes had been assigned. The Bureau planted errors and found that dependent verification missed as many as half of the errors. This and other research caused the Bureau to develop independent verification, in which records are assigned to three coders who do not see each other's work. A "majority rule" is used to determine the best code, and statistics about such errors are used to improve the process and to identify substandard performance.

Imputation was also a necessary tool developed for use in the census. To keep within time and budget parameters, the Bureau developed a "hot-deck" imputation system, based on the assumption that people who live in proximity are likely to resemble each other for many characteristics such as educational attainment and income. Another kind of imputation was also used in 1970, 1980, and 1990 to deal with a small, residual set of addresses left on the mailing list with no information about whether or not they were occupied. No one answered the door, nor did neighbors know if anyone lived there. Thus, based on a model that assumed a high correlation between the characteristics of neighboring households, the Bureau imputed occupancy or vacancy status, and to those imputed as occupied, a number of people were imputed. In 1980, only 762,000 persons were imputed, about .003 of the total census count, but they were not spread evenly over all the States. As a result of the imputation, Indiana lost a Congressional seat to Florida. However, it should be acknowledged that doing nothing about the unclassified units would have been equivalent to imputing them all as vacant. There was information available that showed that over half of these units could be expected to be occupied so the data based on imputation were more accurate than data based on counts alone with no imputation.

5. ADDITIONAL USES OF STATISTICAL TOOLS

Statistical tools can be used to correct the census for the undercount. The Waksberg-Hanson root mean-square error model estimates the amount of error in the census assuming a relative response bias in the overall census of 2 percent. (The 1990 estimate was 1.6 percent.) Also assume a response variance in both the adjusted and unadjusted census equal to one-fourth the sampling variance of a 25-percent sample. That estimate may now be too low since decreasing mail-back rates have driven enumerator variances higher. However, to be on the conservative side, we shall use the 1960 and 1970 measurements.

The model is the simple mean-square error model used frequently by the Census Bureau.

$$\mathrm{MSE}(T) = \mathrm{Var}(T) + B_T^2$$

Assume T is a cell size or a size of interest in the census in an area where N is the population size. $T = NP$ where P is the proportion of the population having a certain characteristic. B is the bias in the census count. So, for example, in an area of 2,500 people, one might be interested in knowing the number of children under 10 years of age. $N = 2{,}500$ and $T = NP$.

Now the variance of an estimated proportion, p is:

$$V(p) = \frac{N-n}{N-1} \cdot \frac{1}{n} \cdot PQ$$

If we have a 25 percent sample, this reduces to

$$V(p) = \frac{3}{4} \cdot \frac{1}{n} \cdot PQ = \frac{3PQ}{N}$$

$$V(T) = V(Np) = N^2 V(p) = 3NPQ$$
$$= 3TQ$$

Relative bias = (.02) so Bias = $.02T$

Now we are dealing with a census, so there is no sampling variance, but the response variance is equal to 1/4 of what the sampling variance would be. So

$$\mathrm{MSE}(T) = (.02T)^2 + (.25)(3)TQ$$

and

$$\mathrm{RMSE}(T) = \sqrt{(.02T)^2 + (.25)(3)TQ}$$

This formula has been used as the basis of the calculations in Table 2. For an unadjusted census, RMSE(T) would have both the bias and variance components. For an adjusted census, the relative bias is zero, so only the response variance term remains. However, this analysis presumes that the adjustment factors themselves are free from any kind of variance and bias, and that the same adjustment factors can be uniformly applied within the demographic groups.

For example, Table 2 shows that for a total of 500 in an area of 2,500, the RMSE for an unadjusted census is 20 while the RMSE for an adjusted census is 17. For the unadjusted census, the contribution from the bias term is small, $[(.02)(500)]^2=100$. The contribution from the response variance is $(.25)(3)(500)(.8)=300$. So RMSE $=\sqrt{400}=20$. For the adjusted census, the bias term, 100. is removed, so the RMSE$=\sqrt{300}=17$. However, if one considers that the estimated bias term has both variance and bias, there may be little difference between the adjusted and unadjusted results for a small area. As the total, T, gets larger, the bias term is more dominant, and the adjustment removes more error.

Table 2 shows that for a small area of 2,500 persons there is no gain for small totals, but a gain of 43 percent in accuracy for a large total of 1,500 persons. In a somewhat larger area of 10,000 persons, there is little reduction in error until a total of 1,000 is of interest, where there is a gain in accuracy of 21 percent and for a large total of 5,000, there is a gain of 61 percent. Thus, if we were talking about the number of men or women in an area of 10,000, a total that might be expected to be around half the population, there would be a large gain in the accuracy of the total from using adjusted census figures. For an area of 50,000 the bias term dominates the mean square error, even at smaller totals such as 1,000. Here the gain is 21 percent, which grows to 81 percent for a very large total of 25,000.

This illustration shows is that adjusting the census does not add to the error of the census, even for small areas and small cells, if one assumes that the bias term is measured without error. For smaller area sizes and smaller cells, the response variance dominates the mean square error, but the total error is never less than the response variance. When the census is adjusted, the bias term goes to zero, and the gains in accuracy are dramatic.

One virtue of this model is that it was developed by the Census Bureau long before the current debate on adjustment grew heated. It was used to disabuse people of the idea that the census cells have no error. It was used successfully to show critics that having most of the census questions answered by only a sample would not hurt the data unduly. Such a tried and true census model now shows the real value of adjustment.

Table 2 used the relative response bias of 2 percent based on the 1990 Census overall estimate of the undercount of 1.6 percent. However, since the undercount hits minority populations harder, let's look at a comparison of an adjusted and unadjusted census in which the relative bias is 4 percent. (The 1990 estimates of the undercount were 4.4 percent for African-Americans, 4.5 percent for American Indians, 5.0 percent for Hispanics, and 2.3 percent for Asians.)

Table 3 shows the RMSE for minority communities for the sizes 2,500, 10,000 and 50,000. Though the RMSE's for the adjusted census stay the same, since the bias has been removed, the unadjusted RMSE's are considerably larger. The gains in accuracy from an adjustment are much larger in minority communities, as one would expect. For example, as shown above, the error in the number of males in a non-minority community of 10,000 would be about 109 unadjusted and 43 adjusted. In a minority community, the errors are 205 and 43 respectively. In a larger area of 50,000 the improvement is dramatic even for a small cell of 1,000.

Now, suppose we repeal the 1976 law that specifies that there shall be no sampling for the apportionment numbers. Think about a census in which, after a certain date, the housing units not returning census forms are sampled.

Table 2
Expected Root Mean Square Error (RMSE) of Estimated Cell Frequencies for Population Estimates Based on a Census with No Adjustment for Undercount and with Adjustment

Area of 2,500 Population having RMSE based on			Area of 10,000 Population having RMSE based on			Area of 50,000 Population having RMSE based on		
Cell Frequency	Unadjusted Census	Adjusted Census	Cell Frequency	Unadjusted Census	Adjusted Census	Cell Frequency	Unadjusted Census	Adjusted Census
15	3	3	50	6	6	250	15	14
50	6	6	100	9	9	500	22	19
100	9	8	200	13	12	1,000	34	27
500	20	17	500	21	19	2,500	65	42
750	25	20	1,000	33	26	5,000	116	58
1,000	29	21	2,000	53	35	10,000	214	77
1,500	37	21	5,000	109	43	25,000	509	97

Note: Computations assume a relative response bias of 2 percent in the unadjusted census and 0 percent in the adjusted census. There is a response variance in both the adjusted and unadjusted census equal to 1/4 the sampling variance of a 25 percent sample.

Table 3
Expected Root Mean Square Error (RMSE) of Estimated Cell Frequencies for Population Estimates in African-American, American Indian, and Hispanic Communities Based on a Census with No Adjustment for Undercount and with Adjustment

Area of 2,500 Population having RMSE based on			Area of 10,000 Population having RMSE based on			Area of 50,000 Population having RMSE based on		
Cell Frequency	Unadjusted Census	Adjusted Census	Cell Frequency	Unadjusted Census	Adjusted Census	Cell Frequency	Unadjusted Census	Adjusted Census
15	3	3	50	6	6	250	17	14
50	6	6	100	9	9	500	28	19
100	9	8	200	15	12	1,000	48	27
500	26	17	500	21	19	2,500	109	42
750	31	20	1,000	48	26	5,000	208	58
1,000	45	21	2,000	87	35	10,000	407	77
1,500	64	21	5,000	205	43	25,000	1,004	97

Note: Computations assume a relative response bias of 4 percent in the adjusted census and 0 percent in the adjusted census. The response variance in both the adjusted and unadjusted census equal to 1/4 the sampling variance of a 25 percent sample.

In this model, there are two components of variance, the response variance and the sampling variance. The sampling variance is based only on the nonresponse universe.

Let R be the nonresponse rate, and M the population of nonresponse households. Then $M = RN$. The total for which we are trying to estimate the sampling variance is $S = PM$. The relationship between S, the sampled part of the total, and T, the total, is through R. $S = PM = P(RN) = RT$.

So the sampling variance = $3MPQ = 3PQRN$, assuming a 25-percent sample of the nonrespondents. This sampling rate could easily be changed for a larger rate, but for purposes of illustration, it suffices.

In Table 4, there are three contributors to the RMSE. Two of them are the terms we saw in the earlier description when sampling of the non-mail returns was not a consideration. Now we have a third term, expressing the sampling variance arising from the sample of non-mail returns. In an adjustment, only the bias term goes to zero, while the two variance terms remain. Each of the variance terms gets smaller as the cell size gets larger, but they do not vanish.

Table 4 shows the RMSE's for a census with no sampling of non-mail return households, with and without adjustment, for a 25 percent sample of non-mail return households when only half of the population mails them back and when 70 percent mail them back for the three sizes of area we have looked at before: 2,500 population, 10,000 population, and 50,000. The no sampling case is what we will have in the 2000 Census because the use of sampling for follow-up is prohibited. Look first at Section A for a population of 2,500. Where there is no sampling of non-mail return households, we see the numbers from Table 2. When half of the population mails back the census form, and the remaining half is sampled, the variance component keeps the adjusted and unadjusted RMSE's very close together. At maximum, there is a 20 percent reduction in error. There is somewhat more gain when the mailback rate is .70 and only 30 percent of the remaining population is sampled. The maximum gain in this case is 28 percent.

Small areas, such as those of 2,500 may be greatly affected by sampling, especially at a 25-percent rate if the mailback rate is low. Whether a decrease in accuracy is acceptable depends on the uses for the data. Since providing small area data is an important objective of the census, it may be that there would need to be a much larger sampling rate, if not complete follow-up for small areas. The Census Bureau has done this before with some characteristics, such as income, so that there would be less variability in the income data for areas of 2,500 or fewer persons. Following that same principle, it could be specified that there would be no sample follow-up in places of 2,500 persons or fewer, and variable follow-up rates depending on place size. Another strategy would be to use the information abundantly available about coverage error and to specify larger samples in places that have characteristics highly correlated with the undercount.

For areas of 10,000 population, we see a definite improvement from the adjustment for the bias, but the adjusted numbers with sampling are still considerably larger than the adjusted figures without sampling. However, if there is no adjustment, the sampling adds to the RMSE, but the unadjusted numbers are not much different. There is a 15 percent increase in the RMSE when only half the population returns the census form and an increase of 9 percent when 70 percent return it.

Finally, when we look at an area of 50,000 we see that the bias dominates the RMSE for all but the smallest cell sizes. When the total we are trying to estimate is 5,000 or larger, sampling adds to the RMSE, but an adjustment, with sampling, is still superior to unadjusted numbers with no sampling.

Table 5 is similar, but geared to a predominantly minority population. As in Table 3, the relative bias is 4 percent, reflecting an average undercount rate for minorities. In this table, the RMSE's for unadjusted totals are much more similar, even for smaller areas, because of the larger effect of the bias term on the RMSE.

The results for areas of 50,000, which exist in most large cities, show the devastating effects of not adjusting for the large minority undercount. The sampling variance for the larger totals has practically no effect on the RMSE, but the improvement from adjustment for all cases, sampling or no sampling is 83 percent or higher. The added error because of sampling is negligible.

Unfortunately, in many minority communities, low mail-return rates and undercounting occur together. Such communities have a 50 percent mail return rate or lower. It may be that the sample size will need to be increased in these areas.

Table 4

Expected Root Mean Square Error (RMSE) of Estimated Cell Frequencies for Population Estimates Based on an Unadjusted Census, an Adjusted Census, and on a 25 Percent Sample of Non-Mail Return Households

A. Area of 2,500 population having RMSE based on

	No sampling of non-mail return HH's		25% sample and .50 mailback rate		25% sample, and .70 mailback rate	
Cell Frequency	Unadjusted	Adjusted	Unadjusted	Adjusted	Unadjusted	Adjusted
15	3	3	6	6	5	5
50	6	6	11	11	9	9
100	9	8	15	15	13	13
500	20	17	32	30	28	26
750	25	20	38	34	33	29
1,000	29	21	42	37	37	31
1,500	37	21	47	37	43	31

B. Area of 10,000 population having RMSE based on

	No sampling of non-mail return HH's		25% sample and .50 mailback rate		25% sample, and .70 mailback rate	
Cell Frequency	Unadjusted	Adjusted	Unadjusted	Adjusted	Unadjusted	Adjusted
50	6	6	11	11	9	9
100	9	9	15	15	13	13
200	13	12	21	21	18	18
500	21	19	34	33	30	28
1,000	33	26	49	45	43	39
2,000	53	35	72	60	65	51
5,000	109	43	125	75	119	64

C. Area of 50,000 population having RMSE based on

	No sampling of non-mail return HH's		25% sample and .50 mailback rate		25% sample, and .70 mailback rate	
Cell Frequency	Unadjusted	Adjusted	Unadjusted	Adjusted	Unadjusted	Adjusted
250	15	14	24	24	21	20
500	22	19	35	33	30	29
1,000	34	27	51	47	45	40
2,500	65	42	89	73	80	63
5,000	116	58	142	101	132	86
10,000	214	77	241	134	231	115
25,000	509	97	527	168	520	144

Table 5
Expected Root Mean Square Error (RMSE) of Estimated Cell Frequencies for Population Estimates in African-American, American Indian, and Hispanic Communities Based on an Unadjusted Census, an Adjusted Census, and on a 25 Percent Sample of Non-Mail Return Households

A. Area of 2,500 population having RMSE based on

	No sampling of non-mail return HH's		25% sample, and .50 mailback rate		25% sample, and .70 mailback rate	
Cell Frequency	Unadjusted	Adjusted	Unadjusted	Adjusted	Unadjusted	Adjusted
15	3	3	6	6	5	5
50	6	6	11	11	9	9
100	9	8	15	15	13	13
500	26	17	36	30	33	26
750	31	20	46	34	42	29
1,000	45	21	54	37	51	31
1,500	64	21	70	37	68	31

B. Area of 10,000 population having RMSE based on

	No sampling of non-mail return HH's		25% sample, and .50 mailback rate		25% sample, and .70 mailback rate	
Cell Frequency	Unadjusted	Adjusted	Unadjusted	Adjusted	Unadjusted	Adjusted
50	6	6	11	11	9	9
100	9	9	15	15	13	13
200	15	12	22	21	18	18
500	21	19	38	33	34	28
1,000	48	26	60	45	56	39
2,000	87	35	100	60	95	51
5,000	205	43	214	75	210	64

C. Area of 50,000 population having RMSE based on

	No sampling of non-mail return HH's		25% sample, and .50 mailback rate		25% sample, and .70 mailback rate	
Cell Frequency	Unadjusted	Adjusted	Unadjusted	Adjusted	Unadjusted	Adjusted
250	17	14	26	23	23	21
500	28	19	39	33	35	29
1,000	48	27	62	47	57	40
2,500	109	42	124	73	118	63
5,000	208	58	224	101	218	86
10,000	407	77	422	134	416	115
25,000	1,004	97	1,014	168	1,010	144

6. CONCLUSION

It has been a tradition for the Census Bureau in the latter half of this century to use statistical techniques, where possible, to make the Decennial Census more accurate and less costly. Using the techniques historically used by the Census Bureau, namely a mean-square error model, one can see that adjustment does improve census totals, even for small areas, when one assumes even a minimal level of response variance. One can also see the need for precaution if sampling is to be used for follow-up. It may be that there should be no sampling in places of 2,500 or fewer people, just as there is no sampling for certain population characteristics in these small places.

In looking at the current census controversy, it is good to remember the spirit of Mahalanobis. Not only did his ingenious use of interpenetrated subsamples give us the ability to estimate the response variance in census statistics, but his insistence that sampling and statistics should be used to solve practical problems has been the hallmark of the U.S. Census. Some of the most fundamental practical problems are those faced by the government and Mahalanobis allocated statistical resources for the solving of these problems. Likewise, the U.S. Census Bureau has a long and rich history of offering practical, cost-efficient solutions to thorny census problems.

REFERENCES

BAILAR, B.A. (1969). Evaluation and Research Program of the U.S. Censuses of Population and Housing, 1960: The Effect of Interviewers and Crew Leaders. Series ER 60 No. 7. Washington, DC: U.S. Bureau of the Census.

EDMONSTON, B., and SCHULTZE, C. (1993). *Modernizing the U.S. Census.* Washington, DC: National Academy Press, 34-35.

HANSON, R.H., and MARKS, E.S. (1958). The influence of the interviewer on the accuracy of survey results. *Journal of the American Statistical Association,* 53, 639-655.

MAHALANOBIS, P.C. (1946). Recent experiments in statistical sampling in the Indian Statistical Institute. *Journal of the Royal Statistical Society,* 325-378.

MAHALANOBIS, P.C. (1950). Why Statistics? *Sankhyā,* 10, 195-228.

MAHALANOBIS, P.C., and LAHIRI, D.B. (1961). Analysis of errors in censuses and surveys with special reference to experience in India. *Bulletin of the International Statistical Institute,* 38, 2, 401-433 (reprinted in *Sankhyā,* 23, 325-358).

MICHAEL, R.T., GAGNON, J.H., LAUMANN, E.O., and KOLATA, G. (1994). *Sex in America, A Definitive Survey.* New York: Little Brown and Co.

RUDRA, A. (1996). *Prasanta Chandra Mahalanobis: A Biography.* New York: Oxford University Press.

WAKSBERG, J., and HANSON, R. (1965). Sampling Applications in Censuses of Population and Housing. U.S. Bureau of the Census, Technical Paper No. 13.

Survey Methodology, June 2000
Vol. 26, No. 1, pp. 31-42
Statistics Canada

Estimation of Census Adjustment Factors

C.T. ISAKI, J.H. TSAY and W.A. FULLER[1]

ABSTRACT

A components-of-variance approach and an estimated covariance error structure were used in constructing predictors of adjustment factors for the 1990 Decennial Census. The variability of the estimated covariance matrix is the suspected cause of certain anomalies that appeared in the regression estimation and in the estimated adjustment factors. We investigate alternative prediction methods and propose a procedure that is less influenced by variability in the estimated covariance matrix. The proposed methodology is applied to a data set composed of 336 adjustment factors from the 1990 Post Enumeration Survey.

KEY WORDS: Components-of-variance; Small area estimation; Undercount; Decennial Census; Smoothing.

1. INTRODUCTION

While the objective of a population census is to record data for all individuals, it has long been recognized that this goal is not achieved in practice. Post enumeration studies associated with the U.S. Census of 1970 and 1980 suggested that the coverage rate was different for different demographic groups. See U.S. Bureau of the Census (1988).

In 1990, a post enumeration survey (PES), using dual system (or capture-recapture) estimation, was used to produce estimates for 1392 subdivisions of the total population of the United States at the time of the 1990 Census. The PES sample contained approximately 377,000 persons in about 5200 sample blocks. Sample persons were divided into post-strata defined by geographic divisions of the country, tenure, size-of-place, race, sex, and age, where the two tenure classes are owners and renters of homes, and size-of-place is a measure of urbanization. The subdivisions were called poststrata. The ratio of the PES estimate to the Census total, called the adjustment factor, was produced for each poststratum. An adjustment factor greater than one is associated with an estimated undercount and a factor less than one is associated with an estimated overcount.

Because relatively large sampling variances were anticipated for individual ratios, a smoothing technique based on components-of-variance and a regression model was used to create the final estimated adjustment factors. The elements of the error covariance matrix used in the prediction model were estimated with a jackknife algorithm, see Fay (1990).

The explanatory variables in the regression model were chosen using a best subsets selection algorithm. Some explanatory variables were forced into the model. For example, in the Midwest region, the ten explanatory variables forced into the model were Black, Hispanic, renter, age group 0-9, age group 10-19, age group 20-29, age group 30-44, age group 45-64, male 10-19 and male 20-64. Most variables were indicator variables, but some were proportions. For example, a variable "percent Black" was used when Black and Hispanic were grouped into a single poststratum. Nine other variables were selected for inclusion in the model based on a best subsets regression algorithm. The variables included mail return rate, substitution rate, type-of-place and six race-by-age and race-by-tenure interaction variables. The mail return rate is the fraction of Census questionnaires returned from the mail distribution, the substitution rate is the fraction of Census households that were entirely replaced with responding households.

The smoothing technique was applied to poststrata ratios by regions of the country. The adjustment factors were designed to be applied to Census counts in the appropriate poststrata to create population estimates adjusted for undercount or overcount. Hogan (1992) contains an overview of the PES. Isaki, Huang and Tsay (1991) provide a detailed description of the results of the smoothing of the poststratum ratios.

Fay (1992) in a manuscript discussing the adjustment factors constructed from the 1990 PES, identified some disturbing results. He noted that some of the estimated regression coefficients in the model differed considerably depending on the form of the estimated covariance matrix used to construct the estimated generalized least squares estimator. Fay conjectured that large differences in coefficients could arise because of an unstable estimator of the error covariance matrix. Although the estimated error variances were smoothed, it was felt that estimated variances of linear combinations might still have large variances. He felt that the estimated variances had large variances because the direct estimates for many blocks were zero.

The Secretary of Commerce ultimately decided to use the unadjusted counts in the Decennial Census. The possible use of adjusted counts for other purposes, such as the Bureau's postcensal estimation program, was left for additional study.

[1] C.T. Isaki and J.H. Tsay, Statistical Research Division, Bureau of the Census, Washington, D.C. 20233, U.S.A; W.A. Fuller, Department of Statistics, Iowa State University, Ames, IA 50010, U.S.A.

We explore alternative smoothed estimators for the adjustment factors, focusing on the effect of estimating the covariance matrix of the vector of the estimated adjustment factors. In the empirical part of our study, we construct estimates based on the 1990 Census data.

2. SMOOTHING MODEL

The model chosen for the construction of predictors is the multivariate components-of-variance model. Closely related models that lead to smoothed estimators for a set of unknowns, have been studied by a number of authors. Fay and Herriot (1979) suggested the use of the model in a small area estimation procedure. Battese, Harter and Fuller (1988) applied the components-of-variance model to crop area estimation. Ericksen and Kadane (1985), Cressie (1992), and Ericksen, Kadane and Tukey (1989) suggested smoothing procedures for census adjustment. Singh, Gambino and Mantel (1994) discuss a range of small area procedures. Efron and Morris (1972) and Morris (1983) contain good discussions of some of the basic theory. Kackar and Harville (1984), Peixoto and Harville (1986), Fay (1987), Fuller and Harter (1987), Hulting and Harville (1991), Ghosh (1992), and Prasad and Rao (1990) discuss estimation and variance estimation for such procedures. Ghosh and Rao (1994) is a review article.

Under the multivariate components-of-variance model, the vector of true values to be predicted is

$$\mathbf{y} = \mathbf{X}\beta + \mathbf{w}, \tag{1}$$

where $\mathbf{y}$ is an n-dimensional column vector, $\mathbf{X}$ is an $n \times k$ matrix of observable characteristics, $\mathbf{w}$ is an n-dimensional column vector of random effects and β is a k-dimensional unknown column vector. The vector $\mathbf{Y}$ is observed, where

$$\mathbf{Y} = \mathbf{y} + \mathbf{e}, \tag{2}$$

$\mathbf{Y}$ is an n-dimensional column vector and $\mathbf{e}$ is the n-dimensional column vector of estimation errors. In our application $\mathbf{Y}$ is the vector of estimated adjustment factors. It is assumed that

$$(\mathbf{w}', \mathbf{e}')' \sim \left(\mathbf{0}, \text{block diag}\left\{\mathbf{I}\sigma^2, \Sigma_{ee}\right\}\right), \tag{3}$$

where Σ_{ee} is the covariance matrix of the estimation errors, and σ^2 is the unknown variance of the random effects.

A class of predictors of $\mathbf{y}$ is defined by

$$\tilde{\mathbf{y}} = \mathbf{XB} + \mathbf{G}'(\mathbf{Y} - \mathbf{XB}), \tag{4}$$

where $\mathbf{B}$ is a k-dimensional vector and $\mathbf{G}$ is an $n \times n$ matrix. Under model (1) with

$$(\mathbf{w}', \mathbf{e}')' \sim N\left(\mathbf{0}, \text{block diag}\left\{\mathbf{I}\sigma^2, \Sigma_{ee}\right\}\right), \tag{5}$$

the conditional expected value of $\mathbf{y}$ given $\mathbf{Y}$ is

$$E\{y \mid Y\} = \mathbf{X}\beta + \mathbf{G}_{zz}'(\mathbf{Y} - \mathbf{X}\beta), \tag{6}$$

where $\mathbf{G}_{zz} = \Sigma_{zz}^{-1}\sigma^2$ and $\sum_{zz} = \mathbf{I}\sigma^2 + \Sigma_{ee}$ is the $n \times n$ covariance matrix of $\mathbf{z} = \mathbf{w} + \mathbf{e}$. Under the normal distribution model defined by (1), (2), and (5) and with the parameters σ^2, Σ_{ee}, β known, the minimum mean square error predictor of $\mathbf{y}$ is given by the right side of equation (6).

Generally, some of the parameters are unknown. Consider first the case in which β is unknown. Let $\hat{\beta}$ be an estimator of β, where

$$\hat{\beta} = (\mathbf{X}'\mathbf{M}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{M}^{-1}\mathbf{Y}, \tag{7}$$

and $\mathbf{M}$ is an $n \times n$ matrix. If $\mathbf{M}$ is fixed

$$\begin{aligned}\tilde{y} - y &= (\mathbf{I} - \mathbf{G})\mathbf{X}(\hat{\beta} - \beta) - (\mathbf{I} - \mathbf{G})\mathbf{w} + \mathbf{Ge} \\ &= (\mathbf{K} - \mathbf{I})\mathbf{w} + \mathbf{Ke},\end{aligned}$$

where $\mathbf{K} = (\mathbf{I} - \mathbf{G}')\mathbf{X}(\mathbf{X}'\mathbf{M}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{M}^{-1} + \mathbf{G}'$. Thus, if $\mathbf{M}$ and $\mathbf{G}$ are fixed,

$$V\{\tilde{y} - y\} = (\mathbf{K} - \mathbf{I})(\mathbf{K} - \mathbf{I})'\sigma^2 + \mathbf{K}\Sigma_{ee}\mathbf{K}'. \tag{8}$$

If model (1), (2), and (3) holds, and if $\sum_{ee}$ and σ^2 are known, then replacing $\mathbf{B}$ with

$$\tilde{\beta} = \left(\mathbf{X}'\Sigma_{zz}^{-1}\mathbf{X}\right)^{-1}\mathbf{X}'\Sigma_{zz}^{-1}\mathbf{Y} \tag{9}$$

and replacing $\mathbf{G}$ with

$$\mathbf{G}_{zz} = \Sigma_{zz}^{-1}\sigma^2 \tag{10}$$

in (4) defines the best linear unbiased predictor of $\mathbf{y}$. See Henderson (1950), Harville (1976), and Robinson (1991). If Σ_{ee} and σ^2 are also unknown, it is natural to use estimators of Σ_{ee} and σ^2 to construct an estimated best linear unbiased predictor. Very often, an estimator of Σ_{ee} is associated with the procedure used to construct the estimator $\mathbf{Y}$. Then σ^2 is estimated from model (1), (2), and (5), treating the estimator of Σ_{ee} as the true Σ_{ee}.

One substitution predictor is

$$\hat{\mathbf{y}} = X\hat{\beta} + \hat{\sigma}^2\hat{\Sigma}_{zz}^{-1}(\mathbf{Y} - \mathbf{X}\hat{\beta}), \tag{11}$$

where

$$\hat{\beta} = \left(\mathbf{X}'\hat{\Sigma}_{zz}^{-1}\mathbf{X}\right)^{-1}\mathbf{X}'\hat{\Sigma}_{zz}^{-1}\mathbf{Y} \tag{12}$$

is the estimated generalized least squares estimator of β,

$$\hat{\Sigma}_{zz} = \mathbf{I}\hat{\sigma}^2 + \hat{\Sigma}_{ee} \tag{13}$$

$\hat{\Sigma}_{ee}$ is an estimator of Σ_{ee}, and $\hat{\sigma}^2$ is an estimator of σ^2. The estimator of σ^2 can be based on likelihood or analysis of variance procedures. Retaining only the terms in the Taylor expansion of the error in (11) that are errors in the basic estimators, we have

$$\begin{aligned}\hat{\mathbf{y}} - \mathbf{y} \doteq\ & \mathbf{e} - \mathbf{H}'\mathbf{z} + \mathbf{H}'\mathbf{X}(\hat{\beta} - \beta) \\ & + (\hat{\sigma}^2 - \sigma^2)\mathbf{H}'\Sigma_{zz}^{-1}\mathbf{z} \\ & - \mathbf{G}'\left(\hat{\Sigma}_{ee} - \Sigma_{ee}\right)\Sigma_{zz}^{-1}\mathbf{z},\end{aligned} \tag{14}$$

where $\mathbf{H}' = \boldsymbol{\Sigma}_{ze}\boldsymbol{\Sigma}_{zz}^{-1}$ and $\mathbf{G}' = \mathbf{I} - \mathbf{H}' = \sigma^2\boldsymbol{\Sigma}_{zz}^{-1}$. If it is assumed that $\hat{\boldsymbol{\Sigma}}_{ee}$ is distributed as a multiple of a Wishart matrix with d_e degrees of freedom, if the covariance between $\hat{\sigma}^2$ and $\hat{\boldsymbol{\Sigma}}_{ee}$ is ignored, if expectations are computed as if $\hat{\sigma}^2$ and $\mathbf{z}$ are independent, and if expectations are computed as if $\mathbf{z}$ and $\hat{\boldsymbol{\Sigma}}_{ee}$ are independent, an approximation to the variance of $\hat{\mathbf{y}} - \mathbf{y}$ obtained from (14) is

$$\mathbf{V}\{\hat{\mathbf{y}} - \mathbf{y}\} \doteq \boldsymbol{\Sigma}_{ee}\mathbf{G} + \mathbf{H}'\mathbf{X}\mathbf{V}_{\beta\beta}\mathbf{X}'\mathbf{H} + \boldsymbol{\Gamma}_{33} + \boldsymbol{\Gamma}_{44}, \quad (15)$$

where

$$\mathbf{V}_{\beta\beta} = \mathbf{V}\{\hat{\boldsymbol{\beta}}\} = \left(\mathbf{X}'\boldsymbol{\Sigma}_{zz}^{-1}\mathbf{X}\right)^{-1} + d_e^{-1}\,\mathrm{tr}\left\{\boldsymbol{\Sigma}_{zz}^{-1}\boldsymbol{\Sigma}_{ee}\right\}\mathbf{L}\boldsymbol{\Sigma}_{ee}\mathbf{L}',$$

$$\boldsymbol{\Gamma}_{33} = \mathbf{H}'\boldsymbol{\Sigma}_{zz}^{-1}\mathbf{H}V_{\sigma\sigma},$$

$$\boldsymbol{\Gamma}_{44} = d_e^{-1}\sigma^4\boldsymbol{\Sigma}_{zz}^{-1}\boldsymbol{\Sigma}_{ee}\boldsymbol{\Sigma}_{zz}^{-1}\left[\mathrm{tr}\left\{\boldsymbol{\Sigma}_{zz}^{-1}\boldsymbol{\Sigma}_{ee}\right\}\right],$$

$$\mathbf{L} = (\mathbf{X}'\boldsymbol{\Sigma}_{zz}^{-1}\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\Sigma}_{zz}^{-1}$$

and $V_{\sigma\sigma} = V\{\hat{\sigma}^2\}$ is the variance of $\hat{\sigma}^2$. The term $\boldsymbol{\Sigma}_{ee}\mathbf{G}$ is the prediction covariance matrix if all parameters are known. The remaining three terms of (15) are the contributions to the variance due to estimating $\boldsymbol{\beta}$, σ^2, and $\boldsymbol{\Sigma}_{ee}$, respectively. The second term in $\mathbf{V}\{\hat{\boldsymbol{\beta}}\}$ is a crude approximation for the increase in the variance of $\hat{\boldsymbol{\beta}}$ due to using an estimator of $\boldsymbol{\Sigma}_{zz}$ in place of $\boldsymbol{\Sigma}_{zz}$ in constructing $\hat{\boldsymbol{\beta}}$.

If the dimension of $\boldsymbol{\Sigma}_{zz}$ is large and the degrees of freedom, d_e, only slightly larger than the dimension, then the second part of the variance of $\hat{\boldsymbol{\beta}}$ and the term $\boldsymbol{\Gamma}_{44}$ can make important contributions to the variance. This is particularly true if σ^2 is small relative to the diagonal elements of $\boldsymbol{\Sigma}_{ee}$. The Monte Carlo study of the next section demonstrates that the contribution to variance approximated by these terms can be important.

A predictor that reduces the effect of the estimation error in $\hat{\boldsymbol{\Sigma}}_{ee}$ uses only diagonal elements of $\boldsymbol{\Sigma}_{ee}$ in the shrinkage component. Let

$$\hat{\mathbf{y}}_d = \mathbf{X}\hat{\boldsymbol{\beta}}_d + \hat{\sigma}^2\hat{\mathbf{D}}_{zz}^{-1}(\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}_d), \quad (16)$$

where

$$\hat{\boldsymbol{\beta}}_d = (\mathbf{X}'\hat{\mathbf{D}}_{zz}^{-1}\mathbf{X})^{-1}\mathbf{X}'\hat{\mathbf{D}}_{zz}^{-1}Y,$$

$$\hat{\mathbf{D}}_{zz} = \mathrm{diag}\left(\hat{\boldsymbol{\Sigma}}_{ee} + \mathbf{I}\hat{\sigma}^2\right),$$

$\hat{\sigma}^2$ is an estimator of σ^2 and $\mathrm{diag}(\mathbf{A})$ is the diagonal matrix composed of the diagonal elements of $\mathbf{A}$. Retaining only the leading terms in the Taylor expansion of the error in (16) gives

$$\hat{\mathbf{y}}_d - \mathbf{y} \doteq -(\mathbf{w} - \mathbf{G}_d'\mathbf{z}) + \mathbf{H}_d'\mathbf{X}(\hat{\boldsymbol{\beta}}_d - \boldsymbol{\beta}) + (\hat{\sigma}^2 - \sigma^2)\mathbf{H}_d'\mathbf{D}_{zz}^{-1}\mathbf{z} - \mathbf{G}_d'(\hat{\mathbf{D}}_{ee} - \mathbf{D}_{ee})\mathbf{D}_{zz}^{-1}\mathbf{z}, \quad (17)$$

where $\mathbf{D}_{zz} = \mathrm{diag}\{\boldsymbol{\Sigma}_{zz}\}$, $\mathbf{G}_d = \mathbf{D}_{zz}^{-1}\sigma^2$, $\mathbf{H}_d = \mathbf{I} - \mathbf{G}_d$, and $\mathbf{D}_{ee} = \mathrm{diag}\{\boldsymbol{\Sigma}_{ee}\}$. If $\mathbf{w}$ and $\mathbf{e}$ are normally distributed, and if $\hat{\sigma}^2$ and $\hat{\mathbf{D}}_{zz}$ are quadratic estimators, then $\hat{\sigma}^2$ and $\hat{\mathbf{D}}_{zz}$ are uncorrelated with $\mathbf{z}$. The i-th element of $\mathbf{w} - \sigma^2\mathbf{D}_{zz}^{-1}\mathbf{z}$ is uncorrelated with the i-th element of $\mathbf{z}$, but is not necessarily uncorrelated with the vector $\mathbf{z}$. If this possible correlation is ignored, if it is assumed that $\hat{\boldsymbol{\Sigma}}_{ee}$ is a Wishart matrix with d_e degrees of freedom, and if the correlation between $\hat{\sigma}^2$ and $\hat{\boldsymbol{\Sigma}}_{ee}$ is ignored, an approximation to the variance of $\hat{\mathbf{y}}_d - \mathbf{y}$ obtained from (17) is

$$\mathbf{V}\{\hat{\mathbf{y}}_d - \mathbf{y}\} = \mathbf{H}_d'\mathbf{H}_d\sigma^2 + \mathbf{G}_d'\boldsymbol{\Sigma}_{ee}\mathbf{G}_d + \mathbf{H}_d'\mathbf{X}\mathbf{V}_{\beta\beta}\mathbf{X}'\mathbf{H}_d + \boldsymbol{\Gamma}_{33dd} + \boldsymbol{\Gamma}_{44dd}, \quad (18)$$

where $\mathbf{G}_d = \mathbf{D}_{zz}^{-1}\sigma^2$, $\mathbf{H}_d = \mathbf{I} - \mathbf{G}_d$,

$$\mathbf{V}_{\beta\beta} = (\mathbf{X}'\mathbf{D}_{zz}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{D}_{zz}^{-1}\boldsymbol{\Sigma}_{zz}\mathbf{D}_{zz}^{-1}\mathbf{X}(\mathbf{X}'\mathbf{D}_{zz}^{-1}\mathbf{X})^{-1}, \quad (19)$$

$$\boldsymbol{\Gamma}_{33dd} = \mathbf{H}_d'\mathbf{D}_{zz}^{-1}\boldsymbol{\Sigma}_{zz}\mathbf{D}_{zz}^{-1}\mathbf{H}_d V_{\sigma\sigma}, \quad (20)$$

$$\boldsymbol{\Gamma}_{44dd} = d_e^{-1}\mathbf{G}_d'\boldsymbol{\Omega}\mathbf{G}_d'$$

and the ij-th element of $\boldsymbol{\Omega}$ is

$$\omega_{ij} = 2\sigma_{eeij}^2\sigma_{zzii}^{-1}\sigma_{zzjj}^{-1}\sigma_{zzij}.$$

The term in $\boldsymbol{\Gamma}_{44dd}$ is an estimator of the contribution to the variance due to using $\hat{\boldsymbol{\Sigma}}_{ee}$ to estimate the covariance matrix. Expression (19) assumes that the contribution of the error in $\hat{\mathbf{D}}_{zz}$ to the variance of $\hat{\boldsymbol{\beta}}$ can be ignored for large d_e. The difference between (15) and (18) is that the multipliers in (19) and (20) do not depend on the dimension of $\boldsymbol{\Sigma}_{zz}$. Therefore, the error in estimating $\boldsymbol{\Sigma}_{zz}$ makes a smaller contribution to the variance. On the other hand, the variance of $\mathbf{w} - \mathbf{G}_d'\mathbf{z}$, the order one term of (17), will be larger than the corresponding term of the error in (14), unless $\boldsymbol{\Sigma}_{zz}$ is diagonal. The first two terms on the right of (18) are the variance of $\mathbf{w} - \mathbf{G}_d'\mathbf{z}$.

3. MONTE CARLO STUDY

To examine the variability in the predictors associated with variability in the estimation of $\boldsymbol{\Sigma}_{ee}$ we conducted a small Monte Carlo study. The model for the study is

$$\mathbf{Y}_j = \mu\mathbf{J} + \mathbf{w} + \mathbf{e}_j, \qquad j = 1, 2, \ldots, r \quad (21)$$

$$\mathbf{w} \sim (\mathbf{0}, \mathbf{I}\sigma^2),$$

$$\mathbf{e}_j \sim \mathrm{ind}\left(\mathbf{0}, \boldsymbol{\Sigma}_{ee}\right),$$

where $\mathbf{J}$ is the k-dimensional column vector of ones, $\mathbf{J} = (1, 1, \ldots, 1)'$, $\mathbf{w}$ is the k-dimensional vector of random small area effects, $\mathbf{e}_j$ is a vector of errors, and $\mathbf{w}$ and $\mathbf{e}_j$ are independent. The model is a simplified version of the model defined in (1), (2), and (3). The mean is the constant function and, hence, we use μ in place of β. To create a vector of correlated variables, we define, for $k = 8$,

$$\begin{bmatrix} e_{1j} \\ e_{2j} \\ e_{3j} \\ e_{4j} \\ e_{5j} \\ e_{6j} \\ e_{7j} \\ e_{8j} \end{bmatrix} = \begin{bmatrix} 1.3u_{1j} \\ 1.5u_{1j} + 0.4u_{2j} \\ 0.9u_{1j} + 0.9u_{3j} \\ 0.9u_{3j} + 1.6u_{4j} \\ 1.6u_{4j} + 0.6u_{5j} \\ 1.0u_{4j} + 1.6u_{6j} \\ 1.0u_{7j} \\ 2.83u_{8j} \end{bmatrix},$$

where u_{ij} are independent random variables. The w_i, $i = 1, 2, ..., 8$, are NI(0, 0.36) random variables, where NI(μ, σ^2) denotes normal independent random variables with mean μ and variance σ^2. This configuration gives a range of error variances and a range of correlations between estimates.

The estimator of σ^2 used in the Monte Carlo study is

$$\hat{\sigma}^2 = \max\left\{(k-1)^{-1} \times\left[\left(\bar{\mathbf{y}} - \mathbf{J}\hat{\mu}_{(0)}\right)'\left(\bar{\mathbf{y}} - \mathbf{J}\hat{\mu}_{(0)}\right) - \operatorname{tr}\left\{r^{-1}\hat{\boldsymbol{\Sigma}}_{ee}\mathbf{A}_0\right\}\right], 0\right\} \quad (22)$$

where tr{**A**} is the trace of the matrix **A**,

$$\mathbf{A}_0 = \mathbf{I} - k^{-1}\mathbf{J}\mathbf{J}'$$

$$\hat{\boldsymbol{\Sigma}}_{ee} = (r-1)^{-1}\sum_{j=1}^{r}(\mathbf{Y}_j - \bar{\mathbf{y}})(\mathbf{Y}_j - \bar{\mathbf{y}})', \quad (23)$$

and

$$\hat{\mu}_0 = k^{-1}\mathbf{J}'\bar{\mathbf{y}}. \quad (24)$$

The estimator $\hat{\sigma}^2$ is a quadratic estimator closely related to the analysis of variance estimator.

Two predictors were compared in the Monte Carlo study. Both are of the form

$$\hat{\mathbf{y}} = \bar{\mathbf{y}} - \hat{\mathbf{H}}'(\bar{\mathbf{y}} - \hat{\mu}\mathbf{J}), \quad (25)$$

where

$$\bar{\mathbf{y}} = r^{-1}\sum_{j=1}^{r} Y_j.$$

They differ in the construction of $\hat{\mathbf{H}}$ and $\hat{\mu}$. The first predictor is of the form (11) and uses the full estimated $\hat{\Sigma}_{ee}$ in $\hat{\mathbf{H}}$ and in the estimator of μ. The predictor is called the general predictor as an abbreviation for estimated generalized least squares predictor. The general predictor is

$$\hat{\mathbf{y}}_g = \bar{\mathbf{y}} - \hat{\mathbf{H}}_g'(\bar{\mathbf{y}} - \hat{\mu}_g\mathbf{J}), \quad (26)$$

where

$$\hat{\mathbf{H}}_g' = r^{-1}\hat{\boldsymbol{\Sigma}}_{ee}\,\hat{\boldsymbol{\Sigma}}_{zz}^{-1},$$

$$\hat{\mu}_g = \left(\mathbf{J}'\hat{\boldsymbol{\Sigma}}_{zz}^{-1}\mathbf{J}\right)^{-1}\mathbf{J}'\boldsymbol{\Sigma}_{zz}^{-1}\bar{\mathbf{y}}, \quad (27)$$

$$\hat{\boldsymbol{\Sigma}}_{zz} = r^{-1}\hat{\boldsymbol{\Sigma}}_{ee} + \mathbf{I}\hat{\sigma}^2, \quad (28)$$

and $\hat{\mu}_g$ is the estimated generalized least squares estimator of μ.

The second predictor is

$$\hat{\mathbf{y}}_d = \bar{\mathbf{y}} - \hat{\mathbf{H}}_d'(\bar{\mathbf{y}} - \hat{\mu}_d\mathbf{J}), \quad (29)$$

where

$$\hat{\mathbf{H}}_d' = r^{-1}\mathbf{M}_{ee}\hat{\mathbf{D}}_{zz}^{-1},$$

$\mathbf{M}_{ee} = \operatorname{diag}\hat{\boldsymbol{\Sigma}}_{ee}$, $\hat{\mathbf{D}}_{zz} = \operatorname{diag}\hat{\boldsymbol{\Sigma}}_{zz}$, and the estimated μ is

$$\hat{\mu}_d = [\mathbf{J}'\hat{\mathbf{D}}_{zz}^{-1}\mathbf{J}]^{-1}\mathbf{J}'\hat{\mathbf{D}}_{zz}^{-1}\bar{\mathbf{y}}.$$

This predictor might be called the diagonal predictor because only the diagonal elements of $\hat{\boldsymbol{\Sigma}}_{ee}$ are used in the construction.

The entries in Table 1 are for $r = 14$. Each sample is composed of a random selection of **w** and a random sample of 14 **e**-vectors. Results are given for errors $u_{ij} \sim$ NI(0, 2) and errors that are centered one-degree-of-freedom chi-square random variables. Thus, in both cases the errors have zero means and variances equal to two. The mean μ was set equal to zero. The second column of Table 1 contains the variance of the sample mean as an estimator of the w_i. Column three of Table 1 contains the ratio of the Monte Carlo variance of an element of $\hat{\mathbf{y}}_g$, where $\hat{\mathbf{y}}_g$ is defined by (28), to the Monte Carlo variance of the corresponding element of $\bar{\mathbf{y}}$ for normal errors. The ratios for elements one through four and element 7 are greater than one. The last two elements of $\mathbf{Y}_j$ are uncorrelated with other elements. Element seven has a small variance and element eight has a large variance. There is a large loss for the predictor relative to the simple mean for element seven and a large gain for element eight.

The fourth column of Table 1 contains the ratios of the variance of the predictor of (29) to the variance of the mean for normal errors. In all cases the diagonal predictor is superior to the general predictor defined in (28). The difference is relatively constant at about 30%. The diagonal predictor is not always superior to the simple mean but the loss is small for elements one, three, and seven. On the other hand, the gains relative to the simple mean are large for elements six and eight. The Monte Carlo variances for both predictors are larger than the approximations associated with equations (15) and (18) except for element 8.

It is somewhat surprising that the diagonal procedure did better relative to the simple mean for chi-square errors than for normal errors. With the chi-square error, the estimated mean and estimated variance are correlated. Hence, on the average, the large positive mean deviations are pulled toward the mean by a larger amount than the smaller negative deviation. The Associate Editor conjectured, and we concur, that this is one reason for the superior performance of the diagonal predictor. On the other hand, the general

prediction procedure is poorer relative to the simple mean for chi-square errors than for normal errors. As the last column of Table 1 demonstrates, the diagonal predictor procedure uniformly dominates both the mean and the general prediction procedure for this parametric configuration with chi-square errors.

Table 1
Monte Carlo Variance Ratios for Alternative Small Area Predictors
(10,000 samples, $r = 14$)

		Normal Errors		Chi-square Errors	
i	$V\{\bar{y}_i - w_i\}$	$\frac{\hat{V}\{\hat{y}_{gi} - w_i\}}{\hat{V}\{\bar{y}_i - w_i\}}$	$\frac{\hat{V}\{\hat{y}_{di} - w_i\}}{\hat{V}\{\bar{y}_i - w_i\}}$	$\frac{\hat{V}\{\hat{y}_{gi} - w_i\}}{\hat{V}\{\bar{y}_i - w_i\}}$	$\frac{\hat{V}\{\hat{y}_{di} - w_i\}}{\hat{V}\{\bar{y}_i - w_i\}}$
1	0.2414	1.277	1.025	1.430	0.899
2	0.3445	1.252	0.875	1.371	0.768
3	0.2268	1.351	1.019	1.480	0.954
4	0.4771	1.003	0.735	1.099	0.686
5	0.4113	0.926	0.876	1.016	0.699
6	0.5121	0.913	0.677	0.975	0.618
7	0.1449	1.366	1.006	2.261	0.896
8	1.1214	0.520	0.384	0.725	0.371

The Monte Carlo variances of $\hat{\mu}_0$, $\hat{\mu}_g$, and $\hat{\mu}_d$ as estimators of μ are 0.150, 0.273, and 0.146 , respectively. If Σ_{ee} and σ^2 are known, the variances of $\hat{\mu}_0$, $\hat{\mu}_g$, and $\hat{\mu}_d$ are 0.149, 0.122, and 0.140, respectively. The use of an estimated covariance matrix for $\hat{\mu}_g$ produced an estimator with larger variance than that of the simple mean.

The predictors are unbiased under the model when the errors are normally distributed. The predictors are biased with chi-square errors because the sample mean is correlated with the sample variance. Table 2 contains the Monte Carlo bias divided by the Monte Carlo standard error of the mean. The bias of the general procedure is 20% to 50% larger than that of the diagonal procedure. In both cases, the squared bias added to the variance produces a mean square error for the procedure that is about 4% to 10% larger than the variance.

This small study demonstrates that use of an estimated covariance matrix with large variability can lead to predictors that are less efficient than the simple mean.

Table 2
Monte Carlo Relative Bias of Alternative Small Area Predictors
(10,000 samples, $r = 14$, chi-square errors)

i	$\frac{\text{Ave.}(\hat{y}_{gi} - w_i)}{[\hat{V}\{\bar{y}_i - w_i\}]^{1/2}}$	$\frac{\text{Ave.}\{\hat{y}_{di} - w_i\}}{[\hat{V}\{\bar{y}_i - w_i\}]^{1/2}}$
1	-0.28	-0.19
2	-0.27	-0.18
3	-0.30	-0.17
4	-0.27	-0.18
5	-0.26	-0.21
6	-0.29	-0.20
7	-0.24	-0.20
8	-0.24	-0.21

4. APPLICATION TO PES DATA FOR POSTCENSAL ESTIMATION

4.1 Postcensal Estimation

The U.S. Bureau of the Census provides annual estimates of the population of small areas based on the decennial censuses and on other sources of information. To consider the possible use of adjusted 1990 Census counts in the postcensal estimation process, the Bureau examined the PES data and defined a new set of 357 poststrata.

The 357 poststrata are composed of 51 poststratum groups, each of which is subdivided into 7 age-sex categories. The seven age-sex categories were (1) both sexes 0-17, (2) males 18-29, (3) males 30-49, (4) females 18-29, (5) females 30-49, (6) males 50+ and (7) females 50+. The factors that define the 51 poststratum groups are race/ethnicity (Non-Hispanic White, Black, Non-Black Hispanic, Asian, American Indian); tenure (owner, renter); type of area (urbanized area of population greater than 250,000, other urbanized area, non-urbanized area) and region (West, South, Midwest, Northeast). Due to sample size limitations, American Indians comprised a single poststratum group and Asians were dichotomized into two poststratum groups – owners and renters. Of the remaining 48 poststratum groups, the first 24 groups reflect a full cross classification of categories for Non-Hispanic White. The next 12 groups are for Black and provide a full cross classification of tenure by region for urbanized areas of population greater than 250,000 but otherwise do not provide regional detail. The same 12 poststratum groups were used for Non-Black Hispanics as were used for Blacks.

A 357 x 357 covariance matrix was obtained with the same jackknife algorithm used for the 1392 poststrata of the 1990 PES. We denote this raw covariance matrix by $\hat{\Sigma}_{ee}$. Hogan (1993) provides a detailed description of the 357 poststrata and gives the motivation for their construction.

4.2 Regression Model

We eliminated Asian and American Indian data from the smoothing process. Hence, minority refers to the combination of Black and Non-Black Hispanic. The data set of interest contains 336 adjustment factors and their estimated raw covariances. The minority by age-sex interaction was included in the regression model after examination of the 1990 data indicated that the net undercount differential between Black and Non-Black varied by sex and age-group. The regression model (1) contains 21 explanatory variables. They are:

1. X_0 = intercept
2. X_j = indicator variable for age-sex categories: $j = 1, 2, ..., 6$ in the order; ages 0-17, male 18-29, male 30-49, *etc.* (female 50+ is the class with no variable)

3. X_7 = indicator variable for renter
4. X_8 = indicator variable for Black
5. X_9 = indicator variable for Non-Black Hispanic
6. X_j = indicator variable for type of place: $j = 10, 11$ for urbanized area 250,000+ and other urban, respectively
7. X_j = indicator variable for region: $j = 12, 13, 14$ for Northeast, South and West, respectively
8. X_j = indicator variable for minority by age-sex interaction: $j = 15, ..., 20$ for minority 0-17, minority male (18-29), *etc.*

The variables X_{12}, X_{13} and X_{14} were the 1990 census proportions of persons in the poststratum group in the particular region for the Black and Non-Black Hispanic poststratum groups that were combined over regions.

A refinement was made in model (3) for the empirical application. On the basis of preliminary analysis, the specified error structure of $\mathbf{w}$, the model error, was changed from $\Sigma_{ww} = \sigma^2\mathbf{I}$ to

$$\Sigma_{ww} = \mathbf{K}_1\sigma_1^2 + \mathbf{K}_2\sigma_2^2, \tag{30}$$

where $\mathbf{K}_1$ is an $n \times n$ diagonal matrix with ones for minority poststrata and zeros elsewhere and $\mathbf{K}_2$ is an $n \times n$ diagonal matrix with ones for nonminority poststrata and zeros elsewhere. The estimated variances are $\hat{\sigma}_1^2 = 0.000506$ (0.000140) and $\hat{\sigma}_2^2 = 0.000112$ (0.000030), where the numbers in parentheses are standard errors. The standard error of the difference is (0.000141). Hence there is evidence that the variances are different for the two groups.

In our discussion of predictors, we considered two predictors, the substitution predictor of (11) and the diagonal predictor of (16). It is natural to consider a compromise predictor of the form

$$\begin{aligned}\hat{\mathbf{y}}_\varphi &= \mathbf{X}\hat{\beta}_\varphi + \hat{\mathbf{G}}_\varphi'(\mathbf{Y} - \mathbf{X}\hat{\beta}_\varphi) \\ &= \mathbf{Y} - \hat{\mathbf{H}}_\varphi'(\mathbf{Y} - \mathbf{X}\hat{\beta}_\varphi),\end{aligned} \tag{31}$$

where $0 \le \varphi \le 1$,

$$\hat{\mathbf{G}}_\varphi = \hat{\Sigma}_{\varphi\varphi}^{-1}\hat{\Sigma}_{ww},$$

$$\hat{\mathbf{H}}_\varphi = \mathbf{I} - \hat{\mathbf{G}}_\varphi = \hat{\Sigma}_{\varphi\varphi}^{-1}\left[\varphi\hat{\mathbf{D}}_{ee} + (1-\varphi)\hat{\Sigma}_{ee}\right],$$

$$\hat{\Sigma}_{\varphi\varphi} = \hat{\Sigma}_{ww} + \varphi\hat{\mathbf{D}}_{ee} + (1-\varphi)\hat{\Sigma}_{ee},$$

$$\hat{\mathbf{D}}_{ee} = \operatorname{diag}\left\{\hat{\Sigma}_{ee}\right\},$$

$$\hat{\beta}_\varphi = \left(\mathbf{X}'\hat{\Sigma}_{\varphi\varphi}^{-1}\mathbf{X}\right)^{-1}\mathbf{X}'\hat{\Sigma}_{\varphi\varphi}^{-1}\mathbf{Y},$$

and

$$\hat{\Sigma}_{ww} = \mathbf{K}_1\hat{\sigma}_1^2 + \mathbf{K}_2\hat{\sigma}_2^2.$$

The predictor (31) with $\varphi = 0$ is the substitution predictor and the predictor (31) with $\varphi = 1$ the diagonal predictor. There should be some φ, $0 < \varphi < 1$, that gives a predictor with smaller prediction variance than either of the extremes.

The PES direct estimate of the total number of persons is the weighted sum of the adjustment factors, where the weights are the census counts in the post strata. The standard error of the direct estimator of the total is relatively small and the direct estimator is judged to be the preferred estimator of the total. Therefore, the model predictors are constructed subject to the constraint that the weighted sum of the predictors is equal to the direct estimate of the total. Thus, the restriction is

$$\hat{Y}_T = \sum_{i=1}^{336} a_i y_i = \sum_{i=1}^{336} a_i \tilde{y}_i,$$

where $\hat{Y}_T$ is PES direct estimator of the total, a_i is the census count in the i-th post stratum, and $\tilde{y}_i$ is the final predictor. In the actual computations the a_i were normalized to sum to one. Battese, Harter and Fuller (1988) made an adjustment in the predictions to create estimators to meet the restriction. Ghosh and Rao (1994) discuss such adjustments. We use a procedure that permits direct estimation of the variance of the restricted predictions.

We imposed the restriction on the initial predictors by a procedure that, approximately, constructed the best predictors of 335 quantities that are estimated to be uncorrelated with $\hat{Y}_T$. Let $\hat{\Sigma}_{zz}$ be the estimated covariance matrix of $\mathbf{Y} = (Y_1, Y_2, ..., Y_{336})'$ and define

$$\mathbf{CY} = (\hat{Y}_T, Y_2 - b_2\hat{Y}_T, ..., Y_{336} - b_{336}\hat{Y}_T)',$$

where

$$\mathbf{C} = \mathbf{BT},$$

$$\mathbf{T} = \begin{pmatrix} \mathbf{a} \\ \mathbf{0} \quad \mathbf{I}_{335} \end{pmatrix},$$

$$\mathbf{a} = (a_1, a_2, ..., a_{336}),$$

$$\mathbf{B} = \begin{pmatrix} 1 & \mathbf{0}' \\ -b_{335} & \mathbf{I}_{335} \end{pmatrix},$$

$$\mathbf{b}_{335} = \begin{pmatrix} \mathbf{0}' \\ \mathbf{I}_{335} \end{pmatrix}' \hat{\Sigma}_{zz}\mathbf{a}'\left(\mathbf{a}\hat{\Sigma}_{zz}\mathbf{a}'\right)^{-1},$$

$\mathbf{I}_k$ is the $k \times k$ identity matrix, and $\mathbf{0}$ is a column vector containing all zeros. The elements of $\mathbf{CY}$ are uncorrelated with $\hat{Y}_T$.

If we let $\hat{\mathbf{y}}$ be the model predictor of $\mathbf{y}$, then the model predictor of $\mathbf{Cy}$ is $\mathbf{C}\hat{\mathbf{y}}$. If we use the model predictor for the last 335 elements of $\mathbf{Cy}$ and use $\hat{Y}_T$ as the estimator for the first element of $\mathbf{Cy}$, the predictor of $\mathbf{y}$ is

$$\tilde{\mathbf{y}} = \mathbf{Y} - \mathbf{C}^{-1}\mathbf{AC}\hat{\ddot{\mathbf{H}}}_\varphi'(\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}_\varphi), \qquad (32)$$

where

$$\mathbf{A} = \begin{pmatrix} 0 & \mathbf{0}' \\ \mathbf{0} & \mathbf{I}_{335} \end{pmatrix}.$$

The estimated variance of $\tilde{\mathbf{y}}$ is

$$\hat{\mathbf{V}}\{\tilde{\mathbf{y}} - \mathbf{y}\}$$

$$= (\mathbf{I} - \ddot{\mathbf{H}}_\varphi')\hat{\boldsymbol{\Sigma}}_{ee}(\mathbf{I} - \ddot{\mathbf{H}}_\varphi')' + \ddot{\mathbf{H}}_\varphi'\hat{\boldsymbol{\Sigma}}_{ww}\ddot{\mathbf{H}}_\varphi$$

$$+ \mathbf{C}^{-1}\mathbf{AC}\left[\hat{\mathbf{H}}_\varphi'\mathbf{X}\hat{\mathbf{V}}_{\beta\beta}\mathbf{X}'\hat{\mathbf{H}}_\varphi + \hat{\boldsymbol{\Gamma}}_{33} + \hat{\boldsymbol{\Gamma}}_{44}\right]\mathbf{C}'\mathbf{AC}^{-1\prime}, \qquad (33)$$

where $\ddot{\mathbf{H}}_\varphi' = \mathbf{C}^{-1}\mathbf{AC}\hat{\mathbf{H}}_\varphi'$, and $\hat{\mathbf{V}}_{\beta\beta}$, $\hat{\boldsymbol{\Gamma}}_{33}$ and $\hat{\boldsymbol{\Gamma}}_{44}$ are defined in Appendix B. The sum of the first two terms on the right of (33) is an estimator of the variance treating $\ddot{\mathbf{H}}_\varphi$ as a fixed matrix. The final term on the right of (33) estimates the increase in variance due to estimating the variance.

4.3 Smoothed Factors

For the vector of 336 observations, we produced smoothed factors using the generalized predictor (32) for several values of φ. Note that $\varphi = 0$ corresponds to the substitution predictor and $\varphi = 1$ corresponds to the diagonal predictor.

The estimated standard errors of the predictors were calculated using the crude variance approximation of Appendix B. The average of the ratios of the standard error of $\tilde{y}_\varphi$ to $\tilde{y}_{0.6}$ for some selected values of φ are given in Table 3. The ordering of the ratios is approximately the same for the 48 stratum groups as for the original 336 poststrata. A poststratum group is formed by combining the seven age-sex cells within a given race-by-tenure-by-urbanity-by-region classification. On the basis of these calculations, a φ of 0.5 or 0.6 is the preferred estimator, although the estimated differences in efficiencies are not large. Any member of the φ-class is much superior to the original Y-estimator. The average estimated variance efficiency is about 400% for the φ-predictors, relative to the original poststratum estimators.

Table 3
Average of Ratio of Standard Error of $\tilde{y}_\varphi$ and of Y to Stardard Error of $\tilde{y}_{0.6}$

Predictor	336 Poststrata	48 Poststratum groups
$\varphi = 0$	1.014	1.045
$\varphi = 0.5$	0.995	1.001
$\varphi = 0.6$	1.000	1.000
$\varphi = 0.7$	1.006	1.001
$\varphi = 0.8$	1.014	1.005
$\varphi = 1.0$	1.046	1.037
Original Y	2.235	2.294

Table 4 presents the raw PES estimates, $\mathbf{Y}$, and the $\tilde{y}_{0.6}$ estimates of net undercount for each of 48 poststratum groups. The net undercount is the difference between the estimated total population in the poststratum and the census count divided by the census count.

We chose $\varphi = 0.6$ as the preferred estimator on the basis of the crude standard error ratios of Table 3. The predictions and standard errors are very similar for $\varphi = 0.5$, 0.6 and 0.7. A φ greater than zero has advantages over a φ of zero. The accuracy of the numerical calculations should be better with φ greater than zero because $\hat{\boldsymbol{\Sigma}}_{\varphi\varphi}$ has larger eigenvalues with $\varphi > 0$ than with $\varphi = 0$. One could make a case for using $\varphi = 1.0$ because of the simplicity of the calculations and of the good estimated relative efficiency.

The estimated standard errors of the predictors are considerably smaller than those of the raw estimates. In addition, the set of predictors contains fewer extreme estimates. For example, for poststratum groups 34, 39 and 48, the $\tilde{y}_{0.6}$ estimates of the percent net undercount are 6.04, 0.17 and 7.51 while the raw estimates are 11.06, -4.14 and 18.76, respectively. Most smoothed estimates differ from the direct estimate by less than one direct estimated standard error. The three largest standardized differences are for Black Owner-Large Urban in the West, Black Renter-Large Urban in the Northeast, and Non-Black Hispanic Owner-Large Urban in the Midwest. In the three cases, the difference between the direct estimate and the smoothed estimate divided by the direct standard error is about 1.8.

ACKNOWLEDGEMENTS

This research was partly supported by Cooperative Agreement 43-3AEU-3-80088 between Iowa State University, the National Agricultural Statistics Service and the U.S. Bureau of the Census. This paper reports the results of research and analysis undertaken by staff of the Bureau of the Census and Iowa State University. It has undergone a more limited review than official Census Bureau publications. This paper is released to inform interested parties of research and to encourage discussion. We thank the referees and editors for many comments that led to improvements in the manuscript.

Table 4
Estimated Percent Net Undercount by Poststratum Group

Poststratum Group	Y	s.e. (Y)	$\tilde{y}_{0.6}$	s.e. $(\tilde{y}_{0.6})$
Non-Hispanic White Owner Large Urban				
1. N.E.	-2.08	1.04	-0.63	0.60
2. South	0.69	0.72	0.38	0.44
3. Midwest	-0.26	0.39	-0.13	0.31
4. West	-0.34	0.64	-0.02	0.44
Non-Hispanic White Owner Other Urban				
5. N.E.	-1.07	0.48	-0.73	0.35
6. South	0.52	0.43	0.53	0.33
7. Midwest	-0.10	0.40	0.01	0.31
8. West	0.63	0.58	0.30	0.40
Non-Hispanic White Owner Non-Urban				
9. N.E.	-0.53	0.69	-0.28	0.47
10. South	0.18	0.69	0.58	0.45
11 Midwest	-0.70	1.16	0.16	0.64
12. West	0.29	0.69	0.38	0.46
Non-Hispanic White Renter Large Urban				
13. N.E.	1.17	1.43	2.07	0.61
14. South	2.62	1.56	3.53	0.64
15. Midwest	2.39	1.70	2.53	0.60
16. West	3.28	1.72	3.10	0.58
Non-Hispanic White Renter Other Urban				
17. N.E.	3.53	1.62	2.29	0.61
18. South	3.30	1.86	3.67	0.67
19. Midwest	1.24	1.13	2.39	0.53
20. West	4.70	1.47	3.20	0.57
Non-Hispanic White Renter Non- Urban				
21. N.E.	6.97	4.67	3.54	0.92
22. South	6.65	1.93	3.60	0.66
23. Midwest	2.93	1.60	2.36	0.66
24. West	6.48	2.06	3.48	0.67
Black Owner Large Urban				
25. N.E.	1.65	1.96	0.97	0.91
26. South	2.20	0.94	2.30	0.70
27. Midwest	0.82	0.88	1.13	0.67
28. West	6.49	2.16	2.54	0.96
Black Owner Other Urban				
29. U.S.	1.36	1.01	2.05	0.72
Black Owner Non- Urban				
30. U.S.	3.64	2.03	2.85	0.98
Black Renter Large Urban				
31. N.E.	9.13	1.93	5.57	0.96
32. South	6.69	2.17	6.42	1.10
33. Midwest	6.38	1.91	5.43	1.03
34. West	11.06	3.35	6.04	1.12
Black Renter Other Urban				
35. U.S.	4.33	1.28	4.99	0.82
Black Renter Non- Urban				
36. U.S.	4.84	5.95	5.90	1.24
Non-Black Hispanic Owner Large Urban				
37. N.E.	0.68	4.44	3.00	1.18
38. South	2.59	0.95	2.52	0.72
39. Midwest	-4.14	2.38	0.17	0.97
40. West	2.98	0.92	2.89	0.68
Non-Black Hispanic Owner Other Urban				
41. U.S.	0.95	1.70	2.32	0.87
Non-Black Hispanic Owner Non-Urban				
42. U.S.	2.80	2.83	2.88	1.16
Non-Black Hispanic Renter Large Urban				
43. N.E.	7.21	4.04	5.85	1.27
44. South	10.30	3.11	7.35	1.15
45. Midwest	7.11	3.74	5.71	1.21
46. West	6.29	2.09	6.45	0.98
Non-Black Hispanic Renter Other Urban				
47. U.S.	7.07	3.10	6.26	1.09
Non-Black Hispanic Renter Non- Urban				
48. U.S.	18.76	7.24	7.51	1.38

APPENDIX A: Estimation of Σ_{ww}

The estimators of σ_1^2 and σ_2^2 of Σ_{ww} are patterned after analysis of variance estimators. The estimation process contains several steps using improved estimators from one step in the next step. We partition the regression problem as

$$\begin{pmatrix} \mathbf{Y}_1 \\ \mathbf{Y}_2 \end{pmatrix} = \begin{pmatrix} \mathbf{X}_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{X}_2 \end{pmatrix} \begin{pmatrix} \boldsymbol{\beta}_1 \\ \boldsymbol{\beta}_2 \end{pmatrix} + \begin{pmatrix} \mathbf{z}_1 \\ \mathbf{z}_2 \end{pmatrix},$$

where $(\mathbf{Y}_1, \mathbf{X}_1)$ contains the observations for minorities and $(\mathbf{Y}_2, \mathbf{X}_2)$ contains the remaining observations. Let $\mathbf{Y}_1$ be an n_1-dimensional column vector and $\mathbf{Y}_2$ be an n_2-dimensional column vector observations. An initial estimator of $(\boldsymbol{\beta}_1', \boldsymbol{\beta}_2')'$ is

$$\begin{pmatrix} \tilde{\boldsymbol{\beta}}_1 \\ \tilde{\boldsymbol{\beta}}_2 \end{pmatrix} = \begin{pmatrix} \left(\mathbf{X}_1'\hat{\boldsymbol{\Sigma}}_{ee11}^{-1}\mathbf{X}_1\right)^{-1} & \mathbf{X}_1'\hat{\boldsymbol{\Sigma}}_{ee11}\mathbf{Y}_1 \\ \left(\mathbf{X}_2'\hat{\boldsymbol{\Sigma}}_{ee22}^{-1}\mathbf{X}_2\right)^{-2} & \mathbf{X}_2'\hat{\boldsymbol{\Sigma}}_{ee22}\mathbf{Y}_2 \end{pmatrix},$$

where

$$\hat{\boldsymbol{\Sigma}}_{ee} = \begin{pmatrix} \hat{\boldsymbol{\Sigma}}_{ee11} & \hat{\boldsymbol{\Sigma}}_{ee12} \\ \hat{\boldsymbol{\Sigma}}_{ee21} & \hat{\boldsymbol{\Sigma}}_{ee22} \end{pmatrix}$$

is partitioned to conform to the partition of $\mathbf{Y}$.

Initial estimators of σ_1^2 and σ_2^2 are

$$\tilde{\sigma}_i^2 = \max\left\{\left[(\mathbf{Y}_i - \mathbf{X}_i\tilde{\boldsymbol{\beta}}_i)'\hat{\boldsymbol{\Sigma}}_{eeii}^{-1}(\mathbf{Y}_i - \mathbf{X}_i\tilde{\boldsymbol{\beta}}_i) - g_{1i}\right]g_{2i}^{-1}, 0\right\},$$

for $i = 1, 2$, where

$$g_{1i} = \mathrm{tr}\left\{\hat{\boldsymbol{\Sigma}}_{ee11}(\mathbf{I}_{ni} - \mathbf{X}_i\mathbf{A}_{Mii})'\hat{\boldsymbol{\Sigma}}_{eeii}^{-1}(\mathbf{I}_{ni} - \mathbf{X}_i\mathbf{A}_{Mii}),\right.$$

$$g_{2i} = \mathrm{tr}\left\{(\mathbf{I}_{ni} - \mathbf{X}_i\mathbf{A}_{Mii})'\hat{\boldsymbol{\Sigma}}_{eeii}^{-1}(\mathbf{I}_{ni} - \mathbf{X}_i\mathbf{A}_{Mii}),\right.$$

$$\mathbf{A}_{Mii} = \left(\mathbf{X}_i'\hat{\boldsymbol{\Sigma}}_{eeii}^{-1}\mathbf{X}_i\right)^{-1}\mathbf{X}_i'\hat{\boldsymbol{\Sigma}}_{eeii}^{-1},$$

and $\mathbf{I}_{ni}$ is the $n_i \times n_i$ identity matrix.

The final estimators are

$$\ddot{\sigma}_i^2 = \max\left\{\left[(\mathbf{Y}_i - \mathbf{X}_i\ddot{\boldsymbol{\beta}}_i)'\tilde{\boldsymbol{\Sigma}}_{zzii}^{-1}(\mathbf{Y}_i - \mathbf{X}_i\ddot{\boldsymbol{\beta}}_i) - \ddot{g}_{1i}\right]\ddot{g}_{2i}^{-1}, 0\right\},$$

for $i = 1, 2$, where

$$\ddot{g}_{1i} = \mathrm{tr}\left\{\hat{\boldsymbol{\Sigma}}_{eeii}(\mathbf{I}_{ni} - \mathbf{X}_i\ddot{\mathbf{A}}_{Mii})\tilde{\boldsymbol{\Sigma}}_{zzii}^{-1}(\mathbf{I}_{ni} - \mathbf{X}_i\ddot{\mathbf{A}}_{Mii})\right\}$$

$$\ddot{g}_{2i} = \mathrm{tr}\left\{(\mathbf{I}_{ni} - \mathbf{X}_i\ddot{\mathbf{A}}_{Mii})\tilde{\boldsymbol{\Sigma}}_{zzii}^{-1}(\mathbf{I}_{ni} - \mathbf{X}_i\ddot{\mathbf{A}}_{Mii})\right\}$$

$$\tilde{\boldsymbol{\Sigma}}_{zzii} = \hat{\boldsymbol{\Sigma}}_{eeii} + \tilde{\sigma}_i^2\mathbf{I}_{ni}$$

$$\ddot{\boldsymbol{\beta}}_i = \left(\mathbf{X}_i'\hat{\boldsymbol{\Sigma}}_{zzii}^{-1}\mathbf{X}_i\right)^{-1}\mathbf{X}_i'\hat{\boldsymbol{\Sigma}}_{zzii}^{-1}\mathbf{Y}_i = \ddot{\mathbf{A}}_{Mii}^{-1}\mathbf{Y}_i.$$

Estimators of the variance are

$$\hat{V}\{\ddot{\sigma}_i^2\} = 2\ddot{g}_{2i}^{-2} \times \mathrm{tr}\left\{\left[\tilde{\boldsymbol{\Sigma}}_{zzii}(\mathbf{I}_{ni} - \mathbf{X}_i\ddot{\mathbf{A}}_{Mii})'\tilde{\boldsymbol{\Sigma}}_{zzii}^{-1}(\mathbf{I}_{ni} - \mathbf{X}_i\ddot{\mathbf{A}}_{Mii})\right]^2\right\}$$
$$+ 2\ddot{g}_{2i}^{-2}d_e^{-1} \times \mathrm{tr}\left\{\left[\hat{\boldsymbol{\Sigma}}_{eeii}(\mathbf{I}_{ni} - \mathbf{X}_i\ddot{\mathbf{A}}_{Mii})'\tilde{\boldsymbol{\Sigma}}_{zzii}^{-1}(\mathbf{I}_{ni} - \mathbf{X}_i\ddot{\mathbf{A}}_{Mii})\right]^2\right\},$$

for $i = 1, 2$. The estimated covariance is

$$\hat{C}\{\ddot{\sigma}_1^2, \ddot{\sigma}_2^2\} = 2\mathrm{tr}\left\{\tilde{\boldsymbol{\Sigma}}_{zz}\mathbf{M}_{11}\tilde{\boldsymbol{\Sigma}}_{zz}\mathbf{M}_{22}\right\} + 2d_e^{-1}\mathrm{tr}\left\{\hat{\boldsymbol{\Sigma}}_{ee}\mathbf{M}_{11}\hat{\boldsymbol{\Sigma}}_{ee}\mathbf{M}_{22}\right\},$$

where

$$\mathbf{M}_{11} = \begin{pmatrix} g_{21}^{-1}(\mathbf{I}_{n_1} - \mathbf{X}_1\mathbf{A}_{M11})'\boldsymbol{\Sigma}_{zz11}^{-1}(\mathbf{I}_{n_1} - \mathbf{X}_1\mathbf{A}_{M11}) & \mathbf{0} \\ \mathbf{0}' & \mathbf{0} \end{pmatrix}$$

and

$$\mathbf{M}_{22} = \begin{pmatrix} \mathbf{0} & \mathbf{0}' \\ \mathbf{0} & g_{22}^{-1}(\mathbf{I}_{n_2} - \mathbf{X}_2\mathbf{A}_{M22})'\boldsymbol{\Sigma}_{zz22}^{-1}(\mathbf{I}_{n_2} - \mathbf{X}_2\mathbf{A}_{M22}) \end{pmatrix}.$$

See Searle (1971, Chapter 2 and p. 435).

APPENDIX B: Approximations for the Variance of Predictors

Our model is

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{w} + \mathbf{e}, \tag{B.1}$$

where $\mathbf{Y}$ is an n-dimensional column vector, $\mathbf{X}$ is an $n \times k$ fixed matrix,

$$\begin{pmatrix} \mathbf{w} \\ \mathbf{e} \end{pmatrix} \sim N\left(\begin{pmatrix} \mathbf{0} \\ \mathbf{0} \end{pmatrix}, \begin{pmatrix} \boldsymbol{\Sigma}_{ww} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\Sigma}_{ee} \end{pmatrix}\right), \tag{B.2}$$

and $\boldsymbol{\Sigma}_{ww}$ is defined in (30) of the text.

For purpose of variance estimation, we assume $\hat{\boldsymbol{\Sigma}}_{ee}$ is an unbiased estimator of $\boldsymbol{\Sigma}_{ee}$ distributed as a multiple of a Wishart matrix with d_e degrees of freedom independent of $(\mathbf{w}, \mathbf{e})$. We let $\mathbf{y}$ be the unknown true vector to be predicted and write

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{w} \quad \text{and} \quad \mathbf{z} = \mathbf{w} + \mathbf{e}.$$

By a Taylor expansion

$$\hat{\mathbf{y}}_\varphi - \mathbf{y} = \mathbf{e} - \hat{\mathbf{H}}_\varphi'(\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}_\varphi)$$
$$= \mathbf{e} - \mathbf{H}_\varphi'\mathbf{z} + \mathbf{H}_\varphi'\mathbf{X}(\hat{\boldsymbol{\beta}}_\varphi' - \boldsymbol{\beta}) - (\hat{\mathbf{H}}_\varphi' - \mathbf{H}_\varphi')\mathbf{z} + O_p(n^{-1}), \tag{B.3}$$

where

$$\mathbf{H}_\varphi = \boldsymbol{\Sigma}_{\varphi\varphi}^{-1}\left[\varphi\mathbf{D}_{ee} + (1-\varphi)\boldsymbol{\Sigma}_{ee}\right]$$

and $\hat{\mathbf{H}} = \hat{\mathbf{H}}_\varphi$ is defined in (31). The error in $\hat{\beta}_\varphi$ is

$$\hat{\beta}_\varphi - \beta = \left(\mathbf{X}'\hat{\boldsymbol{\Sigma}}_{\varphi\varphi}^{-1}\mathbf{X}\right)^{-1}\mathbf{X}'\hat{\boldsymbol{\Sigma}}_{\varphi\varphi}^{-1}\mathbf{z} \tag{B.4}$$

Now $\hat{\boldsymbol{\Sigma}}_{ee}$ is independent of $\mathbf{z}$ and $\mathbf{Y} - \mathbf{X}\tilde{\boldsymbol{\beta}}$ is uncorrelated with $\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}$ if the true $\boldsymbol{\Sigma}_{zz}$ is used in place of $\hat{\boldsymbol{\Sigma}}_{\varphi\varphi}$. Therefore

$$E\left\{\mathbf{H}_0'\mathbf{X}(\tilde{\beta}_0 - \beta)\mathbf{z}'(\hat{\mathbf{H}}_0 - \mathbf{H}_0)\right\} = \mathbf{0}, \tag{B.5}$$

where $\hat{\mathbf{H}}_0$ is constructed using $\mathbf{Y} - \mathbf{X}\tilde{\boldsymbol{\beta}}$ in the estimators of the elements of $\hat{\boldsymbol{\Sigma}}_{ww}$ defined in Appendix A and $\mathbf{H}_0 = \boldsymbol{\Sigma}_{zz}^{-1}\boldsymbol{\Sigma}_{ee}$. We set the covariance between $\hat{\beta}$ and $\hat{\mathbf{H}}_\varphi$ equal to zero for all φ. Now

$$\begin{aligned}\hat{\mathbf{H}}_\varphi &= \hat{\boldsymbol{\Sigma}}_{\varphi\varphi}^{-1}\left[\varphi\hat{\mathbf{D}}_{ee} + (1-\varphi)\hat{\boldsymbol{\Sigma}}_{ee}\right] \\ &= \left[\hat{\boldsymbol{\Sigma}}_{ww} + \varphi\hat{\mathbf{D}}_{ee} + (1-\varphi)\hat{\boldsymbol{\Sigma}}_{ee}\right]^{-1}\left[\varphi\hat{\mathbf{D}}_{ee} + (1-\varphi)\hat{\boldsymbol{\Sigma}}_{ee}\right]\end{aligned}$$

and

$$\begin{aligned}\hat{\mathbf{H}}_\varphi - \mathbf{H}_\varphi &= \boldsymbol{\Sigma}_{\varphi\varphi}^{-1}\left[\varphi(\hat{\mathbf{D}}_{ee} - \mathbf{D}_{ee}) + (1-\varphi)\left(\hat{\boldsymbol{\Sigma}}_{ee} - \boldsymbol{\Sigma}_{ee}\right)\right] \\ &\quad - \boldsymbol{\Sigma}_{\varphi\varphi}^{-1}\left[\hat{\boldsymbol{\Sigma}}_{ww} - \boldsymbol{\Sigma}_{ww} + \varphi(\hat{\mathbf{D}}_{ee} - \mathbf{D}_{ee})\right. \\ &\qquad \left. + (1-\varphi)\left(\hat{\boldsymbol{\Sigma}}_{ee} - \boldsymbol{\Sigma}_{ee}\right)\right]\mathbf{H}_\varphi \\ &= \boldsymbol{\Sigma}_{\varphi\varphi}^{-1}\left[\varphi(\hat{\mathbf{D}}_{ee} - \mathbf{D}_{ee} + (1-\varphi)\left(\hat{\boldsymbol{\Sigma}}_{ee} - \boldsymbol{\Sigma}_{ee}\right)\right]\mathbf{G}_\varphi \\ &\quad - \boldsymbol{\Sigma}_{\varphi\varphi}^{-1}\left[\hat{\boldsymbol{\Sigma}}_{ww} - \boldsymbol{\Sigma}_{ww}\right]\mathbf{H}_\varphi,\end{aligned}$$

where $\mathbf{G}_\varphi = \mathbf{I} - \mathbf{H}_\varphi$. The contribution of $\hat{\mathbf{D}}_{ee} - \mathbf{D}_{ee}$ to the variance of $\hat{\mathbf{H}}_\varphi$ is small relative to the contribution of $\hat{\boldsymbol{\Sigma}}_{ee} - \boldsymbol{\Sigma}_{ee}$. Therefore, we omit $\hat{\mathbf{D}}_{ee} - \mathbf{D}_{ee}$ in our variance approximation. Then the expectation

$$\begin{aligned}&E\left\{(\mathbf{I} - \mathbf{H}_\varphi)'\left(\hat{\boldsymbol{\Sigma}}_{ee} - \boldsymbol{\Sigma}_{ee}\right)\boldsymbol{\Sigma}_{\varphi\varphi}^{-1}\mathbf{z}\mathbf{z}'\boldsymbol{\Sigma}_{\varphi\varphi}^{-1}\right. \\ &\qquad \left.\left(\hat{\boldsymbol{\Sigma}}_{ee} - \boldsymbol{\Sigma}_{ee}\right)(\mathbf{I} - \mathbf{H}_\varphi)\right\} \\ &= d_e^{-1}\mathbf{G}_\varphi'\left[\boldsymbol{\Sigma}_{ee}\left(\mathrm{tr}\left\{\boldsymbol{\Lambda}\boldsymbol{\Sigma}_{ee}\right\}\right) + \boldsymbol{\Sigma}_{ee}\boldsymbol{\Lambda}\boldsymbol{\Sigma}_{ee}\right]\mathbf{G}_\varphi,\end{aligned} \tag{B.6}$$

where $\boldsymbol{\Lambda} = \boldsymbol{\Sigma}_{\varphi\varphi}^{-1}\boldsymbol{\Sigma}_{zz}\boldsymbol{\Sigma}_{\varphi\varphi}^{-1}$, because $\mathbf{z}$ is independent of $\hat{\boldsymbol{\Sigma}}_{ee}$. We also omit the term $d_e^{-1}\mathbf{G}_\varphi'\boldsymbol{\Sigma}_{ee}\boldsymbol{\Sigma}_{\varphi\varphi}^{-1}\boldsymbol{\Sigma}_{zz}\boldsymbol{\Sigma}_{\varphi\varphi}^{-1}\boldsymbol{\Sigma}_{ee}\mathbf{G}_\varphi$ in our variance approximation.

The expectation for the term containing $(\hat{\boldsymbol{\Sigma}}_{ww} - \boldsymbol{\Sigma}_{ww})$ is

$$E\left\{\mathbf{H}_\varphi'(\hat{\boldsymbol{\Sigma}}_{ww} - \boldsymbol{\Sigma}_{ww})\boldsymbol{\Sigma}_{\varphi\varphi}^{-1}\mathbf{z}\mathbf{z}'\boldsymbol{\Sigma}_{\varphi\varphi}^{-1}(\hat{\boldsymbol{\Sigma}}_{ww} - \boldsymbol{\Sigma}_{ww})\mathbf{H}_\varphi\right\},$$

where

$$\hat{\boldsymbol{\Sigma}}_{ww} - \boldsymbol{\Sigma}_{ww} = \begin{pmatrix} \mathbf{I}_{n1}(\hat{\sigma}_1^2 - \sigma_1^2) & \mathbf{0} \\ \mathbf{0} & \mathbf{I}_{n2}(\hat{\sigma}_2^2 - \sigma_2^2) \end{pmatrix}.$$

Approximating the expectation by treating $\mathbf{z}$ as independent of $\hat{\boldsymbol{\Sigma}}_{ww}$, we obtain

$$\mathbf{H}_\varphi'\begin{pmatrix} \boldsymbol{\Lambda}_{11}V\{\hat{\sigma}_1^2\} & \boldsymbol{\Lambda}_{12}C\{\hat{\sigma}_1^2, \hat{\sigma}_2^2\} \\ \boldsymbol{\Lambda}_{21}C\{\hat{\sigma}_1^2, \hat{\sigma}_2^2\} & \boldsymbol{\Lambda}_{22}V\{\hat{\sigma}_2^2\} \end{pmatrix}\mathbf{H}_\varphi, \tag{B.7}$$

where

$$\boldsymbol{\Lambda} = \begin{pmatrix} \boldsymbol{\Lambda}_{11} & \boldsymbol{\Lambda}_{12} \\ \boldsymbol{\Lambda}_{21} & \boldsymbol{\Lambda}_{22} \end{pmatrix} = \boldsymbol{\Sigma}_{\varphi\varphi}^{-1}\boldsymbol{\Sigma}_{zz}\boldsymbol{\Sigma}_{\varphi\varphi}^{-1}.$$

The Taylor expansion of $\hat{\beta} - \beta$ is

$$\begin{aligned}\hat{\beta} - \beta &= (\mathbf{X}'\boldsymbol{\Sigma}_{\varphi\varphi}^{-1}\mathbf{X})^{-1}\mathbf{X}'\hat{\boldsymbol{\Sigma}}_{\varphi\varphi}^{-1}\mathbf{z} \\ &= (\mathbf{X}'\boldsymbol{\Sigma}_{\varphi\varphi}^{-1}\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\Sigma}_{\varphi\varphi}^{-1}\mathbf{z} \\ &\quad + (\mathbf{X}'\boldsymbol{\Sigma}_{\varphi\varphi}^{-1}\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\Sigma}_{\varphi\varphi}^{-1}(\hat{\boldsymbol{\Sigma}}_{\varphi\varphi} - \boldsymbol{\Sigma}_{\varphi\varphi}) \\ &\qquad \times \boldsymbol{\Sigma}_{\varphi\varphi}^{-1}\mathbf{X}(\mathbf{X}'\boldsymbol{\Sigma}_{\varphi\varphi}^{-1}\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\Sigma}_{\varphi\varphi}^{-1}\mathbf{z} \\ &\quad - (\mathbf{X}'\boldsymbol{\Sigma}_{\varphi\varphi}^{-1}\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\Sigma}_{\varphi\varphi}^{-1} \\ &\qquad \times (\hat{\boldsymbol{\Sigma}}_{\varphi\varphi} - \boldsymbol{\Sigma}_{\varphi\varphi})\boldsymbol{\Sigma}_{\varphi\varphi}^{-1}\mathbf{z} + \text{Remainder}. \\ &= (\mathbf{X}'\boldsymbol{\Sigma}_{\varphi\varphi}^{-1}\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\Sigma}_{\varphi\varphi}^{-1}\mathbf{z} \\ &\quad + \mathbf{L}(\hat{\boldsymbol{\Sigma}}_{\varphi\varphi} - \boldsymbol{\Sigma}_{\varphi\varphi})\boldsymbol{\Sigma}_{\varphi\varphi}^{-1}\mathbf{Q}\boldsymbol{\Sigma}_{\varphi\varphi}^{-1}\mathbf{z} \\ &\quad - \mathbf{L}(\hat{\boldsymbol{\Sigma}}_{\varphi\varphi} - \boldsymbol{\Sigma}_{\varphi\varphi})\boldsymbol{\Sigma}_{\varphi\varphi}^{-1}\mathbf{z} + \text{Remainder}\end{aligned} \tag{B.8}$$

where $\mathbf{Q} = \mathbf{X}(\mathbf{X}'\boldsymbol{\Sigma}_{\varphi\varphi}^{-1}\mathbf{X})^{-1}\mathbf{X}'$ and $\mathbf{L} = (\mathbf{X}'\boldsymbol{\Sigma}_{\varphi\varphi}^{-1}\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\Sigma}_{\varphi\varphi}^{-1}$.

If $\boldsymbol{\Sigma}_{\varphi\varphi} = \boldsymbol{\Sigma}_{zz}$ and if $\hat{\boldsymbol{\Sigma}}_{zz}$ is distributed as a multiple of a Wishart with d_e degrees of freedom, independent of $\mathbf{z}$, then

$$\begin{aligned}&E\{\mathbf{L}\left(\hat{\boldsymbol{\Sigma}}_{zz} - \boldsymbol{\Sigma}_{zz}\right)\boldsymbol{\Sigma}_{zz}^{-1}\mathbf{Q}\boldsymbol{\Sigma}_{zz}^{-1}\mathbf{z}\mathbf{z}'\boldsymbol{\Sigma}_{zz}^{-1}\mathbf{Q}\boldsymbol{\Sigma}_{zz}^{-1} \\ &\qquad \times \left(\hat{\boldsymbol{\Sigma}}_{zz} - \boldsymbol{\Sigma}_{zz}\right)\mathbf{L}'\} \\ &= d_e^{-1}\mathbf{L}\left[\boldsymbol{\Sigma}_{zz}\mathrm{tr}\left\{\boldsymbol{\Sigma}_{zz}^{-1}\mathbf{Q}\right\} + \mathbf{Q}\right]\mathbf{L}' \\ &= d_e^{-1}\left(\mathbf{X}'\boldsymbol{\Sigma}_{zz}^{-1}\mathbf{X}\right)^{-1}(k+1).\end{aligned}$$

Using a similar approximation

$$\begin{aligned}&E\left\{\mathbf{L}\left(\hat{\boldsymbol{\Sigma}}_{zz} - \boldsymbol{\Sigma}_{zz}\right)\boldsymbol{\Sigma}_{zz}^{-1}\mathbf{X}\mathbf{L}\mathbf{z}\mathbf{z}'\boldsymbol{\Sigma}_{zz}^{-1}|\hat{\boldsymbol{\Sigma}}_{zz} - \boldsymbol{\Sigma}_{zz}\mathbf{L}'\right\} \\ &\doteq E\left\{\mathbf{L}\left(\hat{\boldsymbol{\Sigma}}_{zz} - \boldsymbol{\Sigma}_{zz}\right)\boldsymbol{\Sigma}_{zz}^{-1}\mathbf{X}\mathbf{L}\boldsymbol{\Sigma}_{zz}\boldsymbol{\Sigma}_{zz}^{-1}\left(\hat{\boldsymbol{\Sigma}}_{zz} - \boldsymbol{\Sigma}_{zz}\right)\mathbf{L}'\right\} \\ &= E\left\{\mathbf{L}\left(\hat{\boldsymbol{\Sigma}}_{zz} - \boldsymbol{\Sigma}_{zz}\right)\boldsymbol{\Sigma}_{zz}^{-1}\mathbf{Q}\,\boldsymbol{\Sigma}_{zz}^{-1}\left(\hat{\boldsymbol{\Sigma}}_{zz} - \boldsymbol{\Sigma}_{zz}\right)\mathbf{L}'\right\} \\ &= d_e^{-1}\left(\mathbf{X}'\boldsymbol{\Sigma}_{zz}^{-1}\mathbf{X}\right)^{-1}(k+1).\end{aligned}$$

On the basis of this result, we use the approximation

$$\hat{\beta} - \beta \doteq \left(\mathbf{X}'\boldsymbol{\Sigma}_{\varphi\varphi}^{-1}\mathbf{X}\right)^{-1}\mathbf{X}'\boldsymbol{\Sigma}_{\varphi\varphi}^{-1}\mathbf{z} - \mathbf{L}\left(\hat{\boldsymbol{\Sigma}}_{\varphi\varphi} - \boldsymbol{\Sigma}_{\varphi\varphi}\right)\boldsymbol{\Sigma}_{\varphi\varphi}^{-1}\mathbf{z}.$$

We assume $\hat{\Sigma}_{ee}$ is a multiple of a Wishart matrix with d_e-degrees of freedom and approximate $\hat{\Sigma}_{\varphi\varphi} - \Sigma_{\varphi\varphi}$ with $(1-\varphi)(\hat{\Sigma}_{ee} - \Sigma_{ee})$. We have

$$(1-\varphi)^2 E\left\{\mathbf{L}\left(\hat{\Sigma}_{ee} - \Sigma_{ee}\right)\Sigma_{\varphi\varphi}^{-1}\Sigma_{zz}\Sigma_{\varphi\varphi}^{-1}\left(\hat{\Sigma}_{ee} - \Sigma_{ee}\right)\mathbf{L}'\right\}$$

$$= (1-\varphi)^2 d_e^{-1}\mathbf{L}\left[\Sigma_{ee}\,\mathrm{tr}\left\{\Sigma_{\varphi\varphi}^{-1}\Sigma_{zz}\Sigma_{\varphi\varphi}^{-1}\Sigma_{ee}\right\} + \Sigma_{ee}\Sigma_{\varphi\varphi}^{-1}\Sigma_{zz}\Sigma_{\varphi\varphi}^{-1}\Sigma_{ee}\right]\mathbf{L}'. \quad \text{(B.9)}$$

The dominant term is that associated with the trace and we retain only that term in our approximation. Thus, an approximation to the variance of $\hat{\beta}$ is

$$\mathbf{V}_{\beta\beta} = \mathbf{L}\Sigma_{zz}\mathbf{L}' + d_e^{-1}(1-\varphi)^2\mathrm{tr}\left\{\Sigma_{\varphi\varphi}^{-1}\Sigma_{zz}\Sigma_{\varphi\varphi}^{-1}\Sigma_{ee}\right\}\mathbf{L}\Sigma_{ee}\mathbf{L}' \quad \text{(B.10)}$$

Combining results (B.6), (B.7), and (B.9), a crude estimator of the variance of the predictor (31) is

$$\hat{\mathbf{V}}\{\hat{\mathbf{y}}_\varphi\} = \hat{\mathbf{H}}_\varphi'\hat{\Sigma}_{ww}\hat{\mathbf{H}}_\varphi + \hat{\mathbf{G}}_\varphi'\hat{\Sigma}_{ee}\hat{\mathbf{G}}_\varphi + \hat{\mathbf{H}}_\varphi'\mathbf{X}\hat{\mathbf{V}}_{\beta\beta}\mathbf{X}'\hat{\mathbf{H}}_\varphi' + \hat{\Gamma}_{44} + \hat{\Gamma}_{33}, \quad \text{(B.11)}$$

where

$\hat{\mathbf{H}}_\varphi = \mathbf{I} - \hat{\mathbf{G}}_\varphi$,

$$\hat{\mathbf{V}}_{\beta\beta} = \hat{\mathbf{L}}_\varphi\hat{\Sigma}_{zz}\hat{\mathbf{L}}_\varphi' + d_e^{-1}(1-\varphi)^2 \times \mathrm{tr}\left\{\hat{\Sigma}_{\varphi\varphi}^{-1}\hat{\Sigma}_{zz}\hat{\Sigma}_{\varphi\varphi}^{-1}\hat{\Sigma}_{ee}\right\}\hat{\mathbf{L}}_\varphi\hat{\Sigma}_{ee}(1+\delta_\varphi)\hat{\mathbf{L}}_\varphi',$$

$$\hat{\mathbf{L}}_\varphi = \left(\mathbf{X}'\hat{\mathbf{S}}_{\varphi\varphi}^{-1}\mathbf{X}'\right)^{-1}\mathbf{X}'\hat{\mathbf{S}}_{\varphi\varphi}^{-1},$$

$$\hat{\mathbf{S}}_{\varphi\varphi}^{-1} = \left(\hat{\Sigma}_{\varphi\varphi} + \delta_\varphi\hat{\Sigma}_{\varphi\varphi}\right)^{-1},$$

$$\hat{\Sigma}_{zz} = \hat{\Sigma}_{ww} + \hat{\Sigma}_{ee},$$

$$\delta_\varphi = \left[d_e - \mathrm{tr}\left\{\hat{\Sigma}_{zz}^{-1}\hat{\Sigma}_{ee}\right\}\right]^{-1}\mathrm{tr}\left\{\hat{\Sigma}_{zz}^{-1}\hat{\Sigma}_{ee}\right\},$$

$$\hat{\Gamma}_{44} = d_e^{-1}(1-\varphi)^2\mathrm{tr}\left\{\hat{\Sigma}_{\varphi\varphi}^{-1}\hat{\Sigma}_{zz}\hat{\Sigma}_{\varphi\varphi}^{-1}\hat{\Sigma}_{ee}\right\}\hat{\mathbf{G}}'\hat{\Sigma}_{ee}\hat{\mathbf{G}},$$

$$\hat{\Gamma}_{33} = \hat{\mathbf{H}}_\varphi'\begin{pmatrix} \hat{\Lambda}_{11}\hat{V}\{\hat{\sigma}_1^2\} & \hat{\Lambda}_{12}\hat{C}\{\hat{\sigma}_1^2, \hat{\sigma}_2^2\} \\ \hat{\Lambda}_{21}\hat{C}\{\hat{\sigma}_1^2, \hat{\sigma}_2^2\} & \hat{\Lambda}_{22}\hat{V}\{\hat{\sigma}_2^2\} \end{pmatrix}\hat{\mathbf{H}}_\varphi,$$

$$\hat{\Lambda} = \begin{pmatrix} \hat{\Lambda}_{11} & \hat{\Lambda}_{12} \\ \hat{\Lambda}_{21} & \hat{\Lambda}_{22} \end{pmatrix} = \hat{\Sigma}_{\varphi\varphi}^{-1}\hat{\Sigma}_{zz}\hat{\Sigma}_{\varphi\varphi}^{-1},$$

$\hat{V}\{\hat{\sigma}_j^2\}, j = 1,2$, is the estimated variance of $\hat{\sigma}_j^2$, and $\hat{C}\{\hat{\sigma}_1^2, \sigma_2^2\}$ is the estimated covariance between $\hat{\sigma}_1^2$ and $\hat{\sigma}_2^2$.

See Appendix A. The estimator of the variance of $\hat{\beta}$ contains an adjustment for the fact that $(\mathbf{X}'\hat{\Sigma}_{zz}^{-1}\mathbf{X})^{-1}$ is a biased estimator of $(\mathbf{X}'\Sigma_{zz}^{-1}\mathbf{X})^{-1}$.

REFERENCES

BATTESE, G.E., HARTER, R.M., and FULLER, W.A. (1988). An error components model for prediction of county crop areas using survey and satellite data. *Journal of the American Statistical Association*, 83, 28-36.

CRESSIE, N. (1992). REML Estimation in Empirical Bayes smoothing of census undercount. *Survey Methodology*, 18, 75-94.

EFRON, B., and MORRIS, C. (1972). Limiting the risk of Bayes and Empirical Bayes estimates - Part II: The Empirical Bayes case. *Journal of the American Statistical Association*, 67, 130-139.

ERICKSEN, E.P., and KADANE, J.B. (1985). Estimating the population in a census year (with discussion). *Journal of the American Statistical Association*, 80, 98-131.

ERICKSEN, E.P., KADANE, J.B., and TUKEY, J.W. (1989). Adjusting the 1980 Census of Population and Housing (with discussion). *Journal of the American Statistical Association*, 84, 927-944.

FAY, R.E. (1987). Application of multivariate regression to small domain estimation. In *Small Area Statistics*, (Eds. R. Platek, J.N.K. Rao, C.-E. Särndal and M.P. Singh). New York: Wiley, 91-102.

FAY, R.E. (1990). VPLX: Variance estimates for complex samples. *Proceedings of the Section on Survey Research Method, American Statistical Association*, 266-271.

FAY, R.E. (1992). Inferences for Small Domain Estimates From the 1990 Post Enumeration Survey. Unpublished manuscript, U.S. Bureau of the Census.

FAY, R.E., and HERRIOTT, 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.

FULLER, W.A., and HARTER, R.M. (1987). The multivariate components of variance model for small area estimation. In *Small Area Statistics*, (Eds., R. Platek, J.N.K. Rao, C.-E. Särndal and M.P. Singh), New York: Wiley, 103-123.

GHOSH, M. (1992). Hierarchical and Empirical Bayes multivariate estimation. In *Current Issues in Statistical Inference: Essays in Honor of D. Basu*, (Eds. M. Ghosh and P.K. Pathak), IMS Lecture Notes Monograph Series, 17, 151-177.

GHOSH, M., and RAO, J.N.K. (1994). Small area estimation: an appraisal. *Statistical Science*, 9, 55-93.

HARVILLE, D.A. (1976). Extension of the Gauss-Markov Theorem to include estimation of random effects. *Annals of Statistics*, 4, 384-395.

HENDERSON, C.R. (1950). Estimation of genetic parameters (Abstract). *Annals of Mathematical Statistics*, 21, 309-310.

HOGAN, H. (1992). The 1990 Post Enumeration Survey: an overview. *The American Statistician*, 46, 261-269.

HOGAN, H. (1993). The 1990 Post Enumeration Survey: operations and results. *Journal of the American Statistical Association*, 88, 1047-1060.

HULTING, F.L., and HARVILLE, D.A. (1991). Some Bayesian and non-Bayesian procedures for the analysis of comparative experiments and small area estimation: computational aspects, frequentist properties, and relationships. *Journal of the American Statistical Association*, 86, 557-568.

ISAKI, C.T., HUANG, E.T., and TSAY, J.H. (1991). Smoothing adjustment factors from the 1990 Post Enumeration Survey. *Proceedings of the Social Statistics Section, American Statistical Association*, 338-343.

KACKAR, R.N., and HARVILLE, D.A. (1984). Approximations for standard errors of estimators for fixed and random effects in mixed models. *Journal of the American Statistical Association*, 79, 853-862.

MORRIS, C. (1983). Parametric Empirical Bayes inference: theory and applications (with discussions). *Journal of the American Statistical Association*, 78, 47-65.

PEIXOTO, J.L., and HARVILLE, D.A. (1986). Comparisons of alternative predictors under the balanced one-way random model. *Journal of the American Statistical Association*, 81, 431-436.

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.

ROBINSON, G.K. (1991). That BLUP is a good thing: the estimation of random effects. *Statistical Science*, 6, 15-51.

SEARLE, S.R. (1971). *Linear Models*. New York: Wiley.

SINGH, M.P., GAMBINO, J., and MANTEL, H.J. (1994). Issues and strategies for small area data. *Survey Methodology*, 20, 3-14.

U.S. BUREAU OF THE CENSUS. (1988). The Coverage of Population in the 1980 Census, Evaluation and Research Program, PHC(E)-4.

Survey Methodology, June 2000
Vol. 26, No. 1, pp. 43-52
Statistics Canada

Census Coverage Error: A Demographic Evaluation

RÉJEAN LACHAPELLE and DON KERR[1]

ABSTRACT

The 1996 Canadian Census is adjusted for coverage error as estimated primarily through the Reverse Record Check (RRC). In this paper, we will show how there is a wealth of additional information from the 1996 Reverse Record Check of direct value to population estimation. Beyond its ability to estimate coverage error, it is possible to extend the Reverse Record Check classification results to obtain an alternative estimate of demographic growth – potentially decomposed by component. This added feature of the Reverse Record Check provides promise in the evaluation of estimated census coverage error as well as insight as to possible problems in the estimation of selected components in the population estimates program.

KEY WORDS: Census coverage error; Population estimates; Reverse record check.

1. INTRODUCTION

The Reverse Record Check (RRC), in various forms, has been used by Statistics Canada since the 1960's to estimate coverage error in the Canadian Census (Fellegi 1969; Brackstone and Gosselin 1973; Gosselin 1976; Burgess 1988; Carter 1990; Royce, Germain, Julien, Dick, Switzer and Allard 1994, Statistics Canada 1999). Using the Reverse Record Check, Statistics Canada has produced a long time series of population estimates, from 1971 through to the present, fully adjusted for census undercount. The current paper will demonstrate how there is additional information in the Reverse Record Check, which from a demographic perspective, can be exploited for the purposes of population estimation.

The demographic statistics program at Statistics Canada uses information from vital statistics, the most recent census, and various administrative sources in generating highly accurate and up to date population estimates. Information on births, deaths, immigration, emigration, among other demographic components, can be used to estimate population growth since the previous census. With each quinquennial census, a cycle ends and the accuracy of these estimates are put to the test (Romaniuc 1988). Systematic comparisons can be made between these estimates of growth and estimated growth as implied by comparing subsequent censuses (after adjustment for census coverage error).

The resultant difference (conventionally referred to as the error of closure of the intercensal population estimates) has a far from obvious interpretation. While a large error of closure is suggestive of problems in the population estimates, its specific nature is far from obvious (as to which demographic components are specifically responsible for the error). Furthermore, a honest appraisal of this closure error might suggest not only problems in the population estimates, but also potential problems in census coverage studies themselves (at the beginning and/or end of the intercensal period).

The current paper will demonstrate how an alternative estimate of demographic growth is possible, as based explicitly on the RRC classification results. Additional information is available, which assists greatly in the interpretation and decomposition of this closure error. Three alternative estimates of demographic growth for the intercensal period will be presented in the following section, including growth as estimated as part of the regular program of population estimates, implicit growth as obtained in comparing consecutive censuses, and growth as based explicitly on RRC classification results. Section 3 demonstrates how this RRC based estimate of growth can assist in the decomposition and interpretation of closure error, providing evidence of (i) bias in selected components of the population estimates, and (ii) possible problems in the RRC results. Section 4 presents the results from this decomposition, followed by a brief discussion of its implications for both census coverage error measurement and the population estimates program.

2. ALTERNATIVE ESTIMATES OF DEMOGRAPHIC GROWTH

2.1. Administrative Record Based Estimates of Growth: Post-Censal Estimates

Statistics Canada's regular program of population estimates involves the continuous registration and estimation of demographic events, as based on vital statistics and various administrative data sets. These events are added or subtracted from the population documented in the previous census (component method). In estimating a province's population on Census day 1996 (P_{est96}):

[1] Réjean Lachapelle, Demography Division, Main Building, Tunney's Pasture, Statistics Canada, Ottawa, Ontario, K1A 0T6; Don Kerr, Department of Sociology, University of Western Ontario, London, Ontario, N6A 5C2.

$$P_{est96} = P_{91} + B_{91\text{-}96} - D_{91\text{-}96} + I_{91\text{-}96} - E_{91\text{-}96} + \Delta NPR_{91\text{-}96} + NM_{91\text{-}96}. \quad (1)$$

The baseline population (P_{91}) for this estimate builds on the 1991 Census after adjustment for all forms of coverage error, including net census undercount as measured through the 1991 RRC. The postcensal estimate can be obtained by adding or subtracting from this baseline the number of births between censuses ($B_{91\text{-}96}$), the number of deaths ($D_{91\text{-}96}$), immigrants ($I_{91\text{-}96}$), emigrants ($E_{91\text{-}96}$), net interprovincial migration ($NM_{91\text{-}96}$), and the net gain or loss of nonpermanent residents ($\Delta NPR_{91\text{-}96}$).

Non-permanent residents (NPRs) are persons with legal temporary status in Canada (*e.g.*, persons holding student or employment authorizations, minister's permits, refugee claimants, as well as their non-Canadian born dependents). Unlike with interprovincial migration, net gain or net loss of NPRs is not estimated through "flow" data on the ongoing in and out-flows of non-permanent residents, but rather estimated by comparing over time "stock" data on the total number of non-permanent residents living in the country. Further details of methodology, data sources and data quality issues can be obtained from the quarterly and annual releases of the population estimates program (Statistics Canada 1999; 2000).

2.2. Implicit Estimate of Growth

An implicit estimate of growth can be derived using the 1991 and 1996 Censuses, with both censuses adjusted for net undercount. With the exception of a small number of refusal Indian reserves, whose population figures are estimated independently, gross undercoverage was estimated entirely through the RRC in 1996, whereas gross overcoverage was a combined estimate from three studies (the RRC, the Collective Dwellings Study and the Automated Match Study). In 1991, the RRC was used only in the estimation of gross undercoverage, whereas gross overcoverage was estimated through a smaller study, the Private Dwelling Study, in combination with the 1991 Collective Dwelling and Automated Match studies. In addition, persons missed on refusal Indian reserves were estimated as part of the 1991 Reverse Record Check.

In the early evaluation of the 1996 coverage studies, the implicit growth obtained with the above adjustments was considered unrealistic. It has since been established that part of the 1991 estimate of net undercount was in error, and would have in reality been lower had selected methodological enhancements been introduced as in 1996 (Tourigny, Clark and Provost 1998). It has been shown that (i) a number of persons initially classified as missed in 1991 was too high due to misclassification, and (ii) the 1991 estimate of "overcount" was too low. As a result, 1991 estimates of undercount and overcount have been revised to reflect the impact of these methodological changes ($^{rev}U_{91}$, $^{rev}O_{91}$). In addition, for reasons of consistency with 1996, separate modeled estimates of refusal Indian reserves (independent of the RRC) have been added to the Census in 1991.

More specifically, implicit growth (Δ^I) is obtained as:

$$\Delta^I = P_{96} - P_{91} = \{P_{96}^c + U_{96} - O_{96} + IR_{96M} - IR_{RRC96}\} - \{P_{91}^c + {}^{rev}U_{91} - {}^{rev}O_{91} + IR_{91M} - IR_{RRC91}\} \quad (2)$$

where final population figures (P_{96}, P_{91}) are obtained using previously published census figures (P_{96}^c, P_{91}^c) adjusted for undercoverage (U_{96}, $^{rev}U_{91}$) and gross overcoverage (O_{96}, $^{rev}O_{91}$). In adding independently modeled estimates of refusal Indian reserves (IR_{96M}, IR_{91M}), it is necessary to remove that portion of the RRC estimate of gross undercoverage that corresponds to these reserves (IR_{RRC96}, IR_{RRC91}). The results presented in the current paper take these changes into consideration.

2.3.1. RRC Based Estimates of Growth

The Reverse Record Check (RRC) is a record linkage and matching procedure that attempts to trace all persons in its sample, interview them to obtain a census day address, and match their records to individual census documents. This involves the construction of a sample intended to represent the same target population as the census being evaluated. This sampling frame, obtained in a manner that is totally independent of the census being evaluated, is constructed using the previous census, birth registrations over the intercensal period, administrative lists of intercensal immigrants, and an up-to-date listings of non-permanent residents. Persons missed in the previous census are represented by a sample of cases classified as "missed" in the previous RRC, in the absence of a complete list of such persons.

By working with this sample, the RRC targets all persons who could have potentially been part of the 1996 Census universe. Except for a very small sub-population of returning emigrants (Canadian citizens and landed immigrants who were abroad during the previous census), the RRC sample is complete and fully representative. The subsequent classification (missed, enumerated, emigrated, abroad, deceased or out of scope) is applied in the estimation of "missed" in the current census. At the same time, this classification also holds the potential for further inferences, *i.e.*, an additional estimate of demographic growth for the intercensal period.

To estimate demographic growth using the RRC, it is useful to consider the following two equations. In the first equation, the target population of the 1991 Census (P_{91}^T) is expressed in terms of all potential classification outcomes in 1996. In the second equation, it is possible to move in the opposite direction – by expressing the 1996 census target population (P_{96}^T) in terms of all possible statuses in 1991 (or in the case of births and immigrants, the intercensal period).

$$P_{91}^{T} = {}^{91}PP_{96} + {}^{91}NP_{96} + {}^{91NP}C_{96PP} + {}^{91PP}D_{96} + {}^{91NP}D_{96} + {}^{91PP}E_{96FR} + {}^{91NP}E_{96FR} + {}^{91NP}E_{96EX} \quad (3)$$

$$P_{96}^{T} = {}^{91}PP_{96} + {}^{91}NP_{96} + {}^{91NP}C_{96PP} + {}^{91-96}B_{96} + {}^{91EX}I_{96PP} + {}^{91EX}I_{96NP} + {}^{91FR}RE_{96PP} \quad (4)$$

where:

- ${}^{91}PP_{96}$ - Canadian citizens and landed immigrants in Canada in 1991, also targeted by the 1996 census
- ${}^{91}NP_{96}$ - NPRs in Canada in 1991, also targeted by the 1996 census as NPRs
- ${}^{91NP}C_{96PP}$ - NPRs in Canada in 1991 who became landed immigrants over the intercensal period
- ${}^{91PP}D_{96}$ - Canadian citizens and landed immigrants in Canada in 1991, who died over the intercensal period
- ${}^{91NP}D_{96}$ - NPRs in Canada in 1991, who died over the intercensal period
- *FR* - persons with the right to live permanently in Canada (citizens and landed immigrants) that are not in the designated census target population
- ${}^{91PP}E_{96FR}$ - Canadian citizens and landed immigrants in Canada in 1991, who are outside the 1996 Census target population
- ${}^{91NP}E_{96FR}$ - NPRs in Canada in 1991, who became landed immigrants or citizens, and are outside the 1996 census target population
- *EX* - persons who have never been citizens or landed immigrants, and are not in the designated census target population
- ${}^{91NP}E_{96EX}$ - NPRs in Canada in 1991, who did not become landed immigrants, and are outside the 1996 census target population
- ${}^{91-96}B_{96}$ - births over the 1991-1996 period, and in the 1996 census target population
- ${}^{91EX}I_{96NP}$ - persons not in Canada in 1991, who arrived over the intercensal period, and are NPRs in the 1996 census target population
- ${}^{91EX}I_{96PP}$ - immigrants who landed over the intercensal period, and are in the 1996 Census target population
- ${}^{91FR}RE_{96PP}$ - returning emigrants, *i.e.*, Canadian citizens and landed immigrants outside the census universe in 1991, and in the 1996 Census universe

An estimate of growth (Δ^{RRC}) can be obtained by subtracting the former equation from the latter:

$$\Delta^{RRC} = {}^{91-96}B_{96} + {}^{91EX}I_{96PP} + {}^{91EX}I_{96NP} - {}^{91PP}D_{96} - {}^{91NP}D_{96} - {}^{91PP}E_{96FR} - {}^{91NP}E_{96FR} - {}^{91NP}E_{96EX} + {}^{91FR}RE_{96PP}. \quad (5)$$

With the previously introduced sampling frames and classification outcomes, all terms (with the exception of the last term: returning emigrants) can be directly estimated from the 1996 RRC itself. The census target population in 1991 can be approximated through the sample drawn from the census and missed frames – with the identification of relevant classification outcomes. The census target population in 1996 can be approximated through all persons classified as either enumerated or missed in 1996. The final term (*i.e.*, returning emigrants) can be obtained independent of the RRC using the 1996 Census 5-year mobility variable, in identifying all persons outside the country five years ago (excluding recent immigrants and NPRs). It is possible to express this same RRC based estimate of demographic growth at the provincial level by incorporating an estimate of interprovincial migration. As the RRC relied on Health Care Files in Canada's two northern territories (the Yukon and NWT) with administrative lists of addresses current to census being evaluated, this estimate of growth is not possible for the relatively small populations living in Canada's far north.

A minor problem in the RRC design persists that potentially introduces a slight bias into its classification results. Unfortunately it is not possible to identify all NPRs in the RRC sample, with the potential for an unknown amount of frame overlap (*i.e.*, between the census, NPR and immigrant frames). As NPRs in the census can only be identified through the census long form (which is distributed to about 20% of all households), it is possible that some NPRs living in Canada in 1991, selected in the census frame, were also selected in either the immigrant or NPR frames (without being identified as such). While the RRC attempts to adjust for this overlap by identifying all such persons in the immigrant and NPR frames, an unknown bias exists to the extent that this is unsuccessful. This difficulty in identifying overlap leaves the potential of too many immigrants and/or NPRs in the sample, or too few, if too many persons are removed from the aforementioned frames. The latter outcome can subsequently deflate the estimate of demographic growth, gross undercoverage (among other classification outcomes), whereas the former has the opposite outcome.

2.3.2. RRC based Estimate of Growth: A More Detailed Decomposition

While both the postcensal and RRC based estimates of demographic growth should be highly comparable, the specific terms within each are not meant to be directly

equivalent. For example, births in the postcensal estimates denote all intercensal births occurring to a population – irrespective of whether such births move or die – whereas births in the discrete equation denote all births occurring yet still in Canada at the end of the intercensal period. With this in mind, it is possible to expand on the RRC based equation, to derive terms that are more comparable to those used in the postcensal estimates. The RRC based estimate of demographic growth can then be used in the evaluation of the components of demographic growth that enter into the component method.

To expand on this equation, it is useful to begin with births, again expressed in terms of possible RRC classification outcomes. As previously indicated, the birth term as included in equation (5) is only part of all births occurring over the intercensal period. More comprehensively, all births can be expressed as:

$$B^{91\text{-}96} = {}^{91\text{-}96}B_{96} + {}^{91\text{-}96B}D_{96} + {}^{91\text{-}96B}E_{96FR} \qquad (6)$$

where:

$B^{91\text{-}96}$ = all intercensal births
${}^{91\text{-}96}B_{96}$ = all intercensal births ultimately classified as either enumerated or missed in 1996
${}^{91\text{-}96B}D_{96}$ = deaths of intercensal births
${}^{91\text{-}96B}E_{96FR}$ = persons outside target population in 1996 yet born in Canada over the intercensal period

Similarly, all immigrants can be expressed as:

$$I^{91\text{-}96} = {}^{91EX}I_{96PP} + {}^{91NP}C_{96PP} + {}^{91\text{-}96I}D_{96} + {}^{91\text{-}96I}E_{96FR} \qquad (7)$$

where:

${}^{91EX}I_{96PP}$ = intercensal immigrants ultimately classified as either enumerated or missed in 1996
${}^{91NP}C_{96PP}$ = all NPRs in 1991 who obtain landed immigrant status and are ultimately classified as either enumerated or missed in 1996
${}^{91\text{-}96I}D_{96}$ = deaths occurring to landed immigrants over the intercensal period
${}^{91\text{-}96I}E_{96FR}$ = emigrants among intercensal immigrants (irrespective of whether or not they were living in Canada as NPRs in 1991)

In combining equations 5, 6 and 7, demographic growth can be restated as:

$$\begin{aligned} P_{96}^{T} - P_{91}^{T} = {} & B^{91\text{-}96} - {}^{91PP}D_{96} - {}^{91NP}D_{96} - {}^{91\text{-}96B}D_{96} - \\ & {}^{91\text{-}96I}D_{96} + I^{91\text{-}96} + {}^{91FR}RE_{96PP} - {}^{91NP}C_{96PP} - \\ & {}^{91PP}E_{96FR} - {}^{91NP}E_{96FR} - {}^{91\text{-}96B}E_{96FR} - {}^{91\text{-}96I}E_{96FR} - \\ & {}^{91NP}E_{96EX} - {}^{91EX}I_{96NP}. \end{aligned} \qquad (8)$$

Given that the final term of (8) is equivalent to:

$${}^{91EX}I_{96NP} = NP_{96} - NP_{91} + {}^{91NP}D_{96} + {}^{91NP}E_{96EX} + {}^{91NP}C_{96PP} + {}^{91NP}E_{96FR}. \qquad (9)$$

It follows that:

$$\begin{aligned} P_{96}^{T} - P_{91}^{T} = {} & B^{91\text{-}96} - {}^{91PP}D_{96} - {}^{91NP}D_{96} - {}^{91\text{-}96B}D_{96} - {}^{91\text{-}96I}D_{96} \\ & + I^{91\text{-}96} + {}^{91FR}RE_{96PP} - {}^{91NP}C_{96PP} - \\ & {}^{91PP}E_{96FR} - {}^{91NP}E_{96FR} - {}^{91\text{-}96B}E_{96FR} - \\ & {}^{91\text{-}96I}E_{96FR} - {}^{91NP}E_{96EX} + NP_{96} - (NP_{91} - {}^{91NP}D_{96} - \\ & {}^{91NP}E_{96EX} - {}^{91NP}C_{96PP} - {}^{91NP}E_{96FR}) \end{aligned} \qquad (10)$$

or:

$$\begin{aligned} P_{96}^{T} - P_{91}^{T} = {} & \left(B^{91\text{-}96}\right) - \left({}^{91PP}D_{96} - {}^{91\text{-}96B}D_{96} - {}^{91\text{-}96I}D_{96}\right) + \left(I^{91\text{-}96}\right) - \\ & \left({}^{91PP}E_{96FR} + {}^{91\text{-}96B}E_{96FR} + {}^{91\text{-}96I}E_{96FR} - {}^{91FR}RE_{96PP}\right) + \\ & \left(NP_{96} - NP_{91}\right). \end{aligned} \qquad (11)$$

This expanded version of equation (5) provides a breakdown of demographic growth at the national level, and allows for more meaningful comparisons with components estimated through administrative records. All terms, except for ${}^{91FR}RE_{96PP}$ and NP_{91} can be obtained directly from the 1996 RRC. The aforementioned hole in the RRC sampling frame requires an independent estimate of returning emigrants whereas the nature of the sample frame for NPRs explains the absence of the latter term. Rather than a listing of all NPRs to enter Canada over the intercensal period (as was the case with immigrants), the RRC relies on the most up to date administrative listing of NPRs in the establishment of its sampling frame (with no information on the number of NPRs living in Canada in 1991).

Postcensal estimates document demographic growth through the "continuous" registration and estimation of demographic events over time. The RRC estimates growth via information on the status of individuals as identified on at least two "discrete" dates (at the beginning and end of the intercensal period). Irrespective of this minor conceptual distinction between "continuous" versus "discrete" estimation, each term of equation 11 (within each set of parenthesis) roughly corresponds to a separate component as documented using administrative records. The first term identifies all intercensal births (*i.e.*, the weighted sum of the birth frame), the second term includes deaths (classification results across the birth frame, the missed frame, the census frame and immigrant frame), the third term includes all

immigrants (*i.e.*, the weighted sum of the immigrant frame), the fourth term includes emigrants (classification results across the birth frame, the immigrant frame, the missed frame and census frames, as well as the returning emigrant component), and the fifth term corresponds to net gain or loss of NPRs. As the number of NPRs living in Canada in 1991 is not available in the 1996 RRC, for current purposes, this latter term is obtained using the 1991 census count, after adjustments for undercoverage. Again, it is possible to express this equation at the provincial level.

With equation (11), a detailed evaluation of the postcensal estimation program is possible. For example, if differences persist between RRC based estimates and postcensal estimates, it is possible to determine how much of the difference in estimated growth can be traced back to differences in migration (typically estimated with some difficulty in the postcensal estimates program) and how much can be traced to differences in natural increase. Briefly, Table 1 includes all of the aforementioned estimates of growth, including implicit growth, the growth as based on administrative records, and the two alternate estimates of growth as based on the RRC (simplified and expanded equations). Slight differences exist between the simplified and expanded equations – yet not nearly of the same size as with the other estimates (implicit, postcensal). In explanation of the differences between the two RRC based estimates, the simplified equation does not require the same detailed classification as with the expanded equation, is not biased to the same extent by the aforementioned problem of frame overlap, and does not rely on the 1991 census count of NPRs. The differences observed with the remaining estimates are the focus of the current decomposition.

Table 1
Alternate Estimates of Growth, 1991-1996, Canada and Provinces/Territories

	Implicit Growth	Population Estimates Administrative records	RRC simplified	RRC expanded
NFLD.	-17,997	-9,263	-17,897	-17,751
P.E.I.	5,404	5,483	2,568	1,583
N.S.	15,781	24,271	17,075	16,860
N.B.	7,714	13,097	12,017	11,276
QUE.	206,307	300,849	261,357	252,014
ONT.	659,349	766,568	668,443	655,572
MAN.	23,682	24,981	7,377	6,288
SASK.	15,953	11,098	11,524	9,312
ALTA.	186,594	186,986	151,944	159,907
B.C.	505,025	466,611	465,864	472,342
YUKON	3,085	2,329	N/A	N/A
N.W.T.	6,837	5,864	N/A	N/A
Provinces (excl terr)	1,607,771	1,790,681	1,580,273	1,567,404
Canada	1,617,693	1,798,874	N/A	N/A

3. A DECOMPOSITION OF CLOSURE ERROR

Implicit growth for the 1991-96 period is obtained only after all adjustments have been made to the censuses for coverage error, including revised 1991 figures on gross undercount and overcount and refinements for refusal Indian reserves. Alternatively, the RRC based estimate of growth (simplified version) is obtained in working with approximations of the 1991 and 1996 target populations, *i.e.*, the census and missed frames of the 1996 RRC and all persons classified as either missed or enumerated in this study. For this reason, there are minor differences between the two estimates that need to be more clearly identified in a full decomposition of closure error. In this context, it is useful to express implicit growth obtained with final population figures in terms of these approximations (sampling frames and classification outcomes). In a similar manner, as the error of closure is the difference between implicit growth and the growth associated with the postcensal estimates, the error of closure can also be expressed in terms of these approximations.

To simplify the presentation, let δ represent all possible negative growth terms in equation (5) and η as all possible positive growth terms:

$$\delta = ({}^{91PP}D_{96} + {}^{91NP}D_{96} + {}^{91PP}E_{96FR} + {}^{91NP}E_{96FR} + {}^{91NP}E_{96EX}) \quad (12)$$

$$\eta = ({}^{91\text{-}96}B_{96} + {}^{91EX}I_{96PP} + {}^{91EX}I_{96NP} + {}^{91FR}RE_{96PP}) \quad (13)$$

The population enumerated in both censuses can be represented as:

$${}^{91}P_{96} = \left({}^{91}PP_{96} + {}^{91}NP_{96} + {}^{91NP}C_{96PP}\right). \quad (14)$$

Since the final population figures (P_{91}, P_{96}) used in the estimation of implicit growth involve separate modeled estimates for refusal Indian reserves, it is useful to restate the RRC based estimate of growth after specifically delineating such reserves. In designating persons living in refusal reserves in 1996 that were in the target population in 1991 as ${}^{91}IR_{96}$, the growth of these reserves through either migration or birth as estimated by the RRC by η_{IR}, and redefining ${}^{91}P_{96}$ to exclude all persons associated with these two terms, it is possible to return to equations (3)-(5) as:

$$P_{91}^{T} = {}^{91}P_{96} + {}^{91}IR_{96} + \delta \quad (15)$$

$$P_{96}^{T} = {}^{91}P_{96} + {}^{91}IR_{96} + \eta_{IR} + \eta \quad (16)$$

$$\Delta^{RRC} = P_{96} - P_{91} = \eta + \eta_{IR} - \delta \quad (17)$$

In expressing implicit growth in terms of the RRC sampling frames and classification outcomes, it is useful to build on the RRC estimate of growth (equation 17) in defining total growth beginning with P_{91} rather than P_{91}^{T}. In recognition that the final population estimate (P_{91}) is equivalent to the census and missed frames (P_{91}^{T}) minus overcoverage ($^{rev}O_{91}$) plus refinements for refusal Indian reserves ($IR_{91M} - IR_{RRC91}$), it follows:

$$P_{96}^{T} - P_{91} = \eta + n_{IR} - \delta + {}^{rev}O_{91} + (IR_{RRC91} - IR_{91M}). \quad (18)$$

On the other hand, this target population (P_{96}^{T}) can also be expressed as:

$$P_{96}^{T} = EN_{96} + U_{96} + {}^{91FR}RE_{96PP} \quad (19)$$

where EN_{96} is an estimate of the number of persons enumerated in 1996. In recalling from equation 2 that:

$$P_{96} = P_{96}^{c} + U_{96} - O_{96} + (IR_{96M} - IR_{RRC96}) \quad (20)$$

implicit growth (Δ^{I}) can be expressed in terms of the RRC based estimates of growth, as:

$$\Delta^{I} = P_{96} - P_{91} = (P_{96} - P_{96}^{T}) + (P_{96}^{T} - P_{91}) =$$
$$\{(\eta - \delta)\} + \{\eta_{IR} - (IR_{91M} - IR_{RRC91}) + (IR_{96M} - IR_{RRC96})\} +$$
$$\{(P_{96}^{c} - EN_{96} - {}^{91FR}RE_{96PP} + {}^{rev}O_{91} - O_{96})\}. \quad (21)$$

Implicit growth (Δ^{I}) can be defined as the sum of (i) a RRC based estimate of growth (excluding refusal Indian reserves), (ii) a second term depending on the decision to estimate the refusal Indian reserves by an independent model, and (iii) a third term that involves a comparison of the RRC based estimate of enumerated and the number of persons actually enumerated in the 1996 census.

This latter term (the difference on enumerated) has an interesting interpretation, and is considered fundamental to the evaluation of the RRC (Tourigny, Bureau and Clark 1998; Royce 1993). Significant differences with this term can be read as implying either sampling errors and/or possible biases, as either classification error and/or problems in sample selection. To make this comparison meaningful, 1996 overcoverage and an estimate of returning emigrants are removed from the census counts - as neither can be included in the estimate of enumerated. Similarly, since the RRC selects part of its sample from the previous census, it inevitably carries forward some overcoverage inherent in its weights – which must subsequently be removed from its estimate of enumerated. These adjustments are included in the third term (the third set of brackets) in equation 21.

While the estimate of enumerated is inflated by the weights associated with overcoverage in the 1991 Census frame, only a portion is directly associated with this estimate – with the remainder spread across the other classification outcomes. Consequently, all classification results in the aforementioned equations are also slightly overstated. For the purposes of the current decomposition, this minor distinction is ignored. This is another reason, albeit of minor impact, why the RRC-based estimate of growth is different from the implicit estimate, as the latter is not biased by this overcoverage.

From the above, the error of closure is equivalent to:

$$\Delta_{91-96}^{D} - \Delta_{91-96}^{I} =$$
$$[\Delta_{91-96}^{D} - \{(\eta - \delta)\} -$$
$$\{\eta_{IR} - (IR_{91M} - IR_{RRC91}) + (IR_{96M} - IR_{RRC96})\}] -$$
$$[\{(P_{96}^{c} - EN_{96} - {}^{91FR}RE_{96PP} + {}^{rev}O_{91} - O_{96})\}]. \quad (22)$$

In the decomposition of closure error, the first term inside brackets [ ] highlights the difference between the postcensal estimate of growth and the combined RRC estimate of growth (including refusal reserves, after refinements for modeled estimates). The second term (the difference on enumerated) provides evidence as to possible difficulties in the coverage studies. Theoretically, with the absence of sampling and non-sampling error in the RRC, this latter term should be negligible.

4.1. Decomposition Results: Closure Error

Table 2 presents closure error after finalizing both the 1991 and 1996 estimates of population. By adding net undercount to the 1996 published Census figures, along with independent estimates of refusal Indian reserves, Canada's 1996 Census day population, adjusted for coverage error is estimated at 29,619,539. This figure is appreciably lower than the Census day estimate as generated through the postcensal estimates program of 29,800,720. The difference between the two figures – which is equivalent to the aforementioned difference between implicit growth and growth as based on administrative records – was higher than anticipated given past experience, at 181,181 (or .61% of the 1991 Census Day population).

Across provinces/territories, closure error is found to be particularly pronounced in Newfoundland (1.56%), in Canada's north (at -2.38% in Yukon and -1.44% in the NWT), and somewhat surprisingly, in its three largest provinces (as 1.30% in Quebec, .97% in Ontario and -.99% in British Columbia). Regionally, closure errors larger than the national average are observed across eastern and central Canada (except for P.E.I.) while the western provinces have closure errors lower than the national one. It is specifically these errors that the current decomposition seek to evaluate and explain.

Table 3 presents the results from this decomposition, with closure error decomposed into (i) the difference

between the estimate of growth based on administrative records and the RRC based estimate (simplified version), and (ii) the difference on enumerated. Also included is the sampling error associated with the RRC estimates.

4.2. Comparisons between Estimates of Growth

Across all provinces (with the exception of Saskatchewan), growth estimated on the basis of administrative records is higher than the RRC based estimate. At the national level (excluding the territories), this discrepancy on growth (210,408) appears far more important in explaining closure error than the discrepancy on enumerated (-27,498). While for many provinces the difference on growth fell well within expectations in light of sampling error, selected provinces require further explanation. For example, the difference in growth in Ontario is large (98,125), which is almost one half the difference observed

Table 2
Coverage Study Results, Relative to Population Estimate (1996 - Census Day)

	{1}	{2}	{3}	{4=1+2+3}	{5}	{6=5-4}	{7=6/4*100}
	1996 census count with random additions	1996 net undercount	Indian Reserves	1996 Census RRC adjusted	1996 estimate post-censal (i)	Error of closure	Error of closure (%)
NFLD.	551,792	9,424	0	561,216	569,950	8,734	1.56
P.E.I.	134,557	1,149	175	135,881	135,960	79	0.06
N.S.	909,282	20,821	0	930,103	938,593	8,490	0.91
N.B.	738,133	14,225	518	752,876	758,259	5,383	0.71
QUE.	7,138,795	116,750	12,427	7,267,972	7,362,514	94,542	1.30
ONT.	10,753,573	301,368	20,849	11,075,790	11,183,050	107,260	0.97
MAN.	1,113,898	18,881	315	1,133,094	1,134,393	1,299	0.11
SASK.	990,237	28,051	586	1,018,874	1,014,019	-4,855	-0.48
ALTA.	2,696,826	66,327	11,287	2,774,440	2,774,832	392	0.01
B.C.	3,724,500	142,443	3,136	3,870,079	3,831,665	-38,414	-0.99
YUKON	30,766	1,022	0	31,788	31,032	-756	-2.38
N.W.T.	64,402	3,024	0	67,426	66,453	-973	-1.44
Canada	28,846,761	723,485	49,293	29,619,539	29,800,720	181,181	0.61

(i) Post-Censal Estimates for May 14th, obtained with final components for intercensal estimates. Final Estimates (Sept. 24th, 1998) of Net Undercount, 1991 an 1996.

Table 3
Decomposition of Closure Error

Province/Territory	Error of Closure	Difference between		Difference on enumerated	S.E. of estimates
		Dem. and RRC Estimates of Growth	S.E. of estimates		
NFLD.	8,734	8,634	4,889	100	5,176
P.E.I.	79	2,915	2,425	-2,836	2,462
N.S.	8,490	7,196	9,011	1,294	9,455
N.B.	5,383	1,080	7,793	4,303	7,918
QUE.	94,542	39,492	25,493	55,050	29,310
ONT.	107,260	98,125	41,212	9,135	51,300
MAN.	1,299	17,604	10,108	-16,305	10,370
SASK.	-4,855	-426	9,187	-4,429	10,200
ALTA.	392	35,042	19,067	-34,650	21,618
B.C.	-38,414	747	20,518	-39,161	22,996
YUKON	-756	N/A	N/A	-108	270
N.W.T.	-973	N/A	N/A	-284	464
Canada without Territories	182,910	210,408	43,951	-27,498	58,724
Canada	181,181	N/A	N/A	-27,890	58,762

Table 4
Estimated Components (1991-1996) as Compiled by Demography Division and RRC Discrete (detailed) Measurement

	NFLD	PEI	NS	NB	QUE	ONT	MAN	SASK	ALB	BC	CANADA (without terr)
Births											
Demography	31,748	8,803	55,994	44,444	453,556	730,520	81,485	70,382	199,484	229,511	1,905,927
RRC	31,779	8,782	55,984	44,444	454,332	729,744	81,485	70,382	199,484	229,511	1,905,927
Difference	-31	22	10	0	-776	776	0	0	0	0	0
Deaths											
Demography	-19,286	-5,692	-37,677	-28,567	-252,628	-376,760	-45,858	-40,652	-75,798	-126,935	-1,009,853
RRC	-18,530	-6,913	-43,820	-29,354	-273,617	-400,047	-56,108	-40,143	-74,640	-138,433	-1,081,605
Difference	-756	1,221	6,143	787	20,989	23,287	10,250	-509	-1,158	11,498	71,752
Immigration											
Demography	3,411	771	14,489	3,359	189,905	618,869	22,004	11,282	84,130	213,506	1,161,726
RRC	3,538	820	14,058	3,614	189,905	618,870	22,129	11,157	84,130	216,892	1,165,113
Difference	-127	-49	431	-255	0	-1	-125	125	0	-3,386	-3,387
Emigration											
Demography	-671	-206	-2,297	-2,429	-15,490	-48,609	-5,684	-2,493	-19,718	-17,834	-115,431
RRC	-2,227	-455	-7334	-3,889	-55,766	-168,556	-10,871	-7,133	-33,689	-31,739	-321,659
Difference	1,556	249	5,037	1,460	40,276	119,947	5,187	4,640	13,971	13,905	206,228
Interprovincial Migration											
Demography	-23,074	1,643	-5,288	-3,255	-51,176	-40,850	-25,336	-26,644	7,155	167,809	984
RRC	-32,767	-886	-1,479	-2,933	-49,395	-37,505	-29,765	-25,095	-10,321	191,222	1,076
Difference	9,693	2,529	-3,809	-322	-1,781	-3,345	4,429	-1,549	17,476	-23,413	-92
Non-permanent Residents											
Demography	-1,406	164	-950	-455	-23,353	-116,602	-1630	-777	-8,267	554	-152,722
RRC	455	236	-549	-606	-13,445	-86,934	-582	144	-5,057	4,890	-101,448
Difference	-1,861	-72	-401	151	-9,908	-29,668	-1,048	-921	-3,210	-4,336	-51,274
Total											
Demography	-9,263	5,483	24,271	13,097	300,849	766,568	24,981	11,098	186,986	466,611	1,790,681
RRC	-17,751	1,583	16,860	11,276	252,014	655,572	6,288	9,312	159,907	472,343	1,567,404
Difference	8,488	3,900	7,411	1,821	48,835	110,996	18,693	1,786	27,079	-5,731	223,277

nationally. Similarly, Newfoundland, Quebec, Alberta and Manitoba, together explain a large part of this difference.

In providing some indication as to the factors responsible for these differences, Table 4 presents comparisons using equation 11 (detailed equation). Alternative estimates are provided on (i) births, (ii) deaths, (iii) immigration, (iv) emigration, (v) interprovincial migration and (vi) net change in the number of non-permanent residents. The most important problems in the explanation of closure error are obvious in Table 4, with specific reference to emigration. As Canada does not have a complete border registration system, emigration is clearly the weakest of all the components to enter into the population estimate program. Without access to direct information on the number of persons leaving Canada, the RRC, with its exhaustive tracing, record linkage and direct interviewing procedures, is considered an improvement over any other data sources currently available. Although there are known problems in the RRC (for example, the previously mentioned frame overlap), the current evaluation points to an obvious error in the postcensal estimates, *i.e.*, an understatement of population outflow from Canada. Overall, the difference as observed nationally (206,228) explains the bulk of the closure error documented in 1996. Similarly with Ontario, difficulties in the estimation of emigration appear to be fundamental (with a difference of fully 119,947).

Without being decisive, the current decomposition also suggests other problematic components beyond emigration in the explanation of closure error for specific provinces. For example, the results suggest that estimates of interprovincial migration might be somewhat misstated for British Columbia and Newfoundland (after acknowledging the differences observed on these components and corresponding closure errors). Overall, an acceptance of the RRC on these more difficult to estimate migratory flows – would not only explain the largest part of this difference in growth – but also the largest part of 1996 closure error. With the closure error that remains, it is useful to turn to the observed difference on enumerated. In so doing, the emphasis shifts away from potential problems in the postcensal estimates.

4.3. Comparisons between Estimates of Enumerated

While the difference in enumerated observed nationally is much smaller than the difference documented on growth, for about half the provinces, this difference is of comparable if not larger size. In interpreting this fact, it is recognized that the RRC was never designed to target the "enumerated" population. With the priority of documenting the number "missed" in the census, the sampling design of the RRC over represents "difficult to enumerate groups"

(for example, single young adults), while under representing persons easily "enumerated". Overall, the comparison on enumerated bears well for the accuracy of the RRC – with non-significant differences across all provinces/territories. Nevertheless, the differences observed in a few provinces are reason for concern, being very close to statistical signifi.cance at the 95% level in Quebec (positive difference), and approaching statistical significance in British Columbia, Alberta and Manitoba (negative differences).

In the evaluation of the 1991 coverage study results, two alternative hypotheses have been raised in explanation of differences observed for the enumerated (Royce 1993). At one extreme, it could be argued that all of the difference (for a specific province) be explained in terms of the representativeness of the RRC sample, which implies sampling error or frame deficiencies of one sort or another. At the other extreme, it could be argued that all of the difference be explained due to a failure in documenting the true ratio of enumerated to other classification outcomes, which seems to imply some sort of misclassification error or no trace adjustment bias. A correction for the former of the two hypotheses has a relatively minor impact on the estimate of missed (*i.e.*, all classification outcomes are accordingly inflated or deflated by the proportional difference on enumerated). A correction for the latter could have potentially quite a pronounced impact, as a failure to estimate the true ratio implies that all the difference be assigned to other categories.

If the latter hypothesis applies, a correction potentially reduces the error of closure in nine out of twelve provinces/territories (*i.e.*, in all provinces under which the error of closure is in the same direction as the difference on enumerated). On the other hand, if the differences are due to problems in sample representativeness, a subsequent correction is expected to have negligible impact, if not slightly inflating closure error across most provinces. In addition, the evaluation is complicated by the difficulty in establishing the comparable census figures. Error can be potentially introduced through various sources, including: the census-based estimate of returning emigrants ($^{91FR}RE_{96PP}$), too much or too little correction for frame overlap, sampling and non-sampling error in the estimation of undercoverage in 1991 and 1996, sampling and non-sampling error in the estimation of overcoverage, and potential error in the classification by province of the enumerated. In this context, further research appears justified as to the true character of errors in the RRC estimate of enumerated.

5. CONCLUSION

In this paper, we have shown how there is additional information available through Canada's census coverage measurement program that is of considerable value in population estimation. Beyond the ability to estimate census undercount, it is possible to extend the classification results from these studies in order to obtain an alternative estimate of demographic growth – potentially decomposed by component. Using the most important of the coverage studies (*i.e.*, the 1996 Reverse Record Check), a new method was presented which allows for an independent estimate of demographic growth for the intercensal period. The Reverse Record Check not only provides what are considered highly accurate estimates of census coverage error, avoiding some of the correlation biases that have hindered post-enumeration studies in other countries, but also provides very valuable insight as to the magnitude of selected migratory flows of importance to population estimation.

The key to the Reverse Record Check is that it begins with a representative sample of all persons who could have theoretically been in Canada on census day, with only minor deficiencies due to the high quality of vital statistics and immigration data in Canada. Through exhaustive tracing and interviewing procedures, valuable information is then obtained as to the number and characteristics of persons successfully enumerated, missed, counted more than once, as well as useful information on the numbers leaving the country (whether temporarily or permanently), the numbers dying, living in another province, and so on. With a relatively large sample and considerable expertise and effort directed toward minimizing all forms of error, the resultant classification results can potentially inform the population estimates program. This is particularly true with some of the more difficult to estimate migratory flows.

In planning for the 2001 Census, the goal of minimizing all error in the census coverage measurement program remains a priority. As these studies have been designed with a primary target of estimating the population "missed" rather than other classification outcomes (emigrated, deceased, *etc.*), the new demographic approach presented in the current paper leads to the logical question, as to whether its current design need be reworked somewhat if its current usage is broadened. Of interest in this context is the fact that these coverage studies appear to provide an alternative estimate of growth which rivals that as currently available through the population estimates program, and is likely superior with respect to selected components. Further research about how we might more fully exploit this fact appears justified, in improving the quality of the population estimates program.

ACKNOWLEDGEMENTS

We would like to thank R.G. Carter and P. Dick (both of Statistics Canada) and G. Robinson (U.S. Bureau of the Census) for their comments on an earlier draft of this paper. We also acknowledge the helpful comments and suggestions from the Associate Editor and two referees.

REFERENCES

BRACKSTONE, G.J., and GOSSELIN, J.F. (1973). *Census Evaluation Program, 1971 RRC: Methodology Report*. Statistics Canada. Ottawa, Ontario.

BURGESS, R.D. (1988). Evaluation of Reverse Record Check estimates of undercoverage in the Canadian census of population. *Survey Methodology*, 14, 137-156.

CARTER, R.G. (1990). The Measurement of net coverage error in Canadian censuses. *Proceedings: Symposium 90, Measurement and Improvement of Data Quality*, Statistics Canada.

FELLEGI, I.P. (1969). A theory of record linkage. *Journal of the American Statistical Association*, 64(328), 1183-1210.

GOSSELIN, J.-F. (1976). The methodology of the 1971 Reverse Record Check. *Survey Methodology*, 2, 180-193.

ROMANIUC, A. (1988). A demographic approach to the evaluation of undercoverage in the Canadian Census of Population. *Survey Methodology*, 14,157-172.

ROYCE, D. (1993). Evaluation of the May 1993 Revised Results of the 1991 Census Coverage Studies. Social Survey Methods Division, Working Paper, Statistics Canada, Ottawa, Ontario.

ROYCE, D., GERMAIN, M.-F., JULIEN, C., DICK, P., SWITZER, K., and ALLARD, B. (1994). *Coverage: 1991 Census Technical Report*. Catalogue no. 92-341E. Ottawa: Statistics Canada.

STATISTICS CANADA (1999). *Coverage: 1996 Census Technical Reports*. Catalogue no. 92-370-XPB.

STATISTICS CANADA (2000). *Annual Demographic Statistics*. Catalogue no. 21-213-XPB.

TOURIGNY, J., CLARK C., and PROVOST, M. (1998). Evaluation of the March 1998 Preliminary Results of the 1996 Census Coverage Studies. Social Survey Methods Division, Working Paper, Statistics Canada, Ottawa, Ontario.

TOURIGNY, J., BUREAU, M., and CLARK, C. (1998). Revised Direct Estimates of 1991 Census Coverage Studies. Sept 24th release. Social Survey Methods Division, Working Paper, Statistics Canada, Ottawa, Ontario.

Survey Methodology, June 2000
Vol. 26, No. 1, pp. 53-65
Statistics Canada

Multilevel Modelling of Complex Survey Longitudinal Data With Time Varying Random Effects

MOSHE FEDER, GAD NATHAN and DANNY PFEFFERMANN[1]

ABSTRACT

Longitudinal observations consist of repeated measurements on the same units over a number of occasions, with fixed or varying time spells between the occasions. Each vector observation can be viewed therefore as a time series, usually of short length. Analyzing the measurements for all the units permits the fitting of low-order time series models, despite the short lengths of the individual series. We illustrate this paradigm using simulated data that follow the rotation scheme of the Israel Labor Force Survey (LFS). This survey employs a rotating panel sampling scheme of two quarters in the sample, two quarters out of the sample and then two quarters in again. The model consists of two-level linear models for single time points that are connected by allowing the second level effects (corresponding to households) and the first level residuals (corresponding to individuals) to evolve stochastically over time. The likelihood of the model is easily constructed by employing the time series properties of the combined model. However, in view of the large number of unknown parameters, direct maximization of the likelihood could yield unstable estimators. Therefore, a two-stage procedure is adopted. At the first stage, a separate two-level model is fitted for each time point, thus yielding estimators for the fixed effects and the variances. At the second stage, the time series likelihood is maximized only with respect to the time series model parameters. This two-stage procedure has the further advantage of permitting appropriate first and second level weighting to account for possible informative sampling effects. Empirical results when fitting the model to data collected by the Israel LFS are also presented

KEY WORDS: Informative sampling; Probability weighted IGLS; Rotating panel schemes; State-space models.

1. INTRODUCTION

1.1 Background and Objectives

In recent years there has been a growing interest in fitting models to data collected from longitudinal surveys that use complex sampling designs. This interest reflects expansion in requirements by policy makers and social scientists for in-depth studies of social processes over time, rather than of one-time "snap-shots" provided by cross-sectional analyses. A familiar example is the estimation of gross flows between social and demographic states such as employment states or health and education levels. For discussions of these issues and the problems they raise with respect to the design and analysis of longitudinal surveys, see Duncan and Kalton (1987) and Binder (1998).

Examples of surveys we wish to consider in this paper are of three types:

1. Rotating panel surveys such as labor force surveys carried out in many countries. These surveys were often designed originally for cross-sectional analysis of household and individual data, so as to study labor force and other socio-economic characteristics on a current basis. Complex rotating sampling schemes have later been introduced in order to improve comparisons over time. For example, the quarterly Israel Labor Force Survey (LFS) employs a rotating panel sampling scheme whereby each unit in the sample is interviewed for two consecutive quarters; it is left out of the sample for the next two quarters and then is interviewed again for two more consecutive quarters. In The U.S.A. and Brazil, a more complicated sampling scheme of 4 months in the sample, 8 months out of the sample and then 4 months in again is used. Australia, Canada and the U.K. employ sampling schemes by which sampled units are interviewed over a succession of months or quarters before being dropped from the sample. These kinds of surveys are increasingly used for short-term longitudinal analysis, such as the estimation of gross flows between labor force states or studies of social mobility. This has not always proved simple due to the complexity of the survey designs, difficulties in matching and response errors.
2. Medium term panel surveys, such as the U.S. Survey of Income and Programme Participation (SIPP, Herriot and Kasprzyk 1984), the U.S. Panel Study of Income Dynamics (PSID, Survey Research Center, 1984) and the Canadian Survey of Labor and Income Dynamics (SLID, Webber 1994). These surveys differ from labor force surveys in being specially designed for longitudinal analysis of economic and social characteristics of households and individuals. For example, SIPP includes an intensive investigation in the form of a full retrospective interview every 4 months. It provides a complete work history for the

[1] Moshe Feder, Department of Social Statistics, University of Southampton, Southampton, S017 1BJ, U.K.; Gad Nathan and Danny Pfeffermann, Department of Statistics, Hebrew University, Jerusalem, 91905, Israel.

survey period (30-48 months) by combining the continuous retrospective four-month recall data with a reconciliation of data provided for longer periods.

3. Longitudinal cohort studies characterized by the follow-up of a cohort sample over a long time period. For example, in the British Household Panel Survey, starting from a sample of addresses selected in 1991, data have been collected on the same households in subsequent annual waves for over seven years. A wide range of data is collected on labor force characteristics, economic resources and health and education, with emphasis on longitudinal aspects. In this survey all members of the originally selected households were followed and the sample was supplemented by the addition of entrants to the sample households, including children born to sample household members. Other longitudinal cohort studies such as the British National Child Development Study and the British Cohort Study have surveyed a cohort of births over periods of up to 40 years. See Nathan (1999) for description and discussion of the latter three studies.

Most of the studies associated with these surveys require longitudinal analysis for populations that have a complex hierarchical structure, based on data collected from complex sampling designs. Standard analysis of longitudinal survey data often fails to account for the complex nature of the sampling design such as the use of unequal selection probabilities, clustering, post-stratification and other kinds of weighting used for the treatment of non-response. The effect of sampling on the analysis is due to the fact that the models in use typically do not incorporate all the design variables determining the sample selection, either because there may be too many of them or because they are not of substantive interest. However, if the design is "informative" in the sense that the outcome variable is correlated with the design variables not included in the model, even after conditioning on the model covariates, standard estimates of the model parameters can be severely biased, leading possibly to false inference. Pfeffermann (1993, 1996) reviews many examples reported in the literature that illustrate the effects of ignoring the sampling process when fitting models to survey data and discusses methods that have been proposed to deal with this problem. See also the book edited by Skinner, Holt, and Smith (1989) and the more recent paper by Pfeffermann, Skinner, Goldstein, Holmes, and Rasbash (1998) to which we refer in more detail below. It should be emphasized that standard inference may be biased even when the original sample design is simple random within design strata, due to non-response, attrition, and imperfect frames that result in de facto a posteriori differential inclusion probabilities. Special features of longitudinal studies, such as late additions of individuals who join panel households, can also lead to de facto unequal inclusion probabilities.

In this paper we propose to deal with the problems arising from the hierarchical nature of the target population, the longitudinal aspect of the analysis and the effects of complex sampling designs by combining three separate statistical methodologies. These are multilevel modelling (MLM), time series modelling and methods of analysis under complex informative sampling. Multilevel models are used to deal with the hierarchical structure of many human populations like persons within households, pupils within classes, classes within schools and so forth. The models, extensively employed by social scientists especially in the field of education, account for the effects of observed covariates at the lower and higher levels of the structure, with fixed or random coefficients. Common unobservable random effects within the higher levels capture further unexplained variations. The method of Iterative Generalized Least Squares (IGLS) is commonly used for estimating the model parameters, Goldstein (1986, 1995).

Simple state-space time series models are used to combine the multilevel models operating at different time points via a set of linear transition equations that account for the time series relationships of the random covariate coefficients and the higher level random effects. The Kalman filter is used for estimating the model parameters and predict the random effects for current and future time points. Smoothing algorithms can be used for updating past predictions, Harvey (1989). Methods of model fitting under informative sampling are employed to control the bias resulting from the sample selection process. Such methods have been investigated in recent years in the context of analytic inference from complex sample surveys, mostly for cross-sectional analysis of single-level models, *cf.* Skinner *et al.* (1989). In the present paper we utilize the methodology of sample weighting for multilevel modelling as developed by Pfeffermann *et al.* (1998).

The aims of the present study are then to develop models and methods of estimation for longitudinal analysis of hierarchically structured data, taking unequal sample selection probabilities into account. The main feature of our approach is that the model is fitted at the individual level but it contains common higher level random effects that change stochastically over time. The model enables to predict the higher and lower level random effects (like household and individual person effects in the present application), using the data for all the time points with observations. This should enhance model-based inference from complex survey data since it permits a better understanding of the structure and correlation pattern of the longitudinal measurements. In particular, it is bound to improve the prediction of individual measurements compared to the use of aggregate time series models, which by their nature fail to separate the individual (person) effects from the common higher level (household) effects. These advantages are partly illustrated in the example of section 6 and more so in a related paper by Pfeffermann and Nathan (forthcoming) which focuses on the imputation of missing data. It is

important to emphasize in this regard that although the length of each individual longitudinal record is often very short (4 measurements for each individual in our application), the number of records is usually sufficiently large to warrant the application of classical time series estimation and model diagnostic procedures. In this article we only consider parameter estimation under a given model but the use of test statistics and diagnostic procedures that employ the empirical innovations for model identification follows through with minor modifications by virtue of the use of maximum likelihood estimation methods and the consistency of the parameter estimators.

In section 2 we overview the main features of the aforementioned statistical methodologies that are employed in subsequent sections. In section 3 we propose a model that addresses the longitudinal aspects discussed above. Estimation procedures are discussed in section 4. Section 5 contains the results of a simulation study carried out for assessing the performance of the various estimators under different sampling scenarios. Results obtained when fitting the model to real data collected by the Israel LFS are presented in section 6, followed by a brief summary in section 7 of possible model extensions and applications.

1.2 Literature Review

Previous work in this area deals mostly with longitudinal data in a non-survey context and does not consider hierarchically structured populations. In particular, none of the studies that we have come across permits the second level effects (common household effects in our application) to evolve over time. For example, Goldstein, Healy and Rasbash (1994) consider the analysis of repeated measurements using a two-level model with individuals as second levels and the repeated measurements as the first levels. The model extends the standard two-level model by permitting the first level measurements to be correlated over time. The authors consider several possibilities of modelling the autocorrelation structure, which include autoregressive models when the measurements are taken at equally spaced time points and autocorrelation functions when the observations are taken at unequal time intervals. In the latter case the autocorrelation function is linearized for estimation purposes.

Several authors study the application of time series models for the analysis of longitudinal data. In a series of papers by Jones and his co-authors (Jones and Ackerson 1990, Jones and Boadi-Boating 1991, Jones and Vecchia 1993) and the book by Jones (1993), the authors consider observations taken at unequally spaced time gaps. The observations referring to the same subject are allowed to be serially correlated by postulating continuous autoregressive moving average models. These models contain fixed and random effects, but do not have a hierarchical population structure. Weighted least squares and state space modelling combined with the Kalman filter are used for calculating the likelihood function.

Continuous time autoregressive models for irregularly spaced longitudinal data are considered also by Belcher, Hampton and Tunnicliffe (1984), using linear stochastic differential equations for describing the process generating the data. An Empirical Bayes approach is proposed by Bryant and Day (1991) for the simultaneous analysis of a system of mixed linear models, having linked and serially correlated random effects. Chi and Reinsel (1989) consider a score test for autocorrelation between individual errors under a "conditional independence" random effects model. The authors derive a maximum likelihood estimation procedure and use the estimators for predicting the random effects by application of Empirical Bayes.

Diggle, Liang and Zeger (1994) propose the use of generalized linear models for the analysis of longitudinal data. They consider a transition (Markov) model by considering past values as additional predictor variables. Transitional extensions of the GLM are used for maximum likelihood estimation under linear link functions, whereas for non-linear link functions the estimation is based on conditional score functions. Lawless (1999) uses an event history approach for the analysis of longitudinal data. By this approach, the dependent variable is the number of occurrences of a particular event up to a given time point t, with the limiting transitional probabilities being modelled as functions of the previous history and covariates. Zimmerman and Nunez-Anton (1997) propose a structured antedependence model for longitudinal data, primarily in the context of growth analysis. Neither of the above studies considers a hierarchical structure or a complex sampling design.

Finally, Skinner and Holmes (1999) consider a model for longitudinal observations that consists of a "permanent" random effect at the individual level and autocorrelated transitory random effects corresponding to different waves of investigation. The authors study two approaches for the estimation of the unknown model parameters with both approaches accounting for sampling effects and "non informative" attritions. The first approach treats the repeated observations as correlated multivariate outcomes and derives probability-weighted estimators that account for the correlation structure. The second approach considers the model as a two-level model with "individuals" as the second level units and the repeated measurements as first level units. Estimation of the unknown parameters under this approach is carried out by a modification of the PWIGLS method of Pfeffermann *et al.* (1998, see section 2.2).

2. STATISTICAL METHODOLOGIES UNDERLYING THE PROPOSED APPROACH

2.1 Multilevel Models

In what follows we consider a two-level model for the response variable y in a population consisting of

$i = 1, ..., M$ second level units (household, schools, ...) and $j = 1, ..., N_i$ individuals within second level unit i. The model is,

$$y_{ij} = x'_{ij}\beta + z'_{ij}u_i + z_{0ij}e_{ij},\ i = 1, ..., M;\ j = 1, ..., N_i, \quad (2.1)$$

where x_{ij}, z_{ij}, and z_{0ij} are known covariate values of dimensions p, q and 1 respectively, β is a fixed parameter vector of dimension p and $\mu_i \sim N(0, \Omega)$ and $e_{ij} \sim N(0, \sigma^2)$ are independent random second level effects and first level residuals of orders p and 1 respectively.

The inclusion of the multipliers z_{0ij} allows for first level heteroscedasticity whereas the common second level effects u_i explain the (interclass) correlations between individual measurements corresponding to the same second level unit. In the simple case of the "random intercept model", $y_{ij} = x'_{ij}\beta + u_i + e_{ij}$, these correlations take the familiar form, $\mathrm{Corr}(y_{ij}, y_{ik}) = \sigma_u^2 / (\sigma_u^2 + \sigma^2)$. The random intercept model is often applied for small area estimation (see below).

As stated in the introduction, models like (2.1) are widely used by social scientists for studying the effects of the covariate variables and the interrelationships between observations corresponding to the same higher level unit. In such cases, primary interest is in the estimation of the vector coefficient β and the vector θ of the distinct elements of Ω and σ^2. Another, well-known application of the two-level model is for "small area estimation", in which case the second levels are geographical areas or other domains of study. In small area estimation, the target of the analysis is the prediction of the second level (area) means $\bar{X}'_i\beta + \bar{Z}'_i u_i$, where $\bar{X}_i$ and $\bar{Z}_i$ are the true area covariate means, and the estimation of the model parameters is only an intermediate step. See Rao (1999) for a recent review.

Estimation of the unknown model parameters is carried out most conveniently by use of the Iterative Generalized Least Squares (IGLS) algorithm (Goldstein 1986, 1995). For a random sample of m second level units and n_i first level units within second level unit i, the model holding for the sample data is first written in matrix form as

$$y_i = X_i\beta + d_i,\ i = 1 \ldots m \quad (2.2)$$

where $y_i = [y_{i1}, ..., y_{in_i}]'$, $X_i = [x_{i1}, ..., x_{in_i}]'$ and $d_i = [d_{i1}, ..., d_{in_i}]'$ with $d_{ij} = (z'_{ij}u_i + z_{0ij}e_{ij})$. Then, $d_i \sim N(0, V_i)$, where $V_i = Z_i\,\Omega\,Z'_i + \sigma^2 Z_{0i}^2 = V_i(\theta)$; $Z_i = [z_{i1} \ldots z_{in_i}]'$ and $Z_{0i} = \mathrm{diag}[z_{0i1} \ldots z_{0in_i}]$. The IGLS algorithm iterates between the estimation of β, with θ considered known, and the estimation of θ, with β considered known. At each iteration, the estimate obtained for the other vector parameter on the previous iteration is used as the "known" parameter. This process is a special case of the EM algorithm and it converges to the corresponding maximum likelihood estimators (MLE) under the stated normality assumptions. It is known to provide consistent estimators under more general conditions.

2.2 MLM Estimation Under Informative Sampling

The IGLS algorithm described in section 2.1 assumes that the model defined by (2.2) holds for the sample data. This would be the case if selection of the first and second level units is carried out by simple random sampling. However, as discussed in the introduction, the selection of the sample could be informative so that the model holding for the sample units differs from the model holding in the population. For example, in an educational survey, schools in poor areas could be sampled with higher probabilities. In a household survey, higher selection probabilities could be assigned to households in areas characterized by high proportions of minorities or to persons that are unemployed. As illustrated by Pfeffermann *et al.* (1998) and also in section 5 of the present paper, the use of the IGLS algorithm in such cases could yield severely biased estimators for all the parameters. The authors propose therefore a probability weighted IGLS (PWIGLS) algorithm that protects against informative sampling.

The algorithm is an adaptation of the pseudo-MLE method (Binder 1983, Skinner *et al.* 1989, Pfeffermann 1993). Suppose that the two-level model defined by (2.1) holds for the target population. Had all the population values been observed, the IGLS would converge at the end of the iterative process to the census estimators, $(\hat{\beta}_c, \hat{\theta}_c)$. At each iteration, the intermediate estimators $(\hat{\beta}_{(i)}, \hat{\theta}_{(i)})$ are products of matrices with elements that are functions of sums of the population values. When the IGLS is applied to sample data, the population sums are substituted by the corresponding sample sums. The PWIGLS consists of further replacing the unweighted sample sums by weighted sums. Denote by $\pi_i = \Pr(i \in s)$ the second level sample inclusion probabilities and by $\pi_{j|i} = \Pr(j \in s \mid i \in s)$ the conditional first level inclusion probabilities. The PWIGLS estimators are obtained by, 1- replacing each second level sample sum of the general form $\sum_{i=1}^{n} g_i$ by the weighted sum $\sum_{i=1}^{n} w_i g_i$, where $w_i = \pi_i^{-1}$ and 2- replacing each first level sample sum $\sum_{j=1}^{n_i} g_{ij}$ by the weighted sum $\sum_{j=1}^{n_i} w_{j|i} g_{ij}$ with $w_{j|i} = \pi_{j|i}^{-1}$. Note that the weighting process requires the knowledge of the inclusion probabilities at both stages of the selection process and not just the final overall inclusion probabilities $\pi_{ij} = \pi_{j|i} \times \pi_i$.

As established by Pfeffermann *et al.* (1998), the PWIGLS estimators are consistent for the model parameters when both the first and second level sample sizes increase, but the estimators of the variances are not consistent if the first level sample sizes are bounded. For this case, the authors propose appropriate scaling of the weights $w_{j|i}$ that eliminates the bias, provided that the sample selection within the second level units is noninformative. It is important to emphasize that standard weighting of the sample measurements by the weights $w_{ij} = \pi_{ij}^{-1}$, which is routinely applied for single level models yields consistent estimators only for β.

2.3 State-space Models

State-space models as considered here consist of two sets of equations:

1. The measurement (observations) equation:

$$y_t = X_t\beta_t + L_t\alpha_t + \varepsilon_t;\ E(\varepsilon_t) = 0,$$

$$E(\varepsilon_t\varepsilon_{t-k}') = \delta_k H_t,\ t = 1, ..., T \qquad (2.3)$$

2. The transition (system) equation:

$$\alpha_t = G_t\alpha_{t-1} + \eta_t;\ E(\eta_t) = 0,$$

$$E(\eta_t, \eta_{t-k}') = \delta_k Q_t,\ t = 1, ..., T \qquad (2.4)$$

where $\delta_k = 1$ for $k = 0$ and $\delta_k = 0$ otherwise. We also assume $E(\varepsilon_t\eta_s') = 0$ for all t and s. Note that both y_t and α_t can be multivariate. The measurement equations relate the observations y_t at any given time point to covariate values X_t with fixed (nonstochastic) vector coefficients β_t, and linear functions L_t of an unobservable state vector α_t. The transition equations describe the time series relationships between the components of the state vector. The matrices X_t, L_t and G_t are assumed to be nonstochastic although they may change over time, as is the case with the vector coefficients β_t. Notice that the latter vectors can be included as part of the state vectors by taking their transition matrix to be the zero matrix of corresponding order and defining the corresponding residual variances in Q_t to be very large. See Sallas and Harville (1981) for details.

Although not written here in its most general form, the state-space model defined by (2.3) and (2.4) is known to include as special cases many of the time series and mixed linear models in common use. As important examples we mention the family of ARIMA models and models with random regression coefficients. The MLM defined by (2.1) can also be easily structured in a state-space form. To see this, replace the index i by t and define $L_t = [X_t, Z_t]$, $\alpha_t = [\beta_t', u_t']'$, $H_t = \sigma^2 Z_{0t}$ and $G_t = [I_p, 0_q]$ where I_p and 0_q define the identity matrix and the zero matrix of the appropriate orders. (The matrices Z_t and X_t are defined below (2.2).) The vector coefficient β_t is added for convenience to the state vector. The covariance matrix Q_t is block diagonal with 0_p and $Z_t\Omega Z_t'$ as the two blocks. The use of the zeroes matrix 0_p for the covariance of $(\beta_t - \beta_{t-1})$ guarantees that the β-coefficients are fixed over time, in accordance with (2.1). (The representation of the MLM in a state-space form is not unique.)

For given covariance matrices $\{H_t, Q_t\}$ and assuming that β_t, L_t and G_t are known for all t, the best linear unbiased predictor (BLUP) of the state vector at any given time t, based on all the data accumulated until that time, is conveniently obtained by means of the Kalman Filter. Let $\hat{\alpha}_{t-1}$ define the BLUP of α_{t-1} based on the observations until time $(t\text{-}1)$, with covariance matrix $P_{t-1} = \mathrm{Cov}(\hat{\alpha}_{t-1} - \alpha_{t-1})$. The BLUP of α_t at time $(t\text{-}1)$ is then, $\hat{\alpha}_{t|t-1} = G_t\hat{\alpha}_{t-1}$ with covariance matrix $P_{t|t-1} = \mathrm{Cov}(\hat{\alpha}_{t|t-1} - \alpha_t) = G_t P_{t-1} G_t' + Q_t$. When new observations y_t become available, the predictor $\hat{\alpha}_{t|t-1}$ and the corresponding covariance matrix are updated as

$$\hat{\alpha}_t = \hat{\alpha}_{t|t-1} + P_{t|t-1}L_t'F_t^{-1}(y_t - X_t\beta_t - L_t\hat{\alpha}_{t|t-1})$$

$$P_t = P_{t|t-1} - P_{t|t-1}L_t'F_t^{-1}L_t P_{t|t-1} \qquad (2.5)$$

where $F_t = L_t P_{t|t-1} L_t' + H_t = \mathrm{Var}(y_t - \hat{y}_{t|t-1})$ with $\hat{y}_{t|t-1} = X_t\beta_t + L_t\hat{\alpha}_{t|t-1}$ defining the BLUP of y_t at time $(t\text{-}1)$. The actual application of the Kalman filter requires a proper initialization for $\hat{\alpha}_{1|0}$ and $P_{1|0}$ which depends on the model under study. See section 4 for the initialization under the model proposed in this paper.

The unknown model parameters (β_t, elements of H_t, Q_t and possibly L_t and G_t) are ordinarily estimated by MLE with the likelihood conveniently constructed by use of the "prediction error decomposition". Assuming that $\dim(y_t) = n$, the log-likelihood takes the general form,

$$\log(L) = -\{T\frac{n}{2}\log(2\pi) + \frac{1}{2}\sum_{t=1}^{T}\log|F_t| + \frac{1}{2}(Y_t - \hat{Y}_{t|t-1})'F_t^{-1}(Y_t - \hat{Y}_{t|t-1})\}. \qquad (2.6)$$

For a thorough discussion of state-space models and their applications, see Harvey (1989).

3. A MODEL FOR HIERARCHICAL LONGITUDINAL DATA

In this section we propose a time series multilevel model which combines separate cross-sectional two-level models by modelling the evolution of the first and second level random effects over time. Let S_t define the sample available at time t, composed of m_t level 2 units with n_h level 1 units in level 2 unit h. The formulation of the overall sample in terms of the subsets S_t covers situations where the longitudinal observations are collected at different time periods. The proposed model allows also for the rotation patterns mentioned previously and for wave non-response. Note that the samples observed at different time points are generally not disjoint and that the assumption that n_h is fixed over time is not restrictive. Pfeffermann and Nathan (forthcoming) consider the case of temporal missing data for which this supposition does not hold. As long as the missing data are missing completely at random, generalization of the present methodology to this case is straightforward. We assume the following two-level model to hold for the sample S_t:

$$y_{hjt} = x_{hjt}'\gamma_t + z_{ht}'v_t + z_{ht}'u_{ht} + e_{hjt},$$

$$h = 1, ..., m_t, j = 1, ..., n_h, \qquad (3.1)$$

where y_{hjt} is the outcome for first level unit j in second level unit h, x_{hjt} and z_{ht} are fixed known covariate vectors of dimensions p and q respectively, γ_t and v_t are fixed (unknown) vector coefficients and u_{ht} and e_{hjt} are independent second level and first level random effects. For given time t, The model defined by (3.1) is basically the same as the MLM model defined by (2.1), except that we assume $z_{hjt} = z_{ht}$ for all j and t, thus distinguishing between first level covariates and second level covariates. We assume also for convenience $z_{0hjt} \equiv 1$. The model is quite general in that all the covariate variables, the fixed vector coefficients and the random effects are allowed to vary over time in ways defined below. Notice that by assuming that (3.1) holds for the sample data, it is implicitly assumed that the sampling design is noninformative. See the discussion in section 2.2 and also section 4 below.

As in (2.2), the model defined by (3.1) can be formulated in matrix form as,

$$Y_{ht} = X_{ht}\gamma_t + Z_{ht}v_t + Z_{ht}\mu_{ht} + I_{n_h}e_{ht}, \qquad (3.2)$$

where $Y_{ht} = [y_{h1t},...,y_{hn_ht}]'$, $X_{ht} = [x_{h1t},...,x_{hn_ht}]'$, $Z_{ht} = 1 \otimes z_{ht}$ and $e_{ht} = [e_{h1t}, ..., e_{hn_ht}]'$ with $\otimes$ defining the Kronecker product. The matrix representation (3.2) can be written concisely as,

$$Y_{ht} = \tilde{X}_{ht}\beta_t + \tilde{Z}_{ht}\alpha_{ht}, \qquad (3.3)$$

where $\tilde{X}_{ht} = [X_{ht}, Z_{ht}]$; $\tilde{Z}_{ht} = [Z_{ht}, I_{n_h}]$; $\beta_t = [\gamma_t', v_t']'$; $\alpha_{ht} = [u_{ht}', e_{ht}']'$.

Next we model the time series relationships of the vector coefficients and the random effects. We assume that the vectors β_t, $t = 1, 2...$ are fixed without specifying the way they evolve over time. This assumption is generally not restrictive because in practical applications the overall sample size in any given time point is usually sufficiently large to allow accurate estimation of the vector coefficients without having to borrow information across time. For the random second and first level effects we postulate first order autoregressive [AR(1)] relationships of the form,

$$u_{ht} = Au_{h,t-1} + \delta_{ht}; \quad e_{ht} = \rho e_{h,t-1} + \varepsilon_{ht} \qquad (3.4)$$

where A is a $(q \times q)$ matrix of fixed coefficients, ρ is a fixed scalar and $\delta_{ht} \sim N(0_q, \Delta)$; $\varepsilon_{ht} \sim N(0_{n_h}, \sigma_\varepsilon^2 I_{n_h})$ are independent white noise series. The model defined by (3.4) is rather simple and as a further simplification we assume that A and Δ are diagonal, implying that the second level random effects are independent. It is assumed also that $|\rho| < 1$ and $|A_{kk}| < 1$ for all k to guarantee stationarity. More complex models can be considered in principle but it should be emphasized that unlike in classical (aggregate) time series analysis, longitudinal observations may only be taken over a very short time period in which case the use of models that incorporate lagged values of high order may no longer be operational. For example, in the quarterly Israel LFS described in the introduction, individuals are in the sample for a total of 4 quarters over a time period of 6 quarters which clearly limits the class of time series models that can be postulated for the random effects.

The AR(1) models defined by (3.4) can be written concisely as

$$\alpha_{ht} = G_h\alpha_{h,t-1} + \eta_{ht}, \; h = 1, ..., m_t \qquad (3.5)$$

where,

$$G_h = \begin{bmatrix} A & 0 \\ 0 & \rho I_{n_h} \end{bmatrix}, \; \eta_{ht} = \begin{bmatrix} \delta_{ht} \\ \varepsilon_{ht} \end{bmatrix},$$

$$\eta_{ht} \sim N(0, Q_h), Q_h = \begin{bmatrix} \Delta & 0 \\ 0 & \sigma_\varepsilon^2 I_{n_h} \end{bmatrix}. \qquad (3.6)$$

By writing the proposed model using the equations (3.3), (3.5) and (3.6) and setting $\tilde{Z}_{ht} = L_{ht}$, $H_{ht} = 0$, it is easily seen to belong to the class of state-space models presented in section 2.3, with no residual errors in the measurement equation. The model is defined for distinct second level units h but unlike in classical time series analysis where the data consist of a single long series, the data in our case consist of many independent short (longitudinal) series that could be observed over different time periods. Note that the transition matrix, G_h and the covariance matrix, Q_h depend on h through the second level size n_h but they are time invariant. In situations where the second level sizes are not fixed over time (for example, because of missing data), these matrices also change accordingly.

4. ESTIMATION OF THE MODEL PARAMETERS

In principle, the likelihood function holding for the model defined by (3.3), (3.5) and (3.6) can be maximized to obtain the maximum likelihood estimators (MLE) of all the unknown model parameters. However, the number of estimated parameters would usually be very large, which can intensify the computations and result in statistically unstable estimators. For instance, even for $p = q = 2$ and $T = 10$ there are already 46 unknown parameters. We propose therefore a two-stage estimation procedure that employs MLM estimation for the "cross-sectional parameters" and state-space model estimation for the "time series parameters". The use of this procedure has the further advantage of accommodating appropriate weighting to protect against informative sampling.

The procedure starts off by fitting the MLM defined by (3.1) to each sample S_t separately, to obtain IGLS estimates of the time-dependent fixed effects $\beta_t = [\gamma_t', v_t']'$ and the variances of the random effects u_{ht} and e_{hjt}. Notice that by (3.4),

$$\mathrm{Var}(u_{ht}) = \Delta^* = (I - A^2)^{-1}\Delta;$$

$$\mathrm{Var}(e_{hjt}) = \sigma_e^2 = (1 - \rho^2)\sigma_\varepsilon^2 \qquad (4.1)$$

using familiar relationships holding for AR(1) models. The use of this step yields estimates $\{\hat{\beta}_t, \hat{\Delta}_t^*, \hat{\sigma}_{et}^2\}$ for $\{\beta_t, \Delta^*, \sigma_e^2\}$ respectively. Under the model, the true variances (Δ^*, σ_e^2) are fixed over time and assuming that the sample sizes at the various time points are fairly constant, the estimates $\hat{\Delta}_t^*$ and $\hat{\sigma}_{et}^2$ can be averaged to yield single estimates

$$\bar{\Delta}^* = \sum_{t=1}^{T} \hat{\Delta}_t^* / T; \quad \bar{\sigma}_e^2 = \sum_{t=1}^{T} \hat{\sigma}_{et}^2 / T. \qquad (4.2)$$

In the second stage the remaining parameters are estimated by maximizing the likelihood of the combined model defined by (3.3) (3.5) and (3.6), with the parameters estimated in the first stage held fixed at their estimated values. Since observations on different second level units are independent, the log-likelihood has the form $\log(L) = \sum_h \log(L_h)$ where L_h, the contribution to the likelihood from second level unit h, is defined by (2.6) with the index h added to all the components thus distinguishing between different second level units. As pointed out before, the number of time points for which the second level units are observed and the time periods over which the observations are taken may differ between units so that the notation T in (2.6) for the number of time points needs also to be changed to T_h.

When fitting the model to data obtained from rotating panel sampling designs as in the empirical study of the present paper, a further modification is required to account for the intermediate periods without observations. For example, for the Israel LFS described in the introduction, with rotation pattern of two quarters in the sample, two quarters out of the sample and two quarters in again, $T_h = 4$ but the transition equations from $t = 2$ to $t = 3$ (the next quarter with observations) have to be changed to account for the two quarters with missing observations. Repeated substitutions in (3.5) yield the following relationships:

$$\alpha_{h3} = G_h^3 \alpha_{h2} + \eta_{h3}^*; \quad \eta_{h3}^* \sim N(0, Q_{h3}^*),$$

$$Q_{h3}^* = \begin{bmatrix} (A^4 + A^2 + I)\Delta & 0 \\ 0 & (\rho^4 + \rho^2 + 1)\sigma_\varepsilon^2 I_{n_h} \end{bmatrix}. \qquad (4.3)$$

In order to apply the Kalman filter and compute the likelihood, it is needed to set initial values for α_{110} and P_{110}. This is simple under the present model as $\alpha_{ht} = [u_{ht}', e_{ht}']'$ is stationary with zero mean and covariance matrix defined by (4.1). Thus, the filter is started by setting,

$$\alpha_{h110} = E(u_{h1}', e_{h1}') = 0;$$

$$P_{h110} = \mathrm{Var}[u_{h1}', e_{h1}']$$

$$= \mathrm{diag}\{(I - A^2)^{-1}\Delta, \sigma_\varepsilon^2(1 - \rho^2)^{-1} I_{n_h}\}. \qquad (4.4)$$

In the empirical study described in the next two sections we compare two methods regarding the set of parameters estimated in the second stage.

Method 1: The parameters estimated in Stage 2 are the three AR coefficients ρ, A_{11}, A_{22} and the corresponding residual variances $\sigma_\varepsilon^2 = \mathrm{Var}(\varepsilon_{hjt})$ and $\Delta = \mathrm{Var}(\delta_{ht})$, (equation 3.6, three variances in total). Note that under this method the only estimates utilized from Stage 1 are the fixed parameter estimates $\{\hat{\beta}_t = [\hat{\gamma}_t', \hat{v}_t']'\}$. By (4.1), the variances $\Delta^* = \mathrm{Var}(u_{ht})$ and $\sigma_e^2 = \mathrm{Var}(e_{hjt})$ are estimated as

$$\hat{\Delta}^* = (1 - \hat{A}^2)^{-1}\hat{\Delta}; \quad \hat{\sigma}_e^2 = (1 - \hat{\rho}^2)^{-1}\hat{\sigma}_\varepsilon^2. \qquad (4.5)$$

Method 2: The only parameters estimated in Stage 2 are the AR coefficients ρ, A_{11}, A_{22} (Equation 3.4). Note that with this method the variances Δ and σ_ε^2 are set in the likelihood as, $\Delta = (I - A^2)\bar{\Delta}^*$ and $\sigma_e^2 = (1 - \rho^2)\bar{\sigma}_e^2$ utilizing (4.1), where $\bar{\sigma}_e^2$ and $\bar{\Delta}^*$ are defined by (4.2).

The estimation procedures described so far assume implicitly noninformative sampling. As discussed in the introduction, complex sample surveys often involve selection with unequal probabilities that could be correlated with the values of the response variable. When this is the case, the model holding for the sample data may differ from the model holding in the population. A further advantage of the proposed two-stage estimation method is that it can be adapted to protect against informative sampling. This is done by applying the weighting procedure described in section 2.2 in the first stage, replacing the iterative IGLS algorithm by the PWIGLS procedure. Thus, for each sample S_t, PWIGLS is used for estimating the MLM model parameters instead of using the IGLS.

Comment 1: Informative selection of the first and second level units does not affect the conditional distributions of the random effects as defined by (3.4). Thus, although the distribution of u_{h1} and e_{h1} could be largely distorted because of the sample selection at time $t = 1$, this has no effect on the distributions of $u_{h2} \mid u_{h1}$, or $e_{h2} \mid e_{h1}$. The implication of this property is that the computation of the likelihood in the second stage remains the same, but care should be taken of a proper initialization of the Kalman filter. As defined by (4.4), the filter is initialized by the unconditional means and variances of the random effects under the model, but at time $t = 1$ the moments holding for units in the sample can be different because of the sampling effects. As is well known, for long enough series and under

some regularity conditions, the estimates derived from maximization of the likelihood are not sensitive to the initialization procedure but with short series, improper initialization under informative sampling could distort the estimation process. Nonetheless, as illustrated in section 5, having a moderate number of longitudinal observations even of very short length (at most 4 observations in our application) and weighting the likelihood contributions by the inverse of the sample inclusion probabilities (application of the pseudo likelihood approach) yields approximately unbiased estimators for all the time series model parameters.

5. SIMULATION RESULTS

In this section we report the results of a Monte Carlo study carried out for assessing the performance of the various estimation procedures described in section 4 under noninformative and informative rotating sampling schemes.

5.1 Description of Simulation Study

A) Generation of population data and sample rotation scheme

Population values have been generated for individuals (first level units) within households (second level units), using the model defined by (3.1) and (3.4) (see below). The number of persons n_h observed within household h was selected at random with possible values of 2, 3 or 4. A new panel of households has been generated in each of 11 quarters and a sample of these households has been observed following the Israel Labor Force Survey rotation scheme of two quarters in the sample, two quarters out of the sample and two quarters in again. As easily checked, this process yields a complete sample of four panels in each of the quarters 6-11, with one panel in each quarter observed for the first time, one panel observed for the second time, one for the third time and one for the fourth and last time. (In the first quarter there is only one panel, in the next three quarters there are two panels and in the fifth quarter there are 3 panels.) In what follows we only consider the data observed for quarters 6-11.

B) Population model

The model used for generating the y-values for a given household h is defined by (3.1) and (3.4) with $x'_{hjt} \equiv (x_{hj1}, x_{hj2})$ and $z'_{ht} \equiv (3, z_{h2})$, such that the covariate values are fixed over time. The x-values were generated independently from the uniform distribution $U[1, 2]$. Values z_{h2} were generated from the uniform distribution $U[1, 5]$. In order to simplify the presentation and evaluation of the results, we also set the model coefficients to be time invariant such that $\gamma_t = \gamma = (6, -2)'$ and $\nu_t = \nu = (1, 2)'$. The random error terms were generated independently between households using the model (3.4) with $A = \text{diag}[0.5, 0.7]$, $\Delta = \text{diag}[0.8, 0.5]$, ρ=0.4 and $\sigma_e^2 = 0.25$. Notice from (4.1) that $\text{Var}(u_{ht}) = \Delta^* = \text{diag}[1.067, 0.980]$ and $\text{Var}(e_{hjt}) = \sigma_e^2 = 0.298$.

C) Sample selection

We consider two separate sampling schemes.

C1) Noninformative sampling:

Population values have been generated for panels of 30 households, with all the households belonging to a given panel selected to the sample and observed following the sample rotation scheme described in A above. The total number of sampled households in each of the quarters 6-11 is therefore $m = 120$. All the individuals belonging to a given household have been observed, yielding an expected sample size of $n = 360$ individuals for each of the quarters. This sampling scheme corresponds to simple random sampling of households and individuals within the selected households.

C2) Informative sampling

Population values have been generated for panels of 55 households. Households with random effects $u_{h1,1} < 0$ (the value of the first random effect at the first time point) have been sampled with probability 1, households with random effects $u_{h1,1} > 0$ have been sampled independently (Poisson sampling) with probability 0.1. All the individuals belonging to a sampled household have been observed. This sampling scheme yields an expected sample size of approvimately 30 households per panel and expected sample sizes of approximately $m = 120$ households and $n = 360$ individuals per quarter, similarly to the sampling scheme C1.

Comment 2: It should be emphasized that even though there are 4 panels observed in each of the quarters 6-11, there are only 11 separate panels that are used for estimation of the model parameters. Moreover, out of the 11 panels, only the panel entering the sample in quarter 6 for the first time is observed in 4 quarters, only 2 panels are observed in 3 quarters, 6 panels are observed in 2 quarters and 2 panels are observed in only one quarter. This implies a total of 13 panel transitions, with about 390 household transitions observed for estimation of the time series parameters. (By a panel transition we mean that the same panel is observed on two occasions. For 3 of these panel transitions there is a time gap of 2 quarters between the two observations). We refer to this sample structure when assessing the estimation of the time series model parameters.

The whole process of generating population values and selecting the sample has been repeated 100 times for each of the two sampling schemes C1 and C2, with one sample selected from each population. For each sample we applied

the two estimation procedures described in section 4. The simulations were run using the Gauss software package. Maximization of the likelihood has been carried out using the numerical optimization procedure, OPTMUM.

5.2 Results

The results of the simulation study are summarized in Tables 1-4 as averages over the 100 samples selected under the two sampling schemes. Each table contains the mean estimates of the model parameters, the empirical standard deviations (SD) of the estimators and the conventional t-statistics obtained by dividing the difference between the mean estimates and the true parameter values by the standard errors (SE), computed as SD/10. Notice that the estimates of the fixed vector coefficients $\beta_t = (\gamma_t', \nu_t')'$ are the same under the two estimation methods.

Perhaps the most important outcome of this study, revealed from Table 1, is that under noninformative sampling it is indeed possible to fit successfully simple but nontrivial time series models to very short longitudinal series, provided that the number of observed series is sufficiently large. (The model is not trivial because even after subtracting the fixed effects, the dependent response variable is the sum of three AR(1) processes.) This conclusion is further strengthened by the fact that 8 out of the 11 panels have been observed for at most 2 times, yielding a total of 13 panel transitions, three of which with a gap of 2 quarters. See Comment 2 at the end of section 5.1.

Table 1
Means, Standard Deviations (SD) and t-Statistics of Estimators Under Two Estimation Methods. Noninformative Sampling

		Method 1			Method 2		
Parameter	True Value	Mean	SD	t-statistic	Mean	SD	t-statistic
γ_1	6.000	6.002	0.03	0.677	6.002	0.03	0.677
γ_2	-2.000	-2.000	0.03	0.078	-2.000	0.03	0.078
ν_1	1.000	0.989	0.08	-1.357	0.989	0.08	-1.357
ν_2	2.000	2.008	0.08	0.997	2.008	0.08	0.997
A_{11}	0.500	0.497	0.07	-0.391	0.491	0.07	-1.271
A_{22}	0.700	0.696	0.07	-0.532	0.695	0.07	-0.820
Δ_{11}^*	1.067	1.054	0.08	-1.668	1.045	0.08	-2.677
Δ_{22}^*	0.980	0.991	0.10	1.042	0.990	0.11	0.906
ρ	0.400	0.398	0.02	-0.937	0.397	0.02	-1.637
σ_e^2	0.298	0.298	0.01	-0.062	0.297	0.01	-1.382

Evaluation of the performance of the two sets of estimators in Table 1 shows that all the estimators under Method 1 are highly insignificant based on the conventional t-statistics and only the estimator of Δ_{11}^* is significant under Method 2. Note that even in that case the absolute relative bias is about 2% and considering that MLE of time series parameters are generally not strictly unbiased, such a small bias in one of 10 parameters is expected. Notice also that the standard errors of the mean estimators under the two methods are very similar, a result observed also in the other tables.

Next we consider the case of informative sampling. Table 2 shows the results obtained when ignoring the informative sampling process, using the same estimation procedures as used for the noninformative case. As indicated very clearly by this table, some of the parameter estimates are highly significant, particularly the estimators of the parameters indexing the time series model of the random effects u_{h1t} that define the sample selection probabilities. Thus, we find that the absolute relative bias in estimating ν_1 is about 27%, and large absolute relative biases are also observed for the estimators of A_{11} and Δ_{11}^*. (The model defined by (3.1) can be rewritten as $y_{hjt} = x_{hjt}'\gamma_t + z_{ht}' u_{ht}^* + e_{hjt}$ where $u_{ht}^* = u_{ht} + \nu_t$, such that for $\nu_t \equiv \nu$ as under the simulation model, $\nu_1 = E(u_{h1t}^*)$). Note that the three biases are negative, which is explained by the fact that the selection mechanism utilized for this study oversamples individuals with observations that contain negative random effects $u_{h1,1}$. In this case again, the two estimation methods perform very similarly.

Table 2
Means, Standard Deviations (SD) and t-Statistics of Estimators Under Two Estimation Methods. Informative Sampling, Unweighted Estimators

		Method 1			Method 2		
Parameter	True Value	Mean	SD	t-statistic	Mean	SD	t-statistic
γ_1	6.000	5.998	0.02	-0.768	5.998	0.02	-0.768
γ_2	-2.000	-2.000	0.03	0.104	-2.000	0.03	0.104
ν_1	1.000	0.728	0.09	-34.385	0.728	0.09	-34.385
ν_2	2.000	2.005	0.09	0.564	2.005	0.09	0.564
A_{11}	0.500	0.438	0.09	-6.742	0.434	0.09	-7.453
A_{22}	0.700	0.738	0.09	4.078	0.735	0.09	3.941
Δ_{11}^*	1.067	0.995	0.09	-7.766	0.994	0.09	-7.883
Δ_{22}^*	0.980	1.003	0.10	2.352	0.987	0.10	0.698
ρ	0.400	0.407	0.02	3.184	0.405	0.02	2.218
σ_e^2	0.298	0.298	0.01	0.644	0.296	0.01	-1.800

Table 3 shows the results obtained when using the PWIGLS algorithm for the estimation of the MLM parameters (section 2.2) and weighting the time series likelihood contributions $\log(L_h) = -\{1/2\, T_h n_h \log(2\pi) + 1/2 \sum_{t=1}^{T_h} \log|F_{ht}| + 1/2 (Y_{ht} - \hat{Y}_{ht|t-1})' F_{ht}^{-1} (Y_{ht} - \hat{Y}_{ht|t-1})\}$ by the household sampling weights $w_h = 1 / \Pr(h \in s)$, using the same 100 samples as used for Table 2. Weighting the likelihood contributions by the inverse of the sample inclusion probabilities is an application of the pseudo likelihood approach that is often recommended for fitting single level models to cross-sectional data, see, *e.g.*, Binder (1983), Skinner *et al.* (1989) and Pfeffermann (1993). As revealed from this table, the use of the PWIGLS algorithm and weighting the likelihood eliminates the large biases observed in Table 2, despite the improper initialization of the Kalman filter with very short series. (See the discussion in Comment 1 at the end of section 4.) Here again, the two estimation methods perform quite similarly, yielding one

biased estimator in each case but with both biases being relatively very small.

It is important to mention that the SD's of the weighted estimators shown in Table 3 are always larger than the corresponding SD's of the unweighted estimators displayed in Table 2. As pointed out by one of the referees, this implies that the empirical root mean square errors (RMSE's) of the unweighted estimators in Table 2 are in fact larger than the empirical RMSE's of the corresponding estimators in Table 3. This outcome, however, is due to the relatively small sample sizes employed in this study. For larger samples (larger numbers of households and individuals within the households) the RMSE is dominated by the bias which, unlike the variance, is not reduced as the sample size increases. Thus, it is clear that as the sample size increases the RMSE's of the weighted estimators become smaller than the RMSE's of the unweighted estimators. The fact that probability weighted estimators have larger variances than the corresponding unweighted estimators is well known from many other studies, see Pfeffermann (1993) for discussion and references.

Table 3
Means, Standard Deviations (SD) and *t*-Statistics of Estimators Under Two Estimation Methods. Informative Sampling, Weighted Estimators

		Method 1			Method 2		
Parameter	True Value	Mean	SD	*t*-statistic	Mean	SD	*t*-statistic
γ_1	6.000	5.997	0.04	-0.607	5.997	0.04	-0.607
γ_2	-2.000	-2.000	0.05	-0.007	-2.000	0.05	-0.007
ν_1	1.000	0.978	0.14	-1.518	0.978	0.14	-1.518
ν_2	2.000	2.019	0.14	1.330	2.019	0.14	1.330
A_{11}	0.500	0.490	0.15	-0.695	0.477	0.14	-1.611
A_{22}	0.700	0.699	0.17	-0.066	0.709	0.16	0.545
Δ_{11}^*	1.067	1.055	0.17	-0.664	1.040	0.17	-1.560
Δ_{22}^*	0.980	1.023	0.19	2.199	1.010	0.19	1.571
ρ	0.400	0.401	0.04	0.135	0.397	0.04	-0.813
σ_e^2	0.298	0.297	0.01	-0.486	0.294	0.01	-3.340

As discussed in Comment 1 at the end of section 4, informative sampling distorts the cross-sectional distribution of the sample observations and the initialization of the Kalman filter, but does not affect the conditional distributions of the first and second level random effects defined by (3.4). Thus, it is interesting to test whether the use of the PWIGLS algorithm for estimating the cross-sectional model parameters but without weighting the time series likelihood likewise controls the bias. Table 4 shows the results obtained for this case with the same samples as used for Tables 2 and 3. The estimators of the fixed vector coefficients $\beta_t = (\gamma_t', \nu_t')'$ are the same as in Table 3 and hence are not shown again. Notice that the estimators of Δ_{11}^*, Δ_{22}^* and σ_e^2 under Method 2 are also the same as the corresponding estimators in Table 3.

The interesting result revealed from Table 4 is that the estimators of A_{11} and A_{22} have now a non-negligible bias, unlike the corresponding estimators in Table 3. This result can be explained as follows. Under the informative sampling scheme, the expectation of the random effects $u_{h1,1}$ corresponding to households h in the sample is below zero, $E(u_{h1,1} \mid h \in s) < 0$, and hence the initialization of the Kalman filter by the population expectation ($Eu_{h1,1} = 0$, Equation 4.4) yields biased estimators. On the other hand, by weighting the likelihood contributions L_h by the inverse of the sample selection probabilities, the proportions of likelihoods L_h corresponding to random effects that are below and above the model expectation is balanced to the population proportions and thus the use of the model expectation for the initialization process does not bias the estimation process. As noticed for the previous tables, the SD's of the unweighted estimators in Table 4 are much smaller than the SD's of the corresponding weighted estimators in Table 3.

Table 4
Means, Standard Deviations (SD) and *t*-Statistics of Estimators Under Two Estimation Methods. Informative Sampling, Weighted MLM, Unweighted Likelihood

		Method 1			Method 2		
Parameter	True Value	Mean	SD	*t*-statistic	Mean	SD	*t*-statistic
A_{11}	0.500	0.468	0.09	-3.477	0.453	0.10	-4.569
A_{22}	0.700	0.742	0.11	3.948	0.737	0.11	3.197
Δ_{11}^*	1.067	1.060	0.11	-0.598	1.040	0.17	-1.560
Δ_{22}^*	0.980	1.008	0.11	2.449	1.010	0.19	1.571
ρ	0.400	0.407	0.02	3.021	0.402	0.02	0.894
σ_e^2	0.298	0.298	0.01	1.013	0.294	0.01	-3.340

6. APPLICATION OF THE MODEL TO LFS DATA

We fitted the model defined by (3.1) and (3.4) to an empirical data set extracted from data collected by the Israel LFS for Jerusalem during the years 1990-1994. The data contain complete records for 567 individuals in 475 households, with each individual observed in four quarters according to the rotation pattern described before and used for the simulation study. Out of the 475 households, 385 have one individual record, 88 have 2 individual records and only 2 households have 3 individual records. The outcome variable is y = number of hours worked during the week preceding the interview, ($\bar{y} = 39.8$, sd $(y) = 14.8$; calculated over all individuals and all the quarters). The individual level auxiliary variables are x_1 = years of education, ($\bar{x}_1 = 13.4$, sd $(x_1) = 4.8$) and x_2 = gender, (41% females). The household level auxiliary variables are $z_1 = 1$ and z_2 = number of employed persons in the household ($\bar{z}_2 = 1.48$, sd$(z_2) = 0.56$).

We estimated the model parameters using the two methods described in section 4. The sampling weights attached to these data are very similar across households and individuals so that we only computed the unweighted

estimators. The LGLS algorithm produced negative variance estimates for Δ^*_{22} in some of the quarters and these estimates have been set to zero when averaging the variance estimates under Method 2. The quarterly estimates of the fixed model coefficients have not been averaged as they change significantly over the five years period.

The estimates computed by the two methods for the variances and autoregression coefficients are shown in Table 5 using the same notation as in the previous tables. The two sets of estimates are not very far except for the estimator of Δ^*_{22} which, has already mentioned was found to be negative in some of the separate IGLS runs. Note in this respect that for most of the households there is only a single individual record (see above), and that for almost all of these households $z_2 = 1$. This complicates the estimation process since for such households it is impossible to distinguish the first (individual) level effect from the two household effects, which are likewise confounded. (Note that the sum of the latter two variances is similar under the two methods.) As discussed below, the estimators in Table 5 are dominated by the observations obtained for households with two individual records.

Table 5
Estimates of Variances and Autoregression Coefficients Under Two Estimation Methods. LFS Data

Parameter	A_{11}	A_{22}	Δ^*_{11}	Δ^*_{22}	ρ	σ^2_e
Method 1	0.915	-0.606	73.88	2.541	0.242	102.306
Method 2	0.976	-0.548	56.88	14.753	0.448	101.001

Under the Israel LFS sampling design, each individual record consists of 4 observations taken in quarters 1,2, 5 and 6, with quarter 1 defining the first calendar quarter t that the individual is in the sample. In order to assess the prediction power of the model, we computed for every individual record (h,j) the empirical innovations when predicting the adjusted values $r_{hjq} = (y_{hjq} - x'_{hjq}\hat{\gamma}_q - z'_{hq}\hat{v}_q)$ using the household data observed for the preceding quarters that the individual has been in the sample. Note that by subtracting the fixed effects from the original observations, the distribution of the adjusted values no longer depends on the calendar quarters. The innovation for quarter q is the corresponding prediction error which, by (3.1) is computed as $d_{hjq} = (r_{hjq} - z'_{hq}\hat{u}_{hq|q-m} - \hat{e}_{hjq|q-m}) = r_{hjq} - (z'_{hq}, 1)'\hat{\alpha}_{q|q-m}$, $q = 2, 5, 6$ where $\hat{\alpha}_{q|q-m}$ is the predictor of the state vector $\alpha_q = (u'_{hq}, e_{hjq})'$ using the data observed until quarter q-m, with $m = q - 1$ for $q = 2,6$ and $m = 3$ for $q = 5$. The predictor $\hat{\alpha}_{q|q-m}$ is obtained by application of the Kalman filter with the corresponding estimated parameters (see section 2.3 and Equations 3.5 and 4.3).

Table 6 shows the roots of the means of the square innovations (RMSI) by quarter and the number of household (HH) records, as obtained under the two estimation methods (using the parameter values displayed in Table 5). For comparison, we also show the RMSI's of the innovations obtained by predicting the adjusted value for quarter q by the adjusted value in the preceding quarter. The "naive" predictor $\hat{r}_{hjq} = r_{hj,q-m}$ can be interpreted as being the optimal predictor under the simple random walk model $r_{hjq} = r_{hj,q-m}$ + error. The means of the innovations $(r_{hjq} - r_{hj,q-m})$ for $q = (2, 5, 6)$ are (0.68, 0.24, 0.301) for households with one record, (1.24, -1.20, 0.60) for households with two records and (4.02, -5.82, 7.68) for households with 3 records but recall that the latter means are based on only 2 households. The corresponding means of the empirical innovations computed under the model are smaller in absolute value in all the cases.

Table 6
Root Mean Square of Innovations by Number of Household Records and Quarter Under Two Estimation Methods and Naive Prediction. LFS Data

HH Records	1			2			3		
Quarter	2	5	6	2	5	6	2	5	6
Method 1	11.54	11.16	11.62	12.26	11.71	10.88	9.61	9.98	8.94
Method 2	11.71	11.16	11.49	12.10	11.48	10.91	9.30	9.78	7.90
Naive Pred.	14.00	11.92	13.60	14.71	15.12	13.47	7.50	13.32	11.29

The data analyzed in this section behave much more erratically than the data used for the simulation study generated under the model and we cannot claim that the model employed yields the best possible fit (see also below). Nonetheless, the values displayed in Table 6 illustrate some important features of the model. We mention first the generally much better performance of the model predictors compared to the naive predictor $\hat{r}_{hjq} = r_{hj,q-m}$, with the two estimation methods yielding similar RMSI's. The superiority of the model is explained by the fact that whereas the first order autocorrelations of the two random household effects used for the model predictions are high in absolute value (very high for the first component), the autocorrelations of the adjusted values (the "total" errors) are only of moderate size. The first order autocorrelations of the random components are the corresponding autoregression coefficients, see Table 5. The empirical autocorrelations of the adjusted values, $\mathrm{Corr}(\hat{r}_{hjq}, \hat{r}_{hj,q-m})$, $q = 2, 5, 6$; $m = 1$ for $q = 2,6$; $m = 3$ for $q = 5$ are correspondingly (0.46, 0.59, 0.51) for one record households, (0.48, 0.36, 0.45) for two record households and (0.92, 0.43, 0.63) for three record households (based on 6 individual records).

As already noted, the fact that most households have only one individual record introduces identifiability problems since for such households it is impossible to distinguish between the three random effects. Computation of the correlations $\mathrm{Corr}(\hat{r}_{hjq}, \hat{r}_{hj,q-m})$, under the model using the parameter estimates in Table 5 shows a good fit to the correlations computed for two record households. This in turn illustrates that the estimators in Table 5 are

dominated by these observations and we conclude that the model fits best the observations obtained for the households with two records. Note, however, that the RMSI's obtained for the other household sizes are not higher than the RMSI's computed for the two record households (see also below). It is important to mention in this regard that if the data had been aggregated over all the individuals observed in a given calendar quarter, it would have been impossible to account for the random household effects, resulting in inferior predictions of the individual observations. See the discussion in the introduction. (Modelling the aggregate data is rather complicated in this case since the sample in each calendar quarter consists of 4 different panels as defined by the number of times that individuals are in the sample. This implies that the models holding for these panels are different, depending on the number of observations available for each panel.)

Other interesting results noted in Table 6 are that the RMSI's under the model are generally lower for $q = 6$ than for $q = 2$, as explained by the use of more observed data for the same individual in the prediction process (more observed data for estimating the random effects in the preceding quarter). Also, for $q = 6$ the RMSI's decrease as the number of household records increases, as explained by the use of data observed for other household members. Finally, the RMSI's for households with 3 records are much lower by use of the model than the RMSI's obtained for households with 1 and 2 records but we mention again that there are only 2 households with three records. The unexpected results in Table 6 are that for households with one record the RMSI's are somewhat larger for $q = 6$ than for $q = 5$ (note the relatively high and unexplained correlation of 0.59 between the adjusted values 3 quarters apart computed for these households), and that for $q = 2$ and $q = 5$ the RMSI's for households with 2 records are larger than the corresponding RMSI's for households with 1 record. With empirical data of relatively small size such anomalies are not unusual and they show up even more prominently with the naive predictor. (The fact that for a given number of household records the RMSI's by use of the model for $q = 5$ are of similar magnitude to the other RMSI's is reassuring given that the predictions in this case are 3 quarters ahead.)

7. CONCLUSIONS AND MODEL EXTENSIONS

The results of this paper illustrate that it is possible to fit time series models to longitudinal series of very short length and with missing observations. The model used in the present study is an extension of the standard two level linear model by which both the first and second level random effects evolve stochastically over time. This kind of model is suitable for modelling longitudinal measurements that are taken for hierarchical populations. Application of the PWIGLS algorithm combined with standard probability weighting of the time series likelihood is shown to protect against the effects of informative sampling.

Multilevel models are often fitted to discrete data, in which case the models contain nonlinear components. In principle, the two-stage estimation method proposed in this paper can be applied in this case as well, although with very short longitudinal series the range of models that can be fitted is obviously limited. Moreover, a common procedure for estimating the unknown model parameters in the discrete case consists of linearizing the nonlinear components on each iteration of the IGLS around estimates obtained on the previous iteration, and then applying the standard IGLS for computing the revised estimates. See Goldstein (1995) for details. Thus, it seems feasible to extend the PWIGLS algorithm to the discrete case without major difficulties.

In this paper we have not considered variance estimation. This is no problem under the standard IGLS and Pfeffermann *et al.* (1998) propose simple variance estimators for the PWIGLS procedure. However, estimation of the variances of estimators obtained from maximization of the time series likelihood is more problematic because of two reasons. First, the possibly short length of the longitudinal series may no longer justify the use of the information matrix or permit stable estimation thereof, even with large number of second level units. Second, the MLM estimators are held fixed when maximizing the likelihood, implying that the MLE abstract from the sampling errors in the estimation of the MLM parameters. A possible solution to this problem is the use of re-sampling methods that allow to account for all sources of variation in the estimation process.

Finally, we mention an important application of the proposed model for the imputation of missing data. In a recent article, Pfeffermann and Nathan (forthcoming) illustrate the large reductions in the imputation variance that can be achieved under the model compared to the use of more standard imputation methods that ignore the common household effects.

REFERENCES

BELCHER, J., HAMPTON, J.S., and TUNNICLIFFE WILSON, G. (1994). Parameterization of continuous time autoregressive models for irregularly sampled time series data. *Journal of the Royal Statistical Society*, Series B, 56, 141-155.

BINDER, D.A. (1983). On the variances of asymptotically normal estimators from complex surveys. *International Statistical Review*, 51, 279-292.

BINDER, D.A. (1998). Longitudinal surveys: why are these surveys different from all other surveys? *Survey Methodology*, 24, 101-108.

BRYANT, J., and DAY, R. (1991). Empirical Bayes analysis for systems of mixed models with linked autocorrelated random effects. *Journal of the American Statistical Association*, 86, 1007-1012.

CHI, E.M., and REINSEL, G.C. (1989). Models for longitudinal data with random effects and AR(1) errors. *Journal of the American Statistical Association*, 84, 452-459.

DIGGLE, P.J., LIANG, K.Y., and ZEGER, S.L. (1994). *Analysis of Longitudinal Data*. Oxford: Clarendon Press.

DUNCAN, G.J., and KALTON, G. (1987). Issues of design and analysis of surveys across time. *International Statistical Review*, 55, 97-117.

GOLDSTEIN, H. (1986). Multilevel mixed linear model analysis using iterative generalized least squares. *Biometrika*, 73, 43-56.

GOLDSTEIN, H. (1995). *Multilevel Statistical Models* (2nd edition). New York: Halstead.

GOLDSTEIN, H., HEALY, M.J.R., and RASBASH, J. (1994). Multilevel time series models with applications to repeated measures data. *Statistics in Medicine*, 13, 1643-1655.

HARVEY, A.C. (1989). *Forecasting, Structural Time Series Models and the Kalman Filter*. New York: Cambridge University Press.

HERRIOT, R.A., and KASPRZYK, D. (1984). The survey of income and program participation. *Proceedings of the Social Statistics Section, American Statistical Association*, 107-116.

JONES, R.H., and ACKERSON, L.M. (1990). Serial correlation in unequally spaced longitudinal data. *Biometrika*, 77, 721-731.

JONES, R.H., and BOADI-BOATING, F. (1991). Unequally spaced longitudinal data with AR(1) serial correlation. *Biometrics*, 47, 161-175.

JONES, R.H., and VECCHIA, A.V. (1993). Fitting continuous ARMA models to unequally spaced spatial data. *Journal of the American Statistical Association*, 88, 947-954.

JONES, R.H. (1993). *Longitudinal Data with Serial Correlation. A State-space Approach*. New York: Chapman and Hall.

LAWLESS, J.F. (1999). Event History Analysis and Longitudinal Surveys. Paper presented at the Conference on Analysis of Survey Data, Southampton, United Kingdom.

NATHAN, G. (1999). A Review of Sample Attrition and Representativeness in Three Longitudinal Surveys. GSS Methodology Series No. 13. London Office of National Statistics (ONS), United Kingdom.

PFEFFERMANN, D. (1993). The role of sampling weights when modeling survey data. *International Statistical Review*, 61, 317-337.

PFEFFERMANN, D. (1996). The use of sampling weights for survey data analysis. *Statistical Methods in Medical Research*, 5, 239-261.

PFEFFERMANN, D., SKINNER, C.J., GOLDSTEIN, H., HOLMES, D.J., and RASBASH, J. (1998). Weighting for unequal selection probabilities in multilevel models (with discussion). *Journal of the Royal Statistical Society*, Series B, 60, 23-40.

PFEFFERMANN, D., and NATHAN, G. (forthcoming). Imputation for wave nonresponse: existing methods and a times series approach. To appear in: *Survey Nonresponse*, (Eds, R.M. Groves, D. Dillman, J.L. Eltinge, and R.J.A. Little). New York: John Wiley and Sons.

RAO, J.N.K. (1999). Some recent advances in model-based small area estimation. *Survey Methodology*, 25, 175-187.

SALLAS, W.H., and HARVILLE, D. A. (1981). Best linear recursive estimation for mixed linear models. *Journal of the American Statistical Association*, 76, 860-869.

SKINNER, C.J., HOLT, D., and SMITH, T.M.F. (Eds.) (1989). *Analysis of Complex Surveys*. Chichester: Wiley.

SKINNER, C.J., and HOLMES, D. (1999). Random Effects Models for Longitudinal Survey Data. Paper presented at the Conference on Analysis of Survey Data, Southampton, United Kingdom.

SURVEY RESEARCH CENTER (1984). User Guide to the Panel Study of Income Dynamics. Ann Arbor, Michigan: Inter-university Consortium for Political and Social Research.

WEBBER, M. (1994). The survey of labor and income dynamics: lessons learned in testing. *Proceedings of the Annual Research Conference, US Bureau of the Census*, 85-99.

ZIMMERMAN, D.L., and NUNEZ-ANTON, V. (1997). Structured antedependence models for longitudinal data. In: *Modelling Longitudinal and Spatially Correlated Data: Methods, Applications and Future Directions*, (Eds. T.G. Gregoire *et al.*). Lecture Notes in Statistics, 22. New York: Springer Verlag, 62-76.

Survey Methodology, June 2000
Vol. 26, No. 1, pp. 67-78
Statistics Canada

A Conditional Mean Squared Error of Small Area Estimators

LOUIS-PAUL RIVEST and EVE BELMONTE[1]

ABSTRACT

This paper suggests estimating the conditional mean squared error of small area estimators to evaluate their accuracy. This mean squared error is conditional in the sense that it measures the variability with respect to the sampling design for a particular realization of the smoothing model underlying the small area estimators. An unbiased estimator for the conditional mean squared error is easily constructed using Stein's Lemma for the expectation of normal random variables. This estimator can be calculated for any shrinking strategy; composite and empirical Bayes estimators are considered in this work. It can be calculated when the small area estimators are benchmarked to coincide with direct estimators at high level of aggregation. It can accommodate skewness in the data and estimated variances. The conditional mean squared error estimator does not rely on any smoothing model. The price to pay for this property is a high variance; the new estimator is unstable under heavy shrinking. In these situations, it still provides useful diagnostic information about the shrinking model. It can also be seen as a building block for estimators of unconditional mean squared errors such as Prasad and Rao's (1990). Examples dealing with the estimation of the under-coverage in the Canadian Census illustrate the application of this new estimator.

KEY WORDS: Census under-coverage; Diagnostics; Empirical Bayes estimation; Estimated variances; Skewness; Stein's lemma; Survey sampling.

1. INTRODUCTION

In survey sampling, the need to develop accurate methods of estimation for small areas poses challenging statistical problems. For small areas, direct survey estimates have too large a variance to be reliable. Small area techniques "improve" direct estimates by shrinking them towards model based smoothed values. Simple shrinking estimators are proposed by Purcell and Kish (1979). In a pioneering paper, Fay and Herriot (1979) demonstrate that this can lead to interesting gains in precision. The review papers of Ghosh and Rao (1994) and of Singh, Gambino and Mantel (1994) provide convincing evidence of the vitality of this area.

The estimation of the errors in small area estimation is receiving an increasing attention, see Singh, Stukel and Pfeffermann (1998) and Booth and Hobert (1998). This paper suggests estimating the conditional mean squared errors of small area estimators. The conditional mean squared error can be estimated for all shrinking strategies, either empirical Bayes or decision theoretic (Purcell and Kish 1979). Other mean squared errors, such as Prasad and Rao's (1990), and Singh, Stukel and Pfeffermann (1998) frequentist proposals measure the variability with respect to both, the sampling design and the smoothing model. The mean squared error of this paper is conditional in the sense that it measures variability with respect to the sampling design for a particular realization of the smoothing model. This feature is attractive since the conditional estimator reflects the conditions under which the survey has been carried out (see Särndal, Swensson, and Wretman 1992, ch. 7). The drawback of this property is a high variability. In some instances, the proposed estimator is too variable for practical use.

When shrinking is important, the conditional mean squared error estimators are highly unstable. An unconditional assessment of the precision of small area estimators must be used. In this situation, the conditional estimator proposed in this paper still provides some useful information. It can be looked at as a diagnostic for comparing smoothing models. It can also be a building block for constructing Monte Carlo estimates of unconditional mean squared errors in situations where closed form formulas, such as Prasad and Rao's (1990), are not available.

The assessment of the accuracy of estimators for the under-coverage, at the provincial and sub-provincial levels, of the Canadian Census motivated this work. Alternatives to the direct estimates for provincial under-coverage are discussed by Royce (1992) and Rivest (1995). Dick (1995) applies empirical Bayes methods to sub-provincial under-coverage estimates. These two examples are treated in section 5.

An estimator of the conditional mean squared error is presented in section 2. Its construction relies on the multivariate version of Stein's Lemma for the expectation of normal deviates. Section 3 suggests changes to the conditional estimator to accommodate skewness in the distribution of the direct estimators and estimated variances. Section 4 discusses the application of the new estimator to empirical Bayes estimators. Its relationship with Prasad and Rao (1990) prediction variance is highlighted. Examples are treated in section 5.

[1] Louis-Paul Rivest and Eve Belmonte, Département de mathématiques et de statistique, Université de Laval, Ste-Foy, Québec, Canada, G1K 7P4.

2. A CONDITIONAL MEAN SQUARED ERROR ESTIMATOR

Suppose that there are n small areas and let $\mu = (\mu_1, ..., \mu_n)'$ denote the unknown population characteristics for these small areas. The direct survey estimates for the n small areas are $y = (y_1, ..., y_n)'$ where the distribution of y is $N_n(\mu, \Sigma)$, a n-variate normal distribution with mean vector μ and known variance-covariance matrix Σ. As pointed out by Ghosh and Rao (1994), the normality assumption is likely to hold for many surveys since direct survey estimates are usually functions of sums of variables. The $n \times n$ matrix Σ is a design based measure of precision for y. For the time being, this matrix is assumed to be known. This assumption is relaxed in section 3.2. The uncertainty in y comes from the random selection of the sampling units. Subscript S, for sampling design, denotes expectations taken with respect to the distribution of y.

In a typical application of small area techniques, one has,

$$y_i = \frac{\sum_j w_{ij} y_{ij}}{\sum_j w_{ij}}$$

where y_{ij} is the y-value for the j-th sample unit in small area i, w_{ij} is its sampling weight and the sum is over all the sample units in small area i. In many instances, the variance covariance matrix Σ is diagonal; its (i, i) term, is $\sigma_{ii} = \mathrm{Var}_S(y_i)$; when they are non null, the off diagonal elements of Σ are denoted by $\sigma_{ij}, i, j = 1, ..., n$.

Several methods have been proposed to improve the accuracy of direct survey estimators. They involve shrinking y_i towards some indirect estimator of μ_i. The resulting estimators can be written as

$$\hat{\mu}_i = y_i + g_i(y_1, ..., y_n), \quad i = 1, ..., n \tag{1}$$

where the g_i's are functions depending on the shrinking strategy.

In vector form, one can write (1) as $\hat{\mu} = y + g(y)$ where g, whose i-th component is equal to g_i, is a function defined from R^n to R^n. We assume that for each i, the right partial derivative and the left partial derivative of g_i with respect to y_j exists for any y in R^n. When they are equal, $\partial g_i(y)/\partial y_j$ denotes the common value; if they differ $\partial g_i(y)/\partial y_j$ is the average between the two values. The component of $g(y)$ and their partial derivatives are assumed to have finite variances. A conditional assessment of the precision of $\hat{\mu}$ as an estimator for μ is given by the matrix of the mean product errors which is given by

$$E_S\{(\hat{\mu} - \mu)(\hat{\mu} - \mu)'\} = \Sigma + E_S\{(y - \mu)g(y)'\} + E_S\{g(y)(y - \mu)'\} + E_S\{g(y)g(y)'\}$$

On the right hand side of this equality, the only quantities for which there are no obvious estimators are $E_S\{(y-\mu)g(y)'\}$ and $E_S\{g(y)(y-\mu)'\}$. Their evaluations are eased by the following result which is a multivariate extension of Stein's lemma (Stein 1981). Its proof is given in the appendix together with the proofs for Propositions 2, 3, and 4.

PROPOSITION 1: Let y be a $N_n(\mu, \Sigma)$ random vector then,

$$E_S\{(y - \mu)g(y)'\} = \Sigma E_S\{\nabla g(y)\},$$

where $\nabla g(y)$ is an $n \times n$ matrix whose (i,j)-th element is given by $g_i^j(y) = \partial g_i(y)/\partial y_j$.

Now according to Proposition 1, $\Sigma \nabla g(y)$ is an unbiased estimator for $E_s\{(y - \mu)g(y)'\}$. Thus the conditional estimator (index "c" stands for conditional) for the matrix of the mean product errors is given by

$$\mathrm{mpe}_c(\hat{\mu}) = \Sigma + \Sigma \nabla g(y) + \nabla g(y)' \Sigma + g(y)g(y)'. \tag{2}$$

The diagonal terms of (2) can be negative. Since they estimate mean squared errors, a better estimator for the mean squared error of $\hat{\mu}_i$ is

$$\mathrm{mse}_c^+(\hat{\mu}_i) = \max\left(0, \sigma_{ii} + \sum_j \sigma_{ij}\left\{g_j^i(y) + g_i^j(y)\right\} + g_i(y)^2\right).$$

It generalizes an estimator proposed by Bilodeau and Srivastava (1988) for James-Stein estimator, and by Robert (1992 p. 292) for empirical Bayes estimators. When the y_i's are independent, with $\sigma_{ij} = 0$ when $i \neq j$, then

$$\mathrm{mse}_c(\hat{\mu}_i) = \sigma_{ii} + 2\sigma_{ii}\frac{\partial g_i(y)}{\partial y_i} + g_i(y)^2, \tag{3}$$

and $\mathrm{mse}_c^+(\hat{\mu}_i) = \max\{\mathrm{mse}_c(\hat{\mu}_i), 0\}$.

Kott's (1989) small area estimator has $g_i(y) = \hat{\alpha}_i(\hat{\gamma}_i - y_i)$, where $\hat{\gamma}_i$ is a measure of location for the y's and $\hat{\alpha}_i$ is a smoothing parameter. These two statistics involve variance estimates calculated at the "unit" level, that is using the y_{ij}'s. Kott's (1989) conditional mean squared error is

$$v(\hat{\mu}_i) = \sigma_{ii}(1 - 2\hat{\alpha}) + (\hat{\alpha}_i(y_i - \hat{\gamma}_i))^2.$$

This is equal to (3) when both $(d/dy_i)\hat{\alpha}_i$ and $(d/dy_i)\hat{\gamma}_i$ are null. Thus Kott's (1989) estimator for the conditional mean squared error does not account for the estimation for the variance components. This may account for the biases that it exhibited in the simulations reported by Prasad and Rao (1999).

The estimates mse_c and mpe_c can be evaluated numerically by taking

$$\frac{\partial g_i(y)}{\partial y_j} = \frac{g_i(y_1, ..., y_{j-1}, y_j + \epsilon, y_{j+1}, ..., y_n) - g_i(y_1, ..., y_{j-1}, y_j - \epsilon, y_{j+1}, ..., y_n)}{2\epsilon}$$

where ϵ is a small positive number. Thus mse_c and mpe_c can be calculated in all circumstances, even when g has no explicit form.

To illustrate the flexibility of the conditional estimator, consider $\hat{\mu}^* = \hat{\mu}(\Sigma y_i)/(\Sigma\hat{\mu}_i)$, an estimator bench-marked to agree with the direct estimator for the y-total. One has $\hat{\mu}^* = y + g^*(y)$ where

$$g^*(y) = \frac{\sum y_i}{\sum \hat{\mu}_i} g(y) + \left(\frac{\sum y_i}{\sum \hat{\mu}_i} - 1 \right) y.$$

It might be difficult to derive an analytical formula for $\text{mpe}_c(\hat{\mu}^*)$, however this expression is easily evaluated using numerical derivatives. Modifications of the conditional estimator to account for non-normality in the y_i's and for estimated variances σ_{ii} are given next.

3. SENSITIVITY ANALYSIS

In many surveys, especially those in the business sector, the study variables are skewed. Some of this skewness might still be left in the direct estimators y_i. This section suggests a correction to the conditional mean squared error to account for skewness in the distribution of y. It also proposes ways to account for the estimation of the variances σ_{ii} in the mean squared error calculations.

In practice the variances σ_{ii} are estimated. Several authors (Dick 1995; Hogan 1992) smooth the variances before calculating the small area estimates. They then consider the smoothed variances as the true variances in the small area calculations. Section 3.2 gives a condition under which replacing the estimated variances by their smoothed values yields unbiased mean squared error estimators. It also consider situations where the sampling variances are estimated with random groups (Wolter 1985 ch.2). This method consists in carrying a certain number, say k, of replications of the survey design. This yields, for each i, k estimates of μ_i; $\hat{\sigma}_{ii}$ is then equal to the sampling variance of these k estimates divided by k. Assuming that these k estimates are normally distributed, one can consider that, suitably normalized, the distribution of $\hat{\sigma}_{ii}$ is chi-squared with $k-1$ degrees of freedom. A conditional mean squared error, adjusted for variances estimated with random groups, is proposed in this section. To keep the discussion simple, we assume in this section that Σ is a diagonal matrix; in other words the y_i's are assumed to be independent random variables.

3.1 Non-Normality in the Distribution of y_i

In many applications of small area estimation, the distributions of the y_i's are not exactly normal. A simple adjustment to (3) is proposed to deal with asymmetry in the distribution of the y_i's.

Suppose that the skewness of y_i, $\rho_i = E_S\{(y_i - \mu_i)^3\}/\sigma_{ii}^{3/2}$ is small and non-zero. A first order Edgeworth series for the distribution of y_i is given by (see for instance Reid 1991):

$$f(t) = \frac{\exp\{-(t-\mu_i)^2/(2\sigma_{ii})\}}{\sqrt{(2\sigma_{ii}\pi)}} \times \left[1 + \frac{\rho_i}{6}\left\{ \left(\frac{t-\mu_i}{\sqrt{\sigma_{ii}}} \right)^3 - 3\left(\frac{t-\mu_i}{\sqrt{\sigma_{ii}}} \right) \right\} \right].$$

Such an expansion is used to correct for skewness in the direct estimators (Barndorff-Nielsen and Cox 1989, remark 2 p. 92). Expansions involving additional terms are used for correcting for both skewness and kurtosis; they will not be considered in this section. The evaluation of $E\{(y_i - \mu_i)g_i(y)\}$ under f, needed for the construction of the conditional mean squared error estimator, is given in Proposition 2.

PROPOSITION 2: When y_i distributed according to $f(t)$, as ρ_i tends to 0.

$$E_S\{(y_i - \mu_i)g_i(y)\} = \sigma_{ii} E_S\left\{ \frac{\partial g_i(y)}{\partial y_i} \right\} + \frac{\sigma_{ii}^{3/2}\rho_i}{2} E_S\left\{ \frac{\partial^2 g_i(y)}{\partial y_i^2} \right\} + O(\rho_i).$$

A mean squared error estimator corrected for asymmetry is therefore given by $\text{mse}_c^*(\hat{\mu}_i) = \max\{0, \text{mse}_c(\hat{\mu}_i)\}$ where

$$\text{mse}_c(\hat{\mu}_i) = \sigma_{ii} + 2\sigma_{ii}\frac{\partial g_i(y)}{\partial y_i} + \sigma_{ii}^{3/2}\rho_i\frac{\partial^2 g_i(y)}{\partial y_i^2} + g_i(y)^2.$$

In practice, it might be difficult to find individual skewness coefficients ρ_i for each i. A better strategy might be to combine all the data points to come up with a common ρ-value.

3.2 Estimated Variances

Consider first a survey where the $\hat{\sigma}_{ii}$'s are estimated using k random groups. Assuming normality, one can consider that $\{(k-1)\hat{\sigma}_{ii}/\sigma_{ii}: i = 1, ..., n\}$ is a sequence of independent χ^2_{k-1} random variables which is independent of y. Evaluating the conditional mean squared error (3) with variance estimates $\hat{\sigma}_{ii}$ yields potentially biased estimators, since $g_i(y)$ and its derivatives depend on $\hat{\sigma}_{ii}$. The potential bias can be expressed as

$$2E\left\{ \hat{\sigma}_{ii}\frac{\partial g_i(y)}{\partial y_i} \right\} - 2\sigma_{ii}E\left\{ \frac{\partial g_i(y)}{\partial y_i} \right\} \qquad (4)$$

As shown in the Appendix, this bias is $O(1/k)$. The next proposition suggests a small change to (3) that reduces its bias (4).

PROPOSITION 3: Replacing $\hat{\sigma}_{ii}$ by $(k-1)\hat{\sigma}_{ii}/(k+1)$ in the evaluation of $\partial g_i(y)/\partial y_i$ for calculating the mean squared error estimator (3) yields an estimator with an $O(1/k^2)$ bias.

The correction factor $(k-1)/(k+1)$ has been proposed in a different context by Scott and Smith (1971). Other methods are available for correcting the bias for estimating variances, depending on the way in which σ_{ii} is estimated. For instance if the $\hat{\sigma}_{ii}$ are independent $N\{\sigma_{ii}, \text{var}(\hat{\sigma}_{ii})\}$ random variables distributed independently of y, then by Stein's lemma, (4) is equal to $2\text{var}(\hat{\sigma}_{ii})E\{\partial^2 g_i(y)/\partial y_i \partial \hat{\sigma}_{ii}\}$.

Suppose now that the variances are estimated, not necessarily with random groups. In surveys, such as those considered in Dick (1995) and Hogan (1992), explanatory variables are available to model estimated variances. Small area estimators are then calculated with the predicted variances $\tilde{\sigma}_{ii}$ under the smoothing model; this means that $\tilde{\sigma}_{ii}$ enters in the calculation of g_i in (1). Considering (4), the mean squared error estimated with the smoothed variance,

$$\tilde{\sigma}_{ii} + 2\tilde{\sigma}_{ii}\frac{\partial g_i(y)}{\partial y_i} + g_i(y)^2$$

is unbiased provided that

$$2E\left\{(\tilde{\sigma}_{ii} - \sigma_{ii})\frac{\partial g_i(y)}{\partial y_i}\right\} = 0.$$

When $g_i(y)$ is calculated with smoothed variances, (4) should be small; the above condition holds provided that

$$E_V\left\{(\hat{\sigma}_{ii} - \tilde{\sigma}_{ii})\frac{\partial g_i(y)}{\partial y_i}\right\} = 0, \qquad (5)$$

where index V refers to the model for smoothing the variances. One can easily test whether this condition holds by calculating the correlation between the variance residuals and the partial derivatives of the functions g_i. Since, as shown in Proposition 5 of the next section, unconditional mean squared errors can be derived as expectations of $\text{mse}_c(\hat{\mu}_i)$ testing whether (5) is true is relevant even when unconditional measures of accuracy, such as Prasad and Rao's, are calculated. Indeed, replacing variances by their predicted values biases the mean squared error estimators, conditional or unconditional, when (5) is violated.

4. MEAN SQUARED ERROR ESTIMATION FOR EMPIRICAL BAYES ESTIMATORS

4.1 Model Construction

This section assumes that the y_i's are independent, *i.e*, that Σ is diagonal. In an empirical Bayes setting, the model (M) for smoothing direct estimators expresses the parameters μ_i's as random variables whose distributions depend on a p-variate auxiliary variable x_i (Maritz and Lwin 1989, chapter 3),

$$\mu_i = x_i'\beta + v_i, \qquad (6)$$

where β is a $p \times 1$ vector of unknown regression parameters and the v_i's are independent random variables with mean 0 and variance σ_v^2. Often the v_i's are assumed to be normally distributed; the marginal distribution of y_i, with respect to both the sampling design S and the smoothing model M, is then $N(x_i'\beta, \sigma_{ii} + \sigma_v^2)$. The empirical Bayes estimators are obtained by shrinking the direct estimators y_i towards their predicted values under (6).

The extent of the shrinking depends on estimators of the parameters of (6) calculated from the marginal distribution of y_i. Several methods are available for parameter estimation (Cressie 1992). A popular estimator for σ_v^2 (see Lahiri and Rao (1995)) is

$$\hat{\sigma}_v^2 = \max\left[0, (n-p)^{-1}\left\{\sum_{i=1}^{n}(y_i - x_i'\hat{\beta})^2 - \sum_{i=1}^{n}\sigma_{ii}(1-h_{ii})\right\}\right]$$

where $\hat{\beta} = (X'X)^{-1}X'y$, $h_{ii} = x_i'(X'X)^{-1}x_i$, and $X = (x_1, \ldots, x_n)'$. The weighted least squares estimator of β is

$$\hat{\beta}_w = \hat{A}^{-1}\sum_{i=1}^{n}\frac{x_i y_i}{(\hat{\sigma}_v^2 + \sigma_{ii})},$$

where

$$\hat{A} = \sum_{i=1}^{n}\frac{x_i x_i'}{(\hat{\sigma}_v^2 + \sigma_{ii})}.$$

The empirical Bayes estimator for μ_i is then

$$\hat{\mu}_i = x_i'\hat{\beta}_w + \frac{\hat{\sigma}_v^2}{\hat{\sigma}_v^2 + \sigma_{ii}}(y_i - x_i\hat{\beta}_w) = y_i - \frac{\sigma_{ii}}{\hat{\sigma}_v^2 + \sigma_{ii}}(y_i - x_i'\hat{\beta}_w). \qquad (7)$$

Thus for empirical Bayes estimators, one has

$$g_i(y) = -\frac{\sigma_{ii}}{\hat{\sigma}_v^2 + \sigma_{ii}}(y_i - x_i'\hat{\beta}_w).$$

4.2 The Conditional Mean Squared Error Estimator

An explicit form for (3) can be obtained from the following formula for the derivative of the functions g_i for empirical Bayes estimators,

$$\frac{\partial g_i(y)}{\partial y_i} = \frac{\partial \hat{\sigma}_v^2}{\partial y_i}\frac{\partial g_i(y)}{\partial \hat{\sigma}_v^2} - \frac{\sigma_{ii}}{\hat{\sigma}_v^2 + \sigma_{ii}}\left\{1 - \frac{x_i'\hat{A}^{-1}x_i}{(\hat{\sigma}_v^2 + \sigma_{ii})}\right\}, \qquad (8)$$

The partial derivatives appearing in (8) can be evaluated using standard methods. They are given by

$$\frac{\partial \hat{\sigma}_v^2}{\partial y_i} = \frac{2}{(n-p)}(y_i - x_i'\hat{\beta}),$$

and

$$\frac{\partial g_i(y)}{\partial \hat{\sigma}_v^2} = \frac{\sigma_{ii}}{(\hat{\sigma}_v^2 + \sigma_{ii})^2}(y_i - x_i'\hat{\beta}_w) + \frac{\sigma_{ii}}{(\hat{\sigma}_v^2 + \sigma_{ii})} x_i' \frac{\partial \hat{\beta}_w}{\partial \hat{\sigma}_v^2},$$

where

$$\frac{\partial \hat{\beta}_w}{\partial \hat{\sigma}_v^2} = -\hat{A}^{-1} \sum_1^n \frac{x_k(y_k - x_k'\hat{\beta}_w)}{(\hat{\sigma}_v^2 + \sigma_{kk})^2}.$$

From (8), one has a closed form expression for $\text{mse}_c(\hat{\mu}_i)$. This statistic is an estimator of mean squared error for the empirical Bayes estimator for the i-th small area with respect to the sampling design only. It is valid for any sample size n; it relies on the sole assumption that the direct estimators y_i are normally distributed. When $\hat{\sigma}_v^2 = 0$, $\hat{\mu}_i = x_i'\hat{\beta}_w$ and the derivatives in (8) simplify substantially. Since $\partial \hat{\sigma}_v^2 / \partial y_i = 0$, one has

$$\text{mse}_c(\hat{\mu}_i) = (y_i - x_i'\hat{\beta}_w)^2 - \sigma_{ii} + 2x_i'\hat{A}^{-1}x_i.$$

The properties of the conditional mean squared error estimator are best investigated in the simple situation where all the parameters of the smoothing model are assumed to be known. In this situation, $\partial g_i(y) / \partial y_i = -\sigma_{ii} / (\sigma_{ii} + \sigma_v^2)$ and the conditional mean squared error estimator is equal to $\text{mse}_c^+(\hat{\mu}_i) = \max\{(\text{mse}_c(\hat{\mu}_i), 0\}$ where

$$\text{mse}_c(\hat{\mu}_i) = \frac{\sigma_{ii}\sigma_v^2}{\sigma_{ii} + \sigma_v^2} + \left(\frac{\sigma_{ii}}{\sigma_{ii} + \sigma_v^2}\right)^2 \left\{(y_i - x_i'\beta)^2 - \sigma_{ii} - \sigma_v^2\right\}.$$

The model based alternative to this estimator is the posterior variance, $\sigma_{ii}\sigma_v^2 / (\sigma_{ii} + \sigma_v^2)$, which coincides with $E_M[E_S\{\text{mse}_c(\hat{\mu}_i)\}]$. This estimator is a special case of Prasad and Rao (1990) estimator and is denoted $\text{mse}_{\text{PR}}(\hat{\mu}_i)$. Estimator $\text{mse}_c^+(\hat{\mu}_i)$ is highly variable when σ_v^2 is small. Indeed, when σ_v^2 is close to 0, about 50% of the conditional mean squared error estimates are null. To further compare the 2 mean squared error estimators, conditional and unconditional, observe that when all the parameters of the smoothing model are known, the conditional mean squared error of $\hat{\mu}_i$ is

$$E_S\{\text{mse}_c(\hat{\mu}_i)\} = \frac{\sigma_{ii}\sigma_v^2}{\sigma_{ii} + \sigma_v^2} + \left(\frac{\sigma_{ii}}{\sigma_{ii} + \sigma_v^2}\right)^2 \left\{(\mu_i - x_i'\beta)^2 - \sigma_v^2\right\}.$$

The next proposition compares the average mean squared errors of the estimators, conditional or unconditional, of $E_S\{\text{mse}_c(\hat{\mu}_i)\}$.

PROPOSITION 4: When $\sigma_{ii} = \sigma^2$, for $i = 1, ..., n$ and when the small area means are μ_i's are drawn using (6), the efficiency of the posterior variance with respect to the conditional mean squared error estimator for estimating the conditional mean squared error is

$$\frac{E_M\left[\sum \text{MSE}_S\{\text{mse}_c(\hat{\mu}_i)\} / n\right]}{E_M\left[\sum \text{MSE}_S\{\text{mse}_{\text{PR}}(\hat{\mu}_i)\} / n\right]} = \frac{\sigma^4 + 2\sigma^2\sigma_v^2}{\sigma_v^4}$$

where $\text{MSE}_S(\cdot)$ denote a mean squared error taken with respect to the distribution of the y_i's which are independent $N(\mu_i, \sigma^2)$ random variables.

The above efficiency is larger than 1 provided that $\sigma_v^2/\sigma^2 < 2.41$. Proposition 4 shows under heavy shrinking, the unconditional mean squared error estimator is a better estimator of the conditional mean squared error than the conditional estimator. This surprising result is caused by the large variance of the conditional estimator; when shrinking is extensive, it is a poor estimator.

In some situations, such as that consider in section 5.1, shrinking is light and the use of the conditional mean squared error estimator is appropriate. The conditional efficiency of $\hat{\mu}_i$ with respect to the direct estimator y_i is given by $\sigma_{ii} / \text{mse}_c^+(\hat{\mu}_i)$. This is larger than one provided that $(y_i - x_i'\beta)^2 / (\sigma_{ii} + \sigma_v^2) < 2$. Assuming that the smoothing model holds true, conditional efficiencies less than 1 can be expected for approximatively 16% $(= P[N(0,1)^2 < 2])$ of the small area estimators. This percentage should be higher if the smoothing model is deficient. Conditional efficiencies less than 1 occur in small areas having large residuals. On the other hand, the unconditional efficiencies, calculated with the posterior variance are, in this situation, less than 1 for all small areas. This shows that it is practically impossible for all the conditional efficiencies to be less 1; this had already been noted by Rao and Shinozaki (1978) for James-Stein estimators.

Many of the observations made in the unrealistic situation where all the parameters are known also apply when parameters are estimated. The unconditional alternative to the conditional mean squared error estimator is Prasad and Rao's (1990) estimator,

$$\text{mse}_{\text{PR}}(\hat{\mu}_i) = \frac{\sigma_{ii}\hat{\sigma}_v^2}{\sigma_{ii} + \hat{\sigma}_v^2} + \frac{\sigma_{ii}^2 x_i'\hat{A}^{-1}x_i}{(\sigma_{ii} + \hat{\sigma}_v^2)^2} + 2\frac{\sigma_{ii}^2 \widehat{\text{Var}}(\hat{\sigma}_v^2)}{(\sigma_{ii} + \hat{\sigma}_v^2)^3}, \quad (9)$$

where $\widehat{\text{Var}}(\hat{\sigma}_v^2) = 2\sum (\hat{\sigma}_{ii} + \hat{\sigma}_v^2)^2 / n^2$. To investigate the extent to which Proposition 4 holds when parameters are estimated, a small Monte Carlo study was carried out along the lines of the approach ii) simulation study of Prasad and Rao (1999). In all the simulations, $n = 30$ and $\sigma_{ii} = 1$, for $i = 1, ..., n$. The smoothing model (6) was $\mu_i = \mu + v_i$ and various values of σ_v^2 were used. The results reported in Table 1 are based on $m = 5000$ Monte Carlo replications.

The simulations used 5 sets of μ_i-values whose variances are reported in Table 1. For each set, y_i was simulated repeatedly as a $N(\mu_i, 1)$ random variable, $i = 1, ..., n$. The empirical Bayes estimate $\hat{\mu}_i$ was calculated for each small area and the mean squared error for small area i was calculated as $\text{MSE}_i = \sum^* (\hat{\mu}_i - \mu_i)^2 / m$ where $\sum^*$ denotes the sum on the m Monte Carlo replications. The efficiency of the empirical Bayes estimator for small area i is $1 / \text{MSE}_i$. The mean and the median of the $n = 30$ small area efficiencies are given in Table 1. The 2 mean squared errors, conditional and unconditional, were calculated for

each small area in the m Monte Carlo replications; from (9), $\text{mse}_{PR}(\hat{\mu}_i) = (\hat{\sigma}_v^2 + 5/n)/(1 + \hat{\sigma}_v^2)$ for each small area. Table 1 presents the mean and the median of their absolute relative biases, defined as $\left|\sum^* (\text{mse}_i(\hat{\mu}_i) - \text{MSE}_i)\right| / (m\text{MSE}_i)$ and of their coefficients of variation which are equal to $(\sum^* (\text{mse}_i(\hat{\mu}_i) - \text{MSE}_i)^2/m)^{1/2} / \text{MSE}_i$.

Table 1
Relative Efficiency of the Empirical Bayes Estimators (RE), Absolute Relative Bias (RB) and Coefficient of Variation (CV) of two MSE Estimators ($n = 30$). All Results are Expressed in Percentage

$\sum(\mu_i - \bar{\mu})^2/29$		RE%	RB_c%	RB_{PR}%	CV_c%	CV_{PR}%
1.3	mean	212	1	47	97	51
	median	214	1	40	100	43
2.53	mean	149	2	30	37	31
	median	163	2	31	37	32
3.7	mean	129	2	20	23	20
	median	133	1	21	24	21
4.24	mean	125	2	19	19	20
	median	131	1	22	20	22
4.93	mean	122	1	17	15	18
	median	133	1	17	13	17

As shown in section 2, $\text{mse}_c(\hat{\mu}_i)$ is unbiased; the biases reported in Table 1 are caused by Monte Carlo errors. When $n = 30$, the condition $\sigma_v^2/\sigma^2 > 2.4$ derived in Proposition 4 for the conditional estimator to improve on the unconditional estimator is not sufficient; the stronger condition $\sigma_v^2/\sigma^2 > 4$ is needed. Noteworthy is the fact that in Table 1, for $\sum(\mu_i - \bar{\mu})^2/29 > 2.5$, the CV of $\text{mse}_{PR}(\hat{\mu}_i)$ is only bias. Table 1 confirms that, when $\hat{\sigma}_v^2$ is of the same order of magnitude as σ_{ii} or smaller, the squared residual dominates the distribution of the conditional mean squared error estimator; in such cases Prasad and Rao (1990) unconditional estimator is a better estimator of conditional mean squared error. Even in situations when $\text{mse}_c(\hat{\mu}_i)$ cannot be recommended as an estimator for the conditional mean squared error, it still provides interesting diagnostic information: changes in the conditional estimators give a basis for comparing two smoothing models. This is illustrated in section 5.2.

4.3 Conditional Mean Squared Error and Prediction Variance

This section explores the relationship between the conditional mean squared error proposed in this paper and the prediction variance which is an unconditional measure of accuracy. Using the rotation of (6), the prediction variance is $\text{MSE}(\hat{\mu}_i) = E_M[E_S\{(\hat{\mu}_i - x_i'\beta - v_i)^2\}]$. From the construction of presented in section 2, one has

$$E_S\{\text{mse}_c(\hat{\mu}_i)\} = E_S\{(\hat{\mu}_i - x_i'\beta - v_i)^2\}.$$

Thus we have the following result,

PROPOSITION 5: The conditional mean squared error of empirical Bayes small area estimators satisfies,

$$E_M\left[E_S\{\text{mse}_c(\hat{\mu}_i)\}\right] = \text{MSE}(\hat{\mu}_i),$$

where $\text{MSE}(\hat{\mu}_i)$ is the unconditional prediction variance.

Proposition 5 shows that $\text{mse}_c(\hat{\mu}_i)$ can be looked at as an intermediate step in the evaluation of the unconditional mean squared error of $\hat{\mu}_i$. Consider for instance the calculation of Prasad and Rao (1990) $o(1/n)$ approximation to $\text{MSE}(\hat{\mu}_i)$,

$$\text{MSE}_{PR}(\hat{\mu}_i) = \frac{\sigma_{ii}\sigma_v^2}{\sigma_{ii} + \sigma_v^2} + \frac{\sigma_{ii}^2 x_i' A^{-1} x_i}{(\sigma_{ii} + \sigma_v^2)^2} + \frac{\sigma_{ii}^2 \text{Var}(\hat{\sigma}_v^2)}{(\sigma_{ii} + \sigma_v^2)^3},$$

where $\text{Var}(\hat{\sigma}_v^2) = 2\sum(\sigma_{ii} + \sigma_v^2)^2/n^2$. The standard derivation, as reviewed in section 3.2 of Singh, Stukel, and Pfeffermann (1998), is based on Kackar and Harville (1984). An alternative derivation, presented in Belmonte (1998, 1999), is to take the expectation of $\text{mse}_c(\hat{\mu}_i)$, obtained using (8), with respect to the marginal distribution of the y_i's, which are independent $N(x_i'\beta, \sigma_{ii} + \sigma_v^2)$ deviates and to retain only the higher order terms.

Proposition 5 holds in situations where the small area estimators are bench-marked, or where corrections suggested in section 3 are implemented. These are cases for which there are no closed form formulas for the prediction variances. Proposition 4 suggests a simple method for constructing unconditional Monte Carlo estimates. It suffices to generate a large number of replicates of $\{y_i, i = 1, ..., n\}$ where y_i follows a $N(x_i'\hat{\beta}_w, \hat{\sigma}_v^2 + \sigma_{ii})$ and to calculate $\text{mse}_c(\hat{\mu}_i)$ for each one. Averaging the $\text{mse}_c(\hat{\mu}_i)$'s gives a plug-in unconditional prediction variance, equal to the MSE of Proposition 4 evaluated at estimates $\hat{\beta}_w$, $\hat{\sigma}_v^2$ of the unknown parameters. Unfortunately, this estimate is biased (this is a first order estimate in the terminology of Singh, Stukel and Pfeffermann (1998)). For the empirical Bayes estimator given by (7), according to (9) the bias of the Monte Carlo estimate derived form Proposition 4 is $-\sigma_{ii}^2\text{Var}(\hat{\sigma}_v^2)/(\sigma_{ii} + \hat{\sigma}_v^2)^3$. Further work is needed for constructing, using Proposition 4, unbiased unconditional prediction variance estimators.

5. ESTIMATING THE UNDER-COVERAGE IN THE 1991 CANADIAN CENSUS

In 1991, the under-coverage of the Canadian Census was estimated using two surveys, the Over-coverage Study, which estimates the number of persons double counted or erroneously counted in the Census and the Reverse Record Check (Burgess 1988) for the persons missed in the Census. Combining these figures gives estimates of the under-coverage of the Census. This section investigates several estimators of census under-coverage.

5.1 Provincial Estimations

The 1991 under-coverage rates for the ten Canadian provinces and the two territories with their coefficients of variation, expressed in percentage, are given in Table 2. The proportion p_i of the population living in each province (the word province is used in this section to denote the 10 Canadian provinces and the two territories) is also provided. The coefficients of variation (CV) of Table 2 were calculated from variances estimated with 5 random groups. Thus, one can consider that the sampling variances have a χ^2_4 distribution. Throughout this section, we assume that the provincial under-coverage estimates and their variances are independent.

Several estimators for provincial under-coverage are proposed by Royce (1992). Rivest (1995) proposed a composite estimator that shrinks the provincial under-coverage rate towards the national rate. It is given by:

$$r_i^c = \hat{\alpha} r_i + (1 - \hat{\alpha}) r_N,$$

where $r_N = \sum p_i r_i$ is the national under-coverage rate and the shrinking parameter $\hat{\alpha}$ is given by:

$$\hat{\alpha} = \frac{\sum p_i r_i^2 - r_N^2}{\sum p_i(1 - p_i)\sigma_i^2 + \sum p_i r_i^2 - r_N^2}.$$

This is the value of α that is optimal for loss functions for the estimation of provincial totals and of provincial shares of the population; see Royce (1992) and Rivest (1995) for details. One has $r_i^c = r_i + g_i(r)$, where

$$g_i(r) = -\frac{\sum p_i(1 - p_i)\sigma_i^2}{\sum p_i(1 - p_i)\sigma_i^2 + \sum p_i r_i^2 - r_N^2}(r_i - r_N).$$

A closed form expression for the conditional mean square error estimator can be calculated easily by noting that

$$\frac{\partial g_i(r)}{\partial r_i} = 2p_i(r_i - r_N)^2 \frac{\sum p_i(1 - p_i)\sigma_i^2}{\left[\sum p_i(1 - p_i)\sigma_i^2 + \sum p_i r_i^2 - r_N^2\right]^2} - (1 - p_i)(1 - \hat{\alpha}).$$

The second partial derivative of $g_i(r)$ can also be calculated; it has the same sign as $r_i - r_N$. Thus positive skewness in the under-coverage rate, that is likely when estimating rare events such as being missed by the census, increases the conditional mean squared error in provinces where the under-coverage is above the national rate.

For 1991, $\hat{\alpha} = .874$ and the national under-coverage rate is $r_N = 2.872\%$. Table 2 gives the provincial composite under-coverage estimates, r_i^c together with their efficiencies $\text{eff}_{ic}^c = \sigma_{ii} / \text{mse}_c(r_i^c)$, where $\text{mse}_c(r_i^c)$ is calculated as defined in section 2, with the correction proposed in section 3.2 to account for estimated variances. The composite estimator is an improvement over the direct estimators in all cases except three, that correspond to the provinces with the most extreme under-coverage rates.

Table 2 also gives the empirical Bayes estimator r_i^B calculated with a location smoothing model. Under model (M), the true under-coverage rate θ_i is assumed to be distributed as a $N(\beta, \sigma_v^2)$. The parameter estimates are $\hat{\sigma}_v^2 = 1.45 \times 10^{-4}$ and $\hat{\beta}_w = 2.61\%$. Two efficiencies with respect to direct estimators are presented, eff_{ic}^B which is calculated with the conditional mean squared error estimator for r_i^B, including the adjustment of section 3.2 to account for estimated variances, and eff_{iPR}^B which is calculated with Prasad-Rao unconditional estimator. The large under-coverage rate in the N.W. Territories is responsible for the large estimate for $\hat{\sigma}_v^2$; this makes the empirical Bayes estimators r_i^B much closer to the direct estimators r_i than the composite estimators. Redoing the analysis without the N.W. Territories and Yukon changes the empirical Bayes estimates drastically.

Table 2
Two Estimators of Provincial Under-Coverage and Their Efficiencies

PROVINCE	p_i	r_i	CV	r_i^c	eff_{ic}^c	r_i^B	eff_{ic}^B	eff_{iPR}^B
Newfoundland	2.06	1.994	15.96	2.105	1.12	2.038	1.07	1.04
Prince Edward Island	0.47	0.931	30.00	1.176	0.65	1.025	0.93	1.03
Nova Scotia	3.26	1.889	20.05	2.013	1.11	1.959	1.09	1.06
New Brunswick	2.66	3.245	13.73	3.198	1.29	3.162	1.14	1.09
Québec	25.19	2.605	8.35	2.639	1.16	2.605	1.04	1.02
Ontario	37.24	3.641	8.46	3.544	0.87	3.572	1.02	1.04
Manitoba	3.96	1.86	20.83	1.987	1.10	1.936	1.09	1.06
Saskatchewan	3.58	1.798	18.87	1.933	1.04	1.863	1.06	1.05
Alberta	9.24	1.995	14.57	2.106	1.01	2.032	1.06	1.03
British Columbia	12.01	2.733	9.86	2.751	1.26	2.727	1.07	1.03
Yukon	0.10	3.83	15.99	3.709	1.27	3.56	1.05	1.17
N.W. Territories	0.22	5.439	11.28	5.116	0.96	4.813	0.49	1.18

In Table 2, the composite estimator performs better than the empirical Bayes estimator; it provides gains in conditional efficiency larger than 10% in 7 of 12 provinces. Three efficiencies are smaller than 1; the discussion in section 4.2 suggests that efficiencies less than 1 are unavoidable. The relatively poor precision of $\hat{\sigma}_{ii}$ (they are estimated using only 4 degrees of freedom), lowers the conditional efficiencies of the empirical Bayes estimators. It does not affect the composite estimator as much since it uses the same shrinking parameter for all provinces. The conditional efficiencies capture the poor performances of the r_i^C and r_i^B in the provinces with the most extreme under-coverage rates. This is missed by the Prasad Rao efficiencies. They highlight the gains that smoothing brings to the two territories where the under-coverage rates are highly variable. The Prasad Rao efficiencies are meaningful only if one accepts the hypothesis of provincial exchangeability underlying the smoothing model. This is doubtful since under-coverage tends to be higher in large urban provinces than in small rural areas.

5.2 Sub-Provincial Estimations

Dick (1995) considered the estimation of the adjustment factors for census under-coverage for age × sex categories within each province for the 1991 census. The adjustment factor for a small area is defined as F=1+ (estimated under-coverage)/(census count). With four age categories, 0-19, 20-29, 30-44, 45+, and two sexes, there are 96 small areas. The explanatory variables are interactions between the indicator variables for the 12 provinces, the 4 age groups and the two sexes, and the proportions of renters (R) and of people that do not speak either official language (L) in the 96 small areas. In each one, the estimated variance was given by $\tilde{\sigma}_{ii}$ = (under-coverage variance) / (census count)^2.

Dick (1995) regressed the log-variances on the census count to smooth the variance. He considered the exponentials of the predicted values for the log-variances ($\tilde{\sigma}_{ii}$) as the known variances. This underestimates the variability.

Indeed, the average predicted variance $\tilde{\sigma}_{ii}$ represents only 68% of the average unsmoothed variance. Multiplying $\tilde{\sigma}_{ii}$ by $\exp(\hat{\sigma}_r^2/2) = 1.54$, where $\hat{\sigma}_r^2$ is the residual variance of the smoothing model, corrects this problem. Fitting Dick's (1995) model using the "unbiased" smoothed variance yields $\hat{\sigma}_v^2 = 0$. This is a degenerate situation where empirical Bayes estimators are equal to linear model predicted values. Note also the correlation between the variance residuals and the partial derivatives of g_i, calculated as if $\hat{\sigma}_v^2 > 0$, is 0.25. This suggest that (5) is violated. Using $\tilde{\sigma}_{ii}\exp(\hat{\sigma}_r^2/2)$ in the calculation is likely to overestimate the precision the small area estimates. To illustrate the application of the conditional mean squared error estimator, these problems are ignored and the remainder of this section assumes that the sampling variances σ_{ii} are known and equal to their smoothed values $\tilde{\sigma}_{ii}$.

The model fitted by Dick (1995) has ten independent variables; the weighted least squares estimates and their standard errors, given by the square roots of the elements on the diagonal matrix of $\hat{A}^{-1}$, appear in Table 3. The conditional mean squared errors $\text{mse}_c^*(\hat{\mu}_i)$ for the 96 small areas can be calculated using (8). One had $\text{mse}_c^*(\hat{\mu}_i) = 0$ and $\text{mse}_c^*(\hat{\mu}_i) > \sigma_{ii}$ for respectively 51 and 15 small areas. The 15 small areas with large conditional mean squared errors need special attention: can the prediction model be improved for these areas? Two systematic features among the 15 corresponding residuals are noteworthy: there are 2 large positive residuals in the M/0-19 category and 2 large negative residuals in the F/45+ category. This suggests adding M/0-19 and F/45+ as independent variables. The additional column to the X matrix for M/0-19 contains 1's for the 12 small areas for males between 0 and 19 years old and 0 elsewhere; that for F/45+ is constructed in a similar way. Only F/45+ improves the fit; adding this explanatory variable gives the modified Dick model of Table 3. The absolute value of the t-statistic for F/45+ is 3; this is clearly significant.

It is interesting to compare the conditional mean squared errors obtained with the modified Dick model with those for Dick's model. Using the modified model decreases mse_c^* in 26 small areas and increases it in 21; showing a slight improvement with the modified model.

The sub-provincial empirical Bayes adjustment factors can be aggregated at the provincial level. Provincial adjustment factors F_p are given by

$$\hat{F}_p = \frac{\sum_p C_i \hat{F}_i}{\sum_p C_i}$$

where C_i represents the census count for the i-th small area and $\sum_p$ is the summation over the 8 small areas in province p. A mean squared error for the provincial adjustment factor, either conditional or unconditional, can be calculated using a mean product error matrix mpe as

$$\text{mse}(\hat{F}_p) = \frac{1}{\left(\sum_p C_i\right)^2}\sum_p \sum_p C_i C_j \text{mpe}(\hat{F}_i, \hat{F}_j).$$

Conditional mean squared errors are obtained by using formula (2) for mpe. Lahiri and Rao (1995) give a formula for the off-diagonal terms of the unconditional mean product error matrix whose diagonal is given by Prasad Rao (1990) mean squared errors.

Table 3
Two Linear Models for Small Area Correction Factors: Dick (p=11) and Modified Dick (p=12). Parameter Estimates are Given With Their Standard Errors in Parentheses

Category	Variable	Dick		modified Dick	
mean	intercept	1.0076	(0.0018)	1.0099	(0.0018)
Age* Sex Interaction	M / 20-29	0.0563	(0.0038)	0.0541	(0.0037)
	M / 30-44	0.0207	(0.0036)	0.0185	(0.0035)
	F / 20-20	0.0243	(0.0038)	0.02223	(0.0037)
	F / 45+		-	-0.0102	(0.0037)
Province* Renters Interaction	BC*R	0.0436	(0.0115)	0.0433	(0.0110)
	Ontario*R	0.0791	(0.0100)	0.0789	(0.0102)
	Québec*R	0.0253	(0.0097)	0.0259	(0.0090)
	N.-B*R	0.1039	(0.0194)	0.1032	(0.0186)
	Yukon*R	0.0633	(0.0179)	0.0634	(0.0175)
	NWT*R	0.0687	(0.0117)	0.0680	(0.0285)
Language*Sex*Age Interaction	L*F / 0-19	0.0802	(0.0293)	0.0680	(0.0285)
Variance		3.3681e-05	(2.45e-05)	2.21e-05	(2.30e-05)

Table 4
Direct (F_p) and Empirical Bayes (F_p^b) Estimates of the Provincial Correction Factors With Their Conditional (eff_{pc}) and Their Unconditional (eff_{pPR}) Efficiencies. A Conditional Efficiency is ∞ When the Conditional Mean Squared Error Estimator is Null

PROVINCE	F_p	F_p^b	eff_{pc}	eff_{pPR}
Newfoundland	1.0203	1.0176	6.49	2.94
Prince Edward Island	1.0094	1.0153	1.03	4.52
Nova Scotia	1.0193	1.0171	25.3	2.59
New Brunswick	1.0335	1.0367	0.67	1.11
Québec	1.0268	1.0262	1.12	0.93
Ontario	1.0378	1.0396	0.68	0.93
Manitoba	1.0190	1.0176	∞	2.46
Saskatchewan	1.0183	1.0166	∞	2.54
Alberta	1.0204	1.0187	7.37	1.98
British Columbia	1.0281	1.0293	1.09	1.03
Yukon	1.0396	1.0400	1.41	1.17
N.W. Territory	1.0575	1.0550	1.40	1.32

Direct and empirical Bayes aggregated estimates are presented in Table 4 with two efficiencies. The empirical Bayes estimates retain much of the interprovincial differences. This suggest that the explanatory variables of the smoothing model have captured most of the differences between the provincial under-coverage rates. A notable exception is Prince Edward Island's small correction factor which is not accounted for by the explanatory variables. This is the only province for which the two efficiencies differ substantially. The conditional efficiencies are more unstable than the Prasad Rao efficiencies. Except in Prince Edward Island, both tell similar stories: in New Brunswick, Quebec, Ontario, and British Columbia, the aggregated empirical Bayes estimates do not improve much on the direct estimators.

6. CONCLUSIONS

The estimator of the conditional mean squared error proposed in this paper has several interesting features. It can be implemented with any shrinking strategy. It is conditional on the realization of the smoothing model used to produce the small area characteristics; thus the conditional estimator has a large sampling variance. Simple modifications to the estimator are available to handle skewness in the data and estimated variances. In an empirical Bayes setting, it provides diagnostic information concerning the smoothing model. It can also be used as building blocks for estimators of the prediction variances when this variance has no closed form expression.

ACKNOWLEDGEMENTS

We are grateful to Peter Dick for providing the data set analyzed in section 5.2, and to Jon Rao for pointing out the instability of the conditional estimator under heavy smoothing. The financial contributions of the Fonds pour la formation des chercheurs et l'aide à la recherche du Québec and of the National Science and Engineering Research Council of Canada are gratefully acknowledged.

APPENDIX

Proof of Proposition 1

Let $\Sigma^{1/2}$ be a symmetric square root for Σ, such that $(\Sigma^{1/2})^2=\Sigma$ and $z=\Sigma^{-1/2}(y-\mu)$. Note that z has a $N_n(0,I)$ distribution. In terms of the random vector $z, E\{(y-\mu)\, g(y)^t\} = \Sigma^{1/2} E\{z g(\mu+\Sigma^{1/2}z)\}$. Now the conditional expectation of $z_i g_j(\mu+\Sigma^{1/2}z)$ given $(z_1, ..., z_{i-1}, z_{i+1}, ..., z_n)$ is equal to

$$\int_R \frac{z_i \exp(-z_i^2/2)}{\sqrt{2\pi}} g_j(\mu + \Sigma^{1/2} z) dz_i.$$

Integrating by parts shows that the above integral is equal to

$$\int_R \frac{\exp(-z_i^2/2)}{\sqrt{2\pi}} \frac{\partial g_j(\mu + \Sigma^{1/2} z)}{\partial z_i} dz_i.$$

Observe that

$$\frac{\partial g_j(\mu + \Sigma^{1/2} z)}{\partial z_i} = \sum_{k=1}^{n} \Sigma_{ki}^{1/2} g_j^{\,k}(\mu + \Sigma^{1/2} z).$$

Since $\Sigma^{1/2}$ is symmetric, $\Sigma_{ki}^{1/2} = \Sigma_{ik}^{1/2}$. Thus the above expression is the scalar product between $e_i' \Sigma^{1/2}$, the i-th row of $\Sigma^{1/2}$ (e_i represents a $n \times 1$ vector of 0's except for the i-th component which is 1), and $\nabla g(\mu + \Sigma^{1/2} z) e_j$, the j-th column of $\nabla g(y)$, evaluated at $y = \mu + \Sigma^{1/2} z$. We have

$$E\left\{z_i g_j(\mu + \Sigma^{1/2} z) \mid z_1, \ldots, z_{i-1}, z_{i+1}, \ldots, z_n\right\} =$$

$$e_i' \Sigma^{1/2} E\left\{\nabla g(\mu + \Sigma^{1/2} z) e_j \mid z_1, \ldots, z_{i-1}, z_{i+1}, \ldots, z_n\right\}.$$

This equality also holds unconditionally, $E\{z_i g_j(\mu + \Sigma^{1/2} z)\} = e_i' \Sigma^{1/2} E\{\nabla g(\mu + \Sigma^{1/2} z)\} e_j$. In other words,

$$E\{z g(\mu + \Sigma^{1/2} z)\} = \Sigma^{1/2} E\{\nabla g(\mu + \Sigma^{1/2} z)\}.$$

This completes the proof.

Proof of Proposition 2

Let E_i denote the conditional expectation with respect to y_i, given $(y_1, \ldots, y_{i-1}, y_{i+1}, \ldots, y_n)$ and $h(y_i) = g_i(y)$, for fixed values of $(y_1, \ldots, y_{i-1}, y_{i+1}, \ldots, y_n)$. One has

$$E_i\{(y_i - \mu_i) h(y_i)\} = \int_R (t - \mu_i) h(t) f(t) dt.$$

To evaluate this expression, one can integrate by parts. Integrating $(t - \mu_i) \exp\{-(t - \mu_i)^2/(2\sigma_{ii})\}/(2\pi\sigma_{ii})^{1/2}$ in the above integrand yields

$$E_i\{(y_i - \mu_i) h(y_i)\} = \sigma_{ii} E_i\{h'(y_i)\} + \frac{\sigma_{ii}^{1/2} \rho_i}{2}$$

$$\times \int_R h(t) \left\{\frac{(t - \mu_i)^2}{\sigma_{ii}} - 1\right\} \frac{\exp\{(t - \mu_i)^2/(2\sigma_{ii})\}}{(2\pi\sigma_{ii})^{1/2}} dt,$$

where $h'(t)$ is the derivative of $h(t)$. Repeated integrations by parts show that

$$\int_R h(t) \frac{(t - \mu_i)^2}{\sigma_{ii}} \frac{\exp\{(t - \mu_i)^2/(2\sigma_{ii})\}}{(2\pi\sigma_{ii})^{1/2}} dt$$

$$= \int_R \{h'(t)(t - \mu_i) + h(t)\} \frac{\exp\{(t - \mu_i)^2/(2\sigma_{ii})\}}{(2\pi\sigma_{ii})^{1/2}} dt$$

$$= \int_R \{\sigma_{ii} h''(t) + h(t)\} \frac{\exp\{(t - \mu_i)^2/(2\sigma_{ii})\}}{(2\pi\sigma_{ii})^{1/2}} dt$$

where $h''(t)$ is the second derivative of $h(t)$. This yields

$$E_i\{(y_i - \mu_i) h(y_i)\} =$$

$$\sigma_{ii} E_i\{h'(y_i)\} + \frac{\sigma_{ii}^{3/2} \rho_i}{2} E_i\{h''(y_i)\} + o(\rho_i).$$

Taking, on both sides, the expectation with respect to the distribution of $(y_1, \ldots, y_{i-1}, y_{i+1}, \ldots, y_n)$ completes the proof.

Proof of Proposition 3

Let E_i denote the expectation taken with respect to the distribution of $\hat{\sigma}_{ii}$, given all the other random quantities $(y, \hat{\sigma}_{jj}, j \neq i)$. In this context one can write $(\partial g_i(y))/(\partial y_i) = h(\hat{\sigma}_{ii})$, where h is a function possibly depending on $(y, \hat{\sigma}_{jj}, j \neq i)$. A Taylor series expansion of h gives:

$$h(\hat{\sigma}_{ii}) = h(\sigma_{ii}) + h'(\sigma_{ii})(\hat{\sigma}_{ii} - \sigma_{ii}) + h''(\sigma_{ii}) \frac{(\hat{\sigma}_{ii} - \sigma_{ii})^2}{2} + O((\hat{\sigma}_{ii} - \sigma_{ii})^3).$$

Since $(k - 1)\hat{\sigma}_{ii}/\sigma_{ii}$ follows a χ_{k-1}^2 distribution, $E_i\{(\hat{\sigma}_{ii} - \sigma_{ii})^2\} = 2\sigma_{ii}^2/(k - 1)$, and the centered moments of higher orders are $O(1/k^2)$. The above expansion reduces to,

$$\sigma_{ii} E_i\{\partial g_i(y)/\partial y_i\} = \sigma_{ii} h(\sigma_{ii}) + h''(\sigma_{ii}) \frac{\sigma_{ii}^3}{k - 1} + O(1/k^2)$$

It is clear that the bias of $\hat{\sigma}_{ii} h(\hat{\sigma}_{ii})$ as an estimator of this expression is $O(1/k)$, provided that $h'(\sigma_{ii}) \neq 0$. One has, neglecting $O(1/k^2)$ terms,

$$E_i\left\{\hat{\sigma}_{ii} h\left(\frac{(k - 1)\hat{\sigma}_{ii}}{k + 1}\right)\right\}$$

$$\approx \sigma_{ii} h(\sigma_{ii}) + h'(\sigma_{ii}) E_i\left\{\hat{\sigma}_{ii}\left(\frac{(k - 1)\hat{\sigma}_{ii}}{k + 1} - \sigma_{ii}\right)\right\}$$

$$+ \frac{h''(\sigma_{ii})}{2} E_i\left\{\hat{\sigma}_{ii}\left(\frac{(k - 1)\hat{\sigma}_{ii}}{k + 1} - \sigma_{ii}\right)^2\right\}$$

Elementary manipulations show that, in the above formula, the coefficient of $h'(\sigma_{ii})$ is null and

$$E_i\left\{\hat{\sigma}_{ii}\left(\frac{(k-1)\hat{\sigma}_{ii}}{k+1}-\sigma_{ii}\right)^2\right\}=2\frac{\sigma_{ii}^3}{k-1}+O(1/k^2).$$

This shows that

$$E_i\left\{\hat{\sigma}_{ii}h\left(\frac{(k-1)\hat{\sigma}_{ii}}{k+1}\right)\right\}=\sigma_{ii}E_i\{\partial g_i(y)/\partial y_i\}+O(1/k^2).$$

The proof is completed by noting that this equality holds for the unconditional expectation, taken with respect to the joint distribution of $(y, \hat{\sigma}_{ii}, i = 1, ..., n)$.

Proof of Proposition 4

The mean squared error of the posterior variance as an estimator of the conditional mean squared error has only a bias term,

$$\left(\frac{\sigma^2}{\sigma^2+\sigma_v^2}\right)^4\{(\mu_i-x_i'\beta)^2-\sigma_v^2\}^2,$$

while the mean squared error of $\text{mse}_c(\hat{\mu}_i)$ has only a variance component which is given by

$$\left(\frac{\sigma_2}{\sigma^2+\sigma_v^2}\right)^4\text{Var}_S\{(y_i-x_i'\beta)^2\}$$
$$=\left(\frac{\sigma^2}{\sigma^2+\sigma_v^2}\right)^4\{2\sigma_{ii}^2+4(\mu_i-x_i'\beta)^2\sigma^2\}.$$

The efficiency reported in Proposition 4 can be evaluated as the ratio of the 2 average mean squared errors defined above. It is given by,

$$\frac{2\sigma^4+4\sigma^2\sum(\mu_i-x_i'\beta)^2/n}{\sum\{(\mu_i-x_i'\beta)^2-\sigma_v^2\}^2/n}.$$

Taking expectations of the numerator and of the denominator with respect to model (6) yields the result.

REFERENCES

BARNDORFF-NIELSEN, O.E., and COX, D.R. (1989). *Asymptotic Techniques for Use in Statistics*. New York: Chapman and Hall.

BELMONTE, E. (1998). Estimation dans les petites régions: une nouvelle dérivation de l'erreur quadratique moyenne de Prasad-Rao. *1998 Proceedings of the Survey Methods Section, Statistical Society of Canada*, 165-170.

BELMONTE, E. (1999). *L'estimation dans les petites régions: Survol des méthodes de Bayes et présentation d'un estimateur conditionnel de l'EQM*. Mémoire de maîtrise. Département de mathématiques et de statistique, Université Laval.

BILODEAU, M., and SRIVASTAVA, M.S. (1988). Estimation of the MSE matrix of the Stein estimator. *Canadian Journal of Statistics*, 16, 153-159.

BOOTH, J.G., and HOBERT, J.P. (1998). Standard errors of predictions in generalized linear mixed models. *Journal of the American Statistical Association*, 93, 262-272.

BURGESS, R.D. (1988). Evaluation of the reverse record check estimates of under-coverage in the Canadian Census of Population. *Survey Methodology*, 14, 137-156.

CRESSIE, N. (1992). REML estimation in empirical Bayes smoothing of census undercount. *Survey Methodology*, 18, 75-94.

DICK, P. (1995). Modeling net under-coverage in the 1991 Canadian Census. *Survey Methodology*, 21, 44-55.

FAY, R.E., and HERRIOT, R.A. (1979). Estimates of income for small places: An application of James Stein procedure to census data. *Journal of the American Statistical Association*, 74, 269-277.

GHOSH, M., and RAO, J.N.K. (1994). Small area estimation: An appraisal. *Statistical Science*, 9, 55-93.

HOGAN, H. (1992). The 1990 Post-Enumeration Survey: an overview. *The American Statistician*, 46, 261-269.

KACKAR, R.N., and HARVILLE, D.A. (1984). Approximations for standard errors of estimators for fixed and random effects in mixed models. *Journal of the American Statistical Association*, 79, 853-862.

KOTT, P.S. (1989). Robust small domain estimation using random effect modeling. *Survey Methodology*, 15, 3-12.

LAHIRI, P., and RAO, J.N.K. (1995). Robust estimation of mean squared error of small area estimators. *Journal of the American Statistical Association*, 90, 758-766.

MARITZ, J.S., and LWIN, T. (1989). *Empirical Bayes Methods*. (Second Edition), London: Chapman and Hall.

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 small area estimation using a simple random effect model. *Survey Methodology*, 25, 163-171.

PURCELL, N.J., and KISH, L. (1979). Estimation for small domains. *Biometrics*, 35, 365-384.

RAO, C.R., and SHINOZAKI, N. (1978). Precision of individuals estimators in simultaneous estimation of parameters. *Biometrika*, 65, 23-30.

REID, N. (1991). Approximations and asymptotics. In *Statistical Theory and Modeling. In Honor of Sir David Cox, FRS*, (eds. D.V. Hinkley, N. Reid and E.J. Snell), 287-305.

RIVEST, L.P. (1995). A composite estimator for provincial under-coverage in the Canadian census. 1995 *Proceedings of the Survey Methods Section. Statistical Society of Canada*, 33-38.

ROBERT, C. (1992). *L'Analyse Statistique Bayesienne*. Paris: Economica.

ROYCE, D. (1992). A comparison of some estimators for a set of population totals. *Survey Methodology*, 18, 109-125.

SÄRNDAL, C.-E., SWENSSON, B., and WRETMAN, J. (1992). *Model Assisted Survey Sampling*. New York: Springer-Verlag.

SCOTT, A., and SMITH T.M.F. (1971). Interval estimates for linear combinations of means. *Applied Statistics*, 20, 276-285.

SINGH, M.P., GAMBINO, J., and MANTEL, H.J. (1994). Issues and strategy for small area data. *Survey Methodology*, 20, 1-22.

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.

STEIN, C. (1981). Estimation of the mean of a multivariate normal distribution. *The Annals of Statistics*, 9, 1135-1151.

WOLTER, K.M. (1985). *Introduction to Variance Estimation*. New York: Springer-Verlag.

Survey Methodology, June 2000
Vol. 26, No. 1, pp. 79-85
Statistics Canada

Cold Deck and Ratio Imputation

JUN SHAO[1]

ABSTRACT

Imputation is a common procedure to compensate for nonresponse in survey problems. Using auxiliary data, imputation may produce estimators that are more efficient than the one constructed by ignoring nonrespondents and re-weighting. We study and compare the mean squared errors of survey estimators based on data imputed using three different imputation techniques: the commonly used ratio imputation method and two cold deck imputation methods that are frequently adopted in economic area surveys conducted by the U.S. Census Bureau and the U.S. Bureau of Labor Statistics. A cold deck method imputes a nonrespondent of an item by reported values from anything other than reported values for the same item in the current data set (*e.g.*, values from a covariate and/or from a previous survey). Although sometimes a cold deck imputation method makes use of more auxiliary data than the other imputation methods, it is not always better in terms of the mean squared errors of the resulting survey estimators. In a simple case we compare explicitly the mean squared errors and discuss situations under which one method is better than the other two. In general cases we propose to compare mean squared errors empirically based on some consistent estimates of mean squared errors. Estimation of mean squared errors of survey estimators in the presence of imputed data is itself an important problem in surveys. A numerical example related to the Transportation Annual Survey is presented for illustration.

KEY WORDS: Complex survey; Mean squared error; Nonresponse; Simple random sample; Variance estimation.

1. INTRODUCTION

Imputation is one of the most common procedures to compensate for nonresponse in survey problems. In addition to many practical reasons for imputation, imputation using auxiliary data may produce estimators that are more efficient than the one constructed by ignoring nonrespondents and re-weighting. Suppose that we have a sample $\mathbf{s}$ selected from a finite population $\mathcal{P}$ consisting of some units represented by $i = 1, ..., M$, and that we observe $\{y_i, i \in \mathbf{r}\}$ (respondents), $\mathbf{r} \subset \mathbf{s}$. Suppose also that we have auxiliary data x_i's observed for all $i \in \mathbf{s}$ and $x_i > 0$. The commonly used ratio imputation method (see, for example, Kalton and Kasprzyk 1986) imputes nonrespondents as follows. First, we create K imputation cells $\mathcal{P}_k, \mathcal{P}_1 \cup \mathcal{P}_2 \cup \cdots \cup \mathcal{P}_K = \mathcal{P}$, according to a categorical auxiliary variable (which is observed for every $i \in \mathbf{s}$ and is typically different from x) such that for every k, the following model is assumed to hold:

$$y_i = \beta_k x_i + x_i^{1/2} e_i, \quad i \in \mathcal{P}_k, \qquad (1)$$
$$P(a_i = 1 \mid y_i, x_i) = P(a_i = 1 \mid x_i),$$

where β_k is an unknown parameter, e_i is independent of x_i with $E(e_i) = 0$ and unknown $V(e_i) = \sigma_k^2 > 0$, a_i is the indicator of whether y_i is a respondent, and (a_i, x_i)'s are independent. Then, within imputation cell k, a nonrespondent y_i is imputed by $\hat{\beta}_k x_i$, where

$$\hat{\beta}_k = \sum_{i \in \mathbf{r}_k} w_i y_i \Big/ \sum_{i \in \mathbf{r}_k} w_i x_i \qquad (2)$$

is the best linear unbiased estimator of β_k under model (1), $\mathbf{r}_k$ is $\mathbf{r}$ restricted to the k-th imputation cell, and w_i is the survey weight associated with the i-th sampled unit. Note that model (1) consists of a regression model between y_i and x_i (with no intercept and with error variance proportional to x_i) and a response model which assumes that the response mechanism is independent of y_i's, given x_i's. This response mechanism is termed as missing at random by Rubin (1976) or unconfounded response mechanism by Lee, Rancourt and Särndal (1994). Based on the imputed data set, the Horvitz-Thompson (HT) estimator of Y, the population total of y_i's, is

$$\hat{Y}_R = \sum_k \left(\sum_{i \in \mathbf{r}_k} w_i y_i + \sum_{i \in \mathbf{s}_k - \mathbf{r}_k} w_i \hat{\beta}_k x_i \right), \qquad (3)$$

where $\mathbf{s}_k$ is $\mathbf{s}$ restricted to the k-th imputation cell. The HT estimator of Y obtained by ignoring nonrespondents and re-weighting within each imputation cell is

$$\hat{Y}_W = \sum_k \sum_{i \in \mathbf{r}_k} \tilde{w}_{ik} y_i, \quad \tilde{w}_{ik} = w_i \left(\sum_{i \in \mathbf{s}_k} w_i \Big/ \sum_{i \in \mathbf{r}_k} w_i \right). \qquad (4)$$

It can be seen that if $x_i \equiv 1$, then the estimators in (3) and (4) are the same. Both estimators are unbiased if model (1) holds. (Throughout this paper, the bias and variance are with respect to model (1) and repeated sampling, unless otherwise specified.) Under model (1), however, $\hat{Y}_R$ is more efficient than $\hat{Y}_W$ if the size of $\mathbf{r}$ is substantially smaller than the size of $\mathbf{s}$. Even if the regression model in (1) does not hold, $\hat{Y}_R$ may still be more efficient than $\hat{Y}_W$ in terms of their mean squared errors with respect to repeated sampling (Cochran 1977, Chapter 6) when the response

[1] Jun Shao, Department of Statistics, University of Wisconsin, Madison, WI 53706 U.S.A. E-mail: shao@stat.wisc.edu.

probability is a constant in any given imputation cell (which ensures that $\hat{Y}_R$ and $\hat{Y}_W$ are approximately unbiased with respect to repeated sampling).

The purpose of this note is to compare the efficiency of $\hat{Y}_R$ with other estimators of Y based on data with nonrespondents imputed by using a method called cold deck. A cold deck method imputes a nonrespondent of y-variable by reported values from anything other than y-values (*e.g.*, values from a covariate and/or from a previous survey). Cold deck imputation is opposite to hot deck imputation in which a nonrespondent is imputed by a respondent from the same variable in the current survey. The ratio imputation method uses both reported y-values and auxiliary data and is sometimes called a "warm deck" method. The simplest cold deck imputes a nonrespondent y_i, $i \in \mathbf{s} - \mathbf{r}$, by x_i and the resulting HT estimator of Y is

$$\hat{Y}_C = \sum_{i \in \mathbf{r}} w_i y_i + \sum_{i \in \mathbf{s}-\mathbf{r}} w_i x_i. \tag{5}$$

The use of this simple cold deck is motivated by the fact that under model (1), β_k's are close to 1 in many survey problems, especially when x_i's are y-values from a previous survey. When some β_k's are not equal to 1, $\hat{Y}_C$ in (5) has a bias which does not vanish even if $\mathbf{s} = \mathcal{P}$ (*i.e.*, the sample is a census). However, having a small bias may be paid off by lowering the variance so that the overall mean squared error $\text{mse}(\hat{Y}_C) = E(\hat{Y}_C - Y)^2$ may still be smaller than the mean squared error $\text{mse}(\hat{Y}_R) = E(\hat{Y}_R - Y)^2 = V(\hat{Y}_R - Y)$, where E and V denote the expectation and variance under model (1) and repeated sampling. More details can be found in section 2. The simple cold deck may be improved by another cold deck method, the cold deck-ratio method, which imputes a nonrespondent y_i by $x_i \tilde{y}_i / \tilde{x}_i$, where $\tilde{y}_i$ and $\tilde{x}_i$ are reported values from a previous survey. The corresponding HT estimator of Y is

$$\hat{Y}_{C\text{-}\mathbf{R}} = \sum_{i \in \mathbf{r}} w_i y_i + \sum_{i \in \mathbf{s}-\mathbf{r}} w_i x_i \tilde{y}_i / \tilde{x}_i. \tag{6}$$

The estimator in (6) is unbiased if model (1) holds for $\tilde{y}_i$ and $\tilde{x}_i$ (*i.e.*, $\tilde{y}_i = \beta_k \tilde{x}_i + \tilde{x}_i^{1/2} \tilde{e}_i$) with the same β_k as the one for y_i and x_i. These two cold deck methods are widely used in economic area surveys conducted by the U.S. Census Bureau (King and Kornbau 1994) and the U.S. Bureau of Labor Statistics (Butani, Harter and Wolter 1998). Applying cold deck imputation methods does not require knowing the imputation cells, although model (1) is assumed to ensure the unbiasedness of $\hat{Y}_C$ and $\hat{Y}_{C\text{-}\mathbf{R}}$.

Although the cold deck-ratio method makes use of more auxiliary data, it is not always better than the simple cold deck or the ratio imputation method. In section 2 we compare explicitly the mean squared errors of $\hat{Y}_R$, $\hat{Y}_C$ and $\hat{Y}_{C\text{-}\mathbf{R}}$ in a special case where the sample $\mathbf{s}$ is a simple random sample (SRS) and the response probability is a constant. Situations under which one method is better than the others are discussed. If the sampling design or the response mechanism is complex, then it is not easy to compare the mean squared errors explicitly. One may, however, estimate the mean squared errors of $\hat{Y}_R$, $\hat{Y}_C$ and $\hat{Y}_{C\text{-}\mathbf{R}}$ and make an empirical comparison. Variance or mean squared error estimation is itself an important problem, since it is common to report variance or mean squared error estimates along with the estimated totals. These are discussed in section 3.

Our results can also be applied to the problem related to two-phase sampling or double sampling, which is often employed when it is cheap to take a large sample $\{x_i, i \in \mathbf{s}\}$ and expensive to obtain y-values so that a subsample $\{y_i, i \in \mathbf{r}\}$ is taken in the second-phase, $\mathbf{r} \subset \mathbf{s}$.

A numerical example is discussed in section 4 using data from the Transportation Annual Survey conducted by the U.S. Census Bureau.

2. SRS WITH UNIFORM RESPONSE

To illustrate the idea, we start with the simplest case where $\mathbf{s}$ is an SRS (without replacement from $\mathcal{P}$ but the sampling fraction is negligible); there is only one imputation cell so that we can drop the subscript k for imputation cell; and the response probability is a constant $p > 0$ (uniform response mechanism).

In this case $w_i = N/n$, where n is the size of the sample $\mathbf{s}$ and N is the size of the population $\mathcal{P}$. Since $n/N \approx 0$ is assumed,

$$\text{mse}(\hat{Y}_R) \approx \frac{N^2}{n}\left(\frac{\sigma^2 \mu_x}{p} + \beta^2 v_x\right) \tag{7}$$

for large n, where $\mu_x = E(x_i)$ and $v_x = V(x_i)$ and, throughout the paper, $A \approx B$ means that A is equal to B up to a term which is relatively negligible compared to A and B as all sample sizes in imputation cells increase to infinity. A more detailed derivation of result (7) is given in the Appendix. For $\hat{Y}_W$ in (4), it is easy to see that $\tilde{w}_i = N/r$, where $\mathbf{r}$ is the size of $\mathbf{r}$, and $\hat{Y}_W$ is unbiased. Then

$$\text{mse}(\hat{Y}_W) = V(\hat{Y}_W - Y) \approx V(\hat{Y}_W) = \frac{N^2}{n}\left(\frac{\sigma^2 \mu_x}{p} + \frac{\beta^2 v_x}{p}\right).$$

Hence $\hat{Y}_R$ is more efficient than $\hat{Y}_W$ unless $p = 1$ and $\beta^2 v_x = 0$. The gain in using $\hat{Y}_R$ is proportional to β^2 and v_x, both are measures of usefulness of the auxiliary variable x in explaining y through model (1).

For the simple cold deck,

$$\hat{Y}_C = \frac{N}{n}\left(\sum_{i \in \mathbf{r}} y_i + \sum_{i \in \mathbf{s}-\mathbf{r}} x_i\right) = \frac{N}{n}\left(\sum_{i \in \mathbf{r}} x_i^{1/2} e_i + \beta \sum_{i \in \mathbf{r}} x_i + \sum_{i \in \mathbf{s}-\mathbf{r}} x_i\right),$$

where e_i's are defined in (1). Consequently,

$$V(\hat{Y}_C) = \frac{N^2}{n}\left\{\sigma^2 p \mu_x + (\beta^2 p + 1 - p) v_x + (\beta - 1)^2 p(1-p)\mu_x^2\right\} \tag{8}$$

(see the Appendix). The bias of $\hat{Y}_C$ is

$$E(\hat{Y}_C - Y) = N\mu_x(1-p)(1-\beta)$$

and, hence,

$$\begin{aligned}\text{mse}(\hat{Y}_C) &= V(\hat{Y}_C - Y) + [E(\hat{Y}_C - Y)]^2 \\ &\approx V(\hat{Y}_C) + [E(\hat{Y}_C - Y)]^2 \\ &= \frac{N^2}{n}\left\{\sigma^2 p\mu_x + (\beta^2 p + 1 - p)v_x \right. \\ &\qquad \left. + (\beta-1)^2(1-p)[p+n(1-p)]\mu_x^2\right\}. \qquad (9)\end{aligned}$$

Comparing (7) and (9), we obtain the following conclusions.

1. When $p = 1$ (no nonresponse), $\text{mse}(\hat{Y}_C) = \text{mse}(\hat{Y}_R)$.
2. When $p < 1$ and $\beta = 1$ (y and x have the same mean), $\text{mse}(\hat{Y}_C) < \text{mse}(\hat{Y}_R)$.
3. When $p < 1$ and $\beta \neq 1$, $\text{mse}(\hat{Y}_C) \leq \text{mse}(\hat{Y}_R)$ if and only if

$$(\beta-1)^2[p+n(1-p)]\mu_x + (1-\beta^2)v_x/\mu_x - \sigma^2(p+1)/p \leq 0. \qquad (10)$$

Assume that $\mu_x > 0$. In most economic surveys, the relative variance v_x/μ_x^2 is smaller than $p + n(1-p)$. Hence the left hand side of (10) is a quadratic function of β with a positive coefficient in the β^2 term and, therefore, the simple cold deck is better when β is in the interval with limits

$$\frac{[p+n(1-p)]\mu_x \pm \sqrt{v_x^2/\mu_x^2 + \{[p+n(1-p)]\mu_x - v_x/\mu_x\}\sigma^2(p+1)/p}}{[p+n(1-p)]\mu_x - v_x/\mu_x}.$$

This interval contains 1 since (10) holds if $\beta = 1$. Note that $[p + n(1-p)]\mu_x$ increases to infinity as n increases to infinity. Hence the interval of β's for which the simple cold deck is better shrinks to a single point ($\beta = 1$) as $n \to \infty$.

We now consider the cold deck-ratio. Assume that $\tilde{y}_i = \beta\tilde{x}_i + \tilde{x}_i^{1/2}\tilde{e}_i$, $E(\tilde{e}_i) = 0$, $V(\tilde{e}_i) = \sigma^2$, and that $\tilde{e}_i$, e_i, and $(x_i, \tilde{x}_i)$ are mutually independent. Let $z_i = x_i\tilde{y}_i/\tilde{x}_i$ and $\epsilon_i = y_i - z_i = x_i^{1/2}e_i - \tilde{e}_i x_i/\tilde{x}_i^{1/2}$. Then $E(\hat{Y}_{C\text{-}R} - Y) = 0$ and

$$\text{mse}(\hat{Y}_{C\text{-}R}) = \frac{N^2}{n}\left\{\sigma^2 p\mu_x + \beta^2 v_x + \sigma^2(1-p)\gamma_x\right\}, \qquad (11)$$

where $\gamma_x = E(x_i^2/\tilde{x}_i)$ (see the Appendix). By (7) and (11),

$$\text{mse}(\hat{Y}_R) - \text{mse}(\hat{Y}_{C\text{-}R}) = \frac{N^2\sigma^2(1-p)}{n}\left\{\left(\frac{1}{p}+1\right)\mu_x - \gamma_x\right\} \qquad (12)$$

and, hence, the cold deck-ratio is better than the ratio imputation method if and only if $1/p + 1 \geq \gamma_x/\mu_x$. Note that $\gamma_x \geq \mu_x$ and γ_x is close to μ_x if x_i and $\tilde{x}_i$ are highly and positively related, in which case cold deck-ratio imputation can be much better than ratio imputation.

The comparison between the simple cold deck and the cold deck-ratio is the same as that between the simple cold deck and the ratio imputation method. One only needs to replace $(p+1)/p$ in the third term of the left hand side of (10) by γ_x/μ_x.

The parameters β, σ, μ_x, v_x and γ_x have to be estimated in order to compare the efficiencies of $\hat{Y}_R$, $\hat{Y}_C$ and $\hat{Y}_{C\text{-}R}$. Instead, we can directly compare estimated mean squared errors of $\hat{Y}_R$, $\hat{Y}_C$ and $\hat{Y}_{C\text{-}R}$. This is discussed next.

3. STRATIFIED SAMPLING WITH UNCONFOUNDED RESPONSE

We consider the following stratified sampling design adopted by many U.S. government survey agencies: the finite population $\mathcal{P}$ is stratified into H strata with N_h units in the h-th stratum; $n_h \geq 2$ units are selected without replacement from stratum h, according to some probability sampling plan; and the units are selected independently across the strata.

The survey weights w_i's are constructed so that if all y_i's are observed, the HT estimator $\sum_{i\in s} w_i y_i$ is unbiased for Y under repeated sampling.

We assume model (1). The response probability is no longer a constant, but independent of the y-value. For the cold deck-ratio, we also assume that within the k-th imputation cell, $\tilde{y}_i = \beta_k\tilde{x}_i + \tilde{x}_i^{1/2}\tilde{e}_i$, $E(\tilde{e}_i) = 0$, $V(\tilde{e}_i) = \tilde{\sigma}_k^2$ and e_i, $\tilde{e}_i$, $(x_i, \tilde{x}_i)$ are mutually independent.

Explicit results for the mean squared errors such as (7), (9) and (11) are not easy to obtain. We may, however, make empirical comparisons of the efficiencies of $\hat{Y}_R$, $\hat{Y}_C$ and $\hat{Y}_{C\text{-}R}$, based on their estimated mean squared errors. Estimation of the mean squared errors of $\hat{Y}_R$, $\hat{Y}_C$ and $\hat{Y}_{C\text{-}R}$, is in fact an important part of the sampling theory. It is well known that for imputed data sets, the naive method that applies the standard variance estimation formulas by treating imputed nonrespondents as observed data leads to underestimation. When no correct method (for estimating the mean squared error) is available, the naive method is used in many survey agencies.

We now derive estimators for $V(\hat{Y})$ or $\text{mse}(\hat{Y})$ that are correct under model (1), where $\hat{Y}$ denotes $\hat{Y}_R$, $\hat{Y}_C$ or $\hat{Y}_{C\text{-}R}$.

Let E_m and V_m be the expectation and variance with respect to model (1) and let E_s and V_s be the expectation and variance with respect to repeated sampling (conditional on the model and response). Then

$$V(\hat{Y} - Y) = E_m[V_s(\hat{Y})] + V_m[E_s(\hat{Y}) - Y]. \qquad (13)$$

We first consider $E_m[V_s(\hat{Y})]$, the first variance component in (13). It suffices to obtain an estimator of $V_s(\hat{Y})$, conditional on $\{y_i, x_i, a_i, i \in \mathcal{P}\}$ (and $\{\tilde{y}_i, \tilde{x}_i, i \in \mathcal{P}\}$ for cold deck-ratio), where a_i is the response indicator for y_i.

The estimation of $V_s(\hat{Y}_C)$ and $V_s(\hat{Y}_{C\text{-}\mathbf{R}})$ is simple (which is an advantage of using a cold deck method). Let

$$v_1 = \sum_h \left(1 - \frac{n_h}{N_h}\right) \frac{n_h}{n_h - 1} \sum_{i \in \mathbf{s}(\mathbf{h})} \left(w_i t_i - \frac{1}{n_h} \sum_{i \in \mathbf{s}(\mathbf{h})} w_i t_i\right)^2 \quad (14)$$

be the standard variance estimator for $\sum_{i \in \mathbf{s}} w_i t_i$ when $\{t_i, i \in \mathbf{s}\}$ is treated as an observed sample (from $\{t_i, i \in \mathcal{P}\}$), where $\mathbf{s}(\mathbf{h})$ is $\mathbf{s}$ restricted to stratum h. Then $V_s(\hat{Y}_C)$ can be estimated by using (14) with $t_i = a_i y_i + (1 - a_i)x_i$ and $V_s(\hat{Y}_{C\text{-}\mathbf{R}})$ can be estimated by using (14) with $t_i = a_i y_i + (1 - a_i)x_i \tilde{y}_i / \tilde{x}_i$.

The estimation of $V_s(\hat{Y}_R)$ is slightly more complicated but similar. Assume that in each imputation cell, the number of sampled units is large and the response probabilities are bounded away from 0. Note that

$$\begin{aligned}\hat{Y}_R &= \sum_k \left[\left(\sum_{i \in \mathbf{s}_k} w_i x_i \Big/ \sum_{i \in \mathbf{r}_k} w_i x_i\right) \right. \\ &\qquad \left. \times \sum_{i \in \mathbf{r}_k} w_i (y_i - \beta_k x_i) + \beta_k \sum_{i \in \mathbf{s}_k} w_i x_i\right] \\ &\approx \sum_k \left[\zeta_k \sum_{i \in \mathbf{s}_k} w_i a_i (y_i - \beta_k x_i) + \beta_k \sum_{i \in \mathbf{s}_k} w_i x_i\right] \\ &= \sum_{i \in \mathbf{s}} w_i \left[\zeta_i a_i (y_i - \beta_i x_i) + \beta_i x_i\right],\end{aligned}$$

where $\zeta_k = E(\sum_{i \in \mathbf{s}_k} w_i x_i) / E(\sum_{i \in \mathbf{r}_k} w_i x_i)$ and $\zeta_i = \zeta_k$ and $\beta_i = \beta_k$ for $i \in \mathbf{s}_k$. After estimating β_k by $\hat{\beta}_k$ and ζ_k by $\hat{\zeta}_k = \sum_{i \in \mathbf{s}_k} w_i x_i / \sum_{i \in \mathbf{r}_k} w_i x_i$, we estimate $V_s(\hat{Y}_R)$ by using (14) with $t_i = \hat{\zeta}_i a_i (y_i - \hat{\beta}_i x_i) + \hat{\beta}_i x_i$, where $\hat{\zeta}_i = \hat{\zeta}_k$ and $\hat{\beta}_i = \hat{\beta}_k$ for $i \in \mathbf{s}_k$.

Before we discuss the estimation of $V_m[E_s(\hat{Y}) - Y]$, the second variance component in (13), it should be noted that $V_m[E_s(\hat{Y}) - Y] / E_m[V_s(\hat{Y})] = O(n/N)$. This is because the variance of $E_s(\hat{Y}) - Y$ (if it is nonzero) is typically of the order N, whereas the order of $V_s(\hat{Y})$ is typically N^2/n and thus the order of $E_m[V_s(\hat{Y})]$ is N^2/n under some regularity conditions. Hence, in theory, it is not necessary to estimate $V_m[E_s(\hat{Y}) - Y]$ if the sampling fraction n/N is negligible. However, the constant in $O(n/N)$ is unknown and, hence, one may still want to estimate $V_m[E_s(\hat{Y}) - Y]$ in applications even when n/N is small.

We now consider the estimation of the second variance component in (13). For $\hat{Y}_C$,

$$\begin{aligned}E_s(\hat{Y}_C) - Y &= \sum_{i \in \mathcal{P}} [a_i y_i + (1 - a_i) x_i] - \sum_{i \in \mathcal{P}} y_i = \\ &\qquad - \sum_{i \in \mathcal{P}} (1 - a_i)(y_i - x_i).\end{aligned}$$

Then, under model (1),

$$\begin{aligned}&V_m[E_s(\hat{Y}_C) - Y] = \\ &E_m\left[\sum_k \sigma_k^2 \sum_{i \in \mathcal{P}_k} (1 - a_i) x_i\right] + V_m\left[\sum_{i \in \mathcal{P}} (1 - a_i)(\beta_i - 1) x_i\right].\end{aligned}$$

If we estimate σ_k^2 by

$$\hat{\sigma}_k^2 = \sum_{i \in \mathbf{s}_k} a_i w_i (y_i - \hat{\beta}_k x_i)^2 \Big/ \sum_{i \in \mathbf{s}_k} a_i w_i x_i,$$

then an estimator of $V_m[E_s(\hat{Y}_C) - Y]$ is

$$\begin{aligned}v_{2C} = &\sum_k \hat{\sigma}_k^2 \sum_{i \in \mathbf{s}_k} (1 - a_i) w_i x_i + \\ &\sum_h \frac{N_h}{n_h - 1} \sum_{i \in \mathbf{s}(h)} \left(u_i - \frac{1}{n_h} \sum_{i \in \mathbf{s}(\mathbf{h})} u_i\right)^2,\end{aligned} \quad (15)$$

where $u_i = (1 - a_i)(\hat{\beta}_i - 1)x_i$ and $\hat{\beta}_i = \hat{\beta}_k$ for $i \in \mathbf{s}_k$.

For $\hat{Y}_{C\text{-}\mathbf{R}}$,

$$E_s(\hat{Y}_{C\text{-}\mathbf{R}}) - Y = -\sum_{i \in \mathcal{P}} (1 - a_i)(y_i - x_i \tilde{y}_i / \tilde{x}_i)$$

and

$$\begin{aligned}&V_m\left[E_s(\hat{Y}_{C\text{-}\mathbf{R}}) - Y\right] = \\ &E_m\left[\sum_k \sigma_k^2 \sum_{i \in \mathcal{P}_k} (1 - a_i) x_i + \sum_k \tilde{\sigma}_k^2 \sum_{i \in \mathcal{P}_k} (1 - a_i) x_i^2 / \tilde{x}_i\right].\end{aligned}$$

Hence $V_m[E_s(\hat{Y}_{C\text{-}\mathbf{R}}) - Y]$ can be estimated by

$$v_{2C\text{-}\mathbf{R}} = \sum_k \left[\hat{\sigma}_k^2 \sum_{i \in \mathbf{s}_k} (1 - a_i) w_i x_i + \hat{\tilde{\sigma}}_k^2 \sum_{i \in \mathbf{s}_k} (1 - a_i) w_i x_i^2 / \tilde{x}_i\right], \quad (16)$$

where

$$\hat{\tilde{\sigma}}_k^2 = \sum_{i \in \mathbf{s}_k} w_i (\tilde{y}_i - \hat{\tilde{\beta}}_k \tilde{x}_i)^2 \Big/ \sum_{i \in \mathbf{s}_k} w_i \tilde{x}_i$$

and

$$\hat{\tilde{\beta}}_k = \sum_{i \in \mathbf{s}_k} w_i \tilde{y}_i \Big/ \sum_{i \in \mathbf{s}_k} w_i \tilde{x}_i.$$

For $\hat{Y}_R$,

$$E_s(\hat{Y}_R) - Y \approx \sum_k \left[\left(\sum_{i \in \mathcal{P}_k} x_i \Big/ \sum_{i \in \mathcal{P}_k} a_i x_i\right) \sum_{i \in \mathcal{P}_k} a_i y_i - \sum_{i \in \mathcal{P}_k} y_i\right]$$

and from Taylor's expansion,

$$\begin{aligned}&V_m[E_s(\hat{Y}_R) - Y] \approx \\ &E_m\left\{\sum_k \sigma_k^2 \left[\sum_{i \in \mathcal{P}_k} x_i \sum_{i \in \mathcal{P}_k} (1 - a_i) x_i\right] \Big/ \sum_{i \in \mathcal{P}_k} a_i x_i\right\}.\end{aligned}$$

It can be estimated by

$$v_{2R} = \sum_k \hat{\sigma}_k^2 \left[\sum_{i \in \mathbf{s}_k} w_i x_i \sum_{i \in \mathbf{s}_k} (1 - a_i) w_i x_i\right] \Big/ \sum_{i \in \mathbf{s}_k} a_i w_i x_i. \quad (17)$$

Finally, $\hat{Y}_R$ and $\hat{Y}_{C\text{-}\mathbf{R}}$ are unbiased but $\hat{Y}_C$ has a bias

$$\sum_k (1 - \beta_k) E_m\left[\sum_{i \in \mathcal{P}_k} (1 - a_i) x_i\right],$$

which can be estimated by

$$\sum_k (1-\hat{\beta}_k)\sum_{i\in s_k}(1-a_i)w_i x_i.$$

Thus, we obtain the following estimated mean squared errors: $\text{mse}(\hat{Y}_R)$ can be estimated by

$$\widehat{\text{mse}}(\hat{Y}_R) = v_{1R} + v_{2R},$$

where v_{1R} is obtained using (14) with $t_i = \hat{\zeta}_i a_i(y_i - \hat{\beta}_i x_i) + \hat{\beta}_i x_i$, $\hat{\zeta}_i = \hat{\zeta}_k$ and $\hat{\beta}_i = \hat{\beta}_k$ for $i \in s_k$, and v_{2R} is given by (17); $\text{mse}(\hat{Y}_C)$ by

$$\widehat{\text{mse}}(\hat{Y}_C) = v_{1C} + v_{2C} + \left[\sum_k (1-\hat{\beta}_k)\sum_{i\in s_k - r_k} w_i x_i\right]^2,$$

where v_{1C} is obtained by using (14) with $t_i = a_i y_i + (1-a_i)x_i$ and v_{2C} is given by (15); and $\text{mse}(\hat{Y}_{C\text{-}\mathbf{R}})$ can be estimated by

$$\widehat{\text{mse}}(\hat{Y}_{C\text{-}\mathbf{R}}) = v_{1C\text{-}\mathbf{R}} + v_{2C\text{-}\mathbf{R}},$$

where $v_{1C\text{-}\mathbf{R}}$ is obtained by using (14) with $t_i = a_i y_i + (1-a_i)x_i \tilde{y}_i/\tilde{x}_i$ and $v_{2C\text{-}\mathbf{R}}$ is given by (16).

Under model (1) and the asymptotic settings in Krewski and Rao (1981), Rao and Shao (1992) or Valliant (1993), the derived mean squared error estimators are asymptotically unbiased and consistent as all sample sizes in imputation cell increase to infinity.

For cold deck or cold deck-ratio imputation, the first term (v_{1C} or $v_{1C\text{-}\mathbf{R}}$) in the estimated mean squared error is the same as the one obtained by applying a standard formula (such as (14)) and treating imputed nonrespondents as observed data. For ratio imputation, applying (14) and treating imputed nonrespondents as observed data produces the following estimator of $\text{mse}(\hat{Y}_R)$:

$$\tilde{v}_{1R} = \sum_h \left(1 - \frac{n_h}{N_h}\right)\frac{n_h}{n_h - 1}\sum_{i\in s(h)}\left(w_i z_i - \frac{1}{n_h}\sum_{i\in s(h)} w_i z_i\right)^2 \quad (18)$$

with $z_i = a_i y_i + (1-a_i)\hat{\beta}_i x_i$, which is different from the first term v_{1R} in our estimator $\widehat{\text{mse}}(\hat{Y}_R)$ and, hence, is not asymptotically valid even if n/N is negligible.

4. AN EXAMPLE

We consider an example using a data set from the Transportation Annual Survey (TAS) conducted by the U.S. Census Bureau.

The TAS is a survey of firms with one or more establishments that are primarily engaged in providing commercial motor freight transportation or public warehousing services in U.S. A stratified simple random sample is selected without replacement from employers contained in the Census Bureau's Standard Statistical Establishment List. The strata, which are also the imputation classes in this example, are constructed according to company's size within each industry.

There are various variables in this survey. We consider the estimation of the population totals of the current year annual revenue (y) in four industries. The variable y has nonrespondents. Three covariates without nonrespondents are considered: the current year annual payroll, the previous year annual revenue, and the previous year annual payroll. The sample size, response size for y, and the sampling weight in each stratum and industry are given in Table 1.

Table 1
Sample Sizes, Response Sizes, and Sampling Weights Across Industries and Strata

Industry	Stratum	Sample Size	Response Size	Sampling Weight
1	0	31	24	1.00
	1	14	6	12.43
	2	11	7	8.91
	3	10	4	6.10
	4	11	6	5.73
	5	16	12	2.70
	6	18	13	2.17
2	0	86	82	1.00
	1	8	2	32.91
	2	13	10	9.85
	3	11	9	10.82
	4	12	10	6.08
	5	13	10	3.60
3	0	38	30	1.00
	1	14	9	87.91
	2	11	8	67.39
	3	13	10	44.48
	4	14	13	25.28
	5	16	13	15.57
	6	18	12	9.80
	7	15	11	6.23
	8	15	14	4.68
	9	40	33	2.13
4	0	28	23	1.00
	1	7	5	32.14
	2	13	6	16.75
	3	10	7	12.90
	4	14	12	7.00
	5	13	9	6.18
	6	11	7	4.70
	7	17	12	3.31
	8	19	14	1.89
	9	22	16	1.82

First, we use the previous year annual revenue as the covariate x in simple cold deck imputation and ratio imputation. The current year annual payroll and the previous year annual payroll are used as $\tilde{y}$ and $\tilde{x}$, respectively. For four industries and three imputation methods, Table 2 lists the estimated totals, the proposed estimated MSE's for the estimated totals, the naive estimated MSE's for the estimated totals (obtained by treating imputed values as

observed data), and the MSE ratios (the proposed estimated MSE over the naive estimated MSE). Note that the proposed estimated MSE is the sum of v_1 and v_2 for the ratio and cold deck-ratio methods or the sum of v_1, v_2, and the squared estimated bias for the simple cold deck method. Values of v_1 and v_2 are also included in the table.

Table 2
Estimated Totals and MSE's When x = the Previous Year Annual Revenue, $\tilde{y}$= the Current Year Payroll, and $\tilde{x}$ = the Previous Year Annual Payroll

		Method		
Industry	Estimate	Cold Deck	Cold Deck-Ratio	Ratio
1	Total	5.31×10^9	5.19×10^9	5.42×10^9
	v_1	7.73×10^{14}	8.46×10^{14}	2.60×10^{15}
	v_2	1.39×10^{15}	2.50×10^{15}	1.81×10^{15}
	Proposed MSE	2.30×10^{15}	3.34×10^{15}	4.40×10^{15}
	Naive MSE	7.73×10^{14}	8.46×10^{14}	2.46×10^{15}
	MSE Ratio	2.97	3.95	1.79
2	Total	1.66×10^{10}	1.63×10^{10}	1.67×10^{10}
	v_1	4.00×10^{15}	4.19×10^{15}	5.57×10^{16}
	v_2	6.03×10^{15}	2.88×10^{16}	6.54×10^{15}
	Proposed MSE	1.02×10^{16}	3.30×10^{16}	6.23×10^{16}
	Naive MSE	4.00×10^{15}	4.19×10^{15}	5.58×10^{16}
	MSE Ratio	2.54	7.87	1.12
3	Total	3.54×10^{10}	3.53×10^{10}	3.59×10^{10}
	v_1	1.32×10^{16}	1.80×10^{16}	1.94×10^{17}
	v_2	5.44×10^{16}	8.62×10^{16}	6.77×10^{16}
	Proposed MSE	6.97×10^{16}	1.04×10^{17}	2.62×10^{17}
	Naive MSE	1.32×10^{16}	1.80×10^{16}	1.87×10^{17}
	MSE Ratio	5.27	5.80	1.40
4	Total	1.27×10^{10}	1.22×10^{10}	1.30×10^{10}
	v_1	2.11×10^{16}	2.14×10^{16}	5.13×10^{15}
	v_2	3.91×10^{15}	8.26×10^{15}	5.06×10^{15}
	Proposed MSE	2.59×10^{16}	2.97×10^{16}	1.02×10^{16}
	Naive MSE	2.11×10^{16}	2.14×10^{16}	5.06×10^{15}
	MSE Ratio	1.23	1.39	2.01

Next, to see the effect of using a wrong covariate in using the simple cold deck method, we repeat the previous computations using the current year annual payroll as the covariate x, and the previous year annual revenue and payroll as $\tilde{y}$ and $\tilde{x}$, respectively. The results are reported in Table 3.

The following is a summary of the results in Tables 2 and 3.

1. The simple cold deck method depends heavily on the choice of the covariate x. When x is the previous year annual revenue (Table 2), the difference among the estimated totals provided by three methods is negligible; in terms of the estimated MSE, the simple cold deck method is the best. However, when x is the current year annual payroll (Table 3), the estimates from the simple cold deck is obviously too low; in terms of the estimated MSE, the simple cold deck method is the worst, because of its large bias (shown in Table 3).

Table 3
Estimated Totals and MSE's When x = the Current Year Annual Payroll, $\tilde{y}$= the Previous Year Annual Revenue, and $\tilde{x}$ = the Previous Year Annual Payroll

		Method		
Industry	Estimate	Cold Deck	Cold Deck-Ratio	Ratio
1	Total	4.49×10^9	5.19×10^9	5.39×10^9
	Bias	-8.99×10^8		
	v_1	8.10×10^{14}	8.46×10^{14}	2.85×10^{15}
	v_2	1.38×10^{15}	2.64×10^{15}	1.75×10^{15}
	Proposed MSE	1.03×10^{16}	3.49×10^{15}	4.60×10^{15}
	Naive MSE	8.10×10^{14}	8.46×10^{14}	2.55×10^{15}
	MSE Ratio	12.68	4.12	1.81
2	Total	1.59×10^{10}	1.63×10^{10}	1.71×10^{10}
	Bias	-1.21×10^9		
	v_1	4.36×10^{15}	4.19×10^{15}	5.74×10^{16}
	v_2	8.20×10^{15}	1.48×10^{16}	8.95×10^{15}
	Proposed MSE	2.73×10^{16}	1.90×10^{16}	6.64×10^{16}
	Naive MSE	4.36×10^{15}	4.19×10^{15}	5.62×10^{16}
	MSE Ratio	6.25	4.54	1.18
3	Total	3.10×10^{10}	3.53×10^{10}	3.47×10^{10}
	Bias	-3.62×10^9		
	v_1	1.25×10^{16}	1.80×10^{16}	2.30×10^{17}
	v_2	4.56×10^{16}	9.25×10^{16}	5.41×10^{16}
	Proposed MSE	1.89×10^{17}	1.10×10^{17}	2.84×10^{17}
	Naive MSE	1.25×10^{16}	1.80×10^{16}	1.83×10^{17}
	MSE Ratio	15.13	6.15	1.56
4	Total	1.06×10^{10}	1.22×10^{10}	1.20×10^{10}
	Bias	-1.35×10^9		
	v_1	1.93×10^{16}	2.14×10^{16}	5.84×10^{15}
	v_2	2.67×10^{15}	4.62×10^{15}	3.07×10^{15}
	Proposed MSE	4.03×10^{16}	2.60×10^{16}	8.92×10^{15}
	Naive MSE	1.93×10^{16}	2.14×10^{16}	8.92×10^{15}
	MSE Ratio	2.09	1.22	1.72

2. There is no definite conclusion on the relative performance (in terms of the estimated MSE) of the ratio imputation method and the cold deck-ratio method. In this example, the cold deck-ratio is better for industries 1-3, whereas the ratio imputation method is better for industry 4. Some scatter plots of the data (not shown) indicate that the correlation between x and $\tilde{x}$ in industries 1-3 is higher than that in industry 4, which might be the reason for the difference in relative performance of the two imputation methods. See also the discussion after formula (12).

3. The naive estimated MSE's are much lower than the proposed estimated MSE's and are too optimistic. For example, in Table 3, the naive MSE's for the simple cold deck method are always smaller than those for the cold deck-ratio method, although we know that the simple cold deck does not work well in this case. In this example, v_2/v_1 is not small because of some large sampling fractions. Since the naive estimated MSE is either equal to v_1 (for the cold deck imputation

methods) or not very different from v_1 (for ratio imputation), the underestimation in using the naive estimated MSE is mainly due to treating imputed values as observed values in strata with large sampling fractions (and ignoring the bias of the simple cold deck estimators in the case of Table 3).

ACKNOWLEDGEMENT

The author would like to thank referees for helpful comments and suggestions. The first draft of this paper was finished at the U.S. Census Bureau and the U.S. Bureau of Labor Statistics when the author was an ASA/NSF Senior Research Fellow. The research was also supported by National Science Foundation Grants DMS-9504425 and DMS-9803112 and National Security Agency Grant MDA904-99-1-0032.

APPENDIX

1. **Proof of (7)**: When $n/N \approx 0$, $V(\hat{Y}_R - Y) \approx V(\hat{Y}_R)$. Then (7) follows from

$$V(\hat{Y}_R) = \frac{N^2}{n^2}\left\{\sigma^2 E\left[\left(\sum_{i\in s} x_i\right)^2 \Big/ \left(\sum_{i\in r} x_i\right)\right] + \beta^2 V\left(\sum_{i\in s} x_i\right)\right\}$$
$$\approx \frac{N^2}{n}\left(\frac{\sigma^2\mu_x}{p} + \beta^2 v_x\right)$$

for large n, where the last approximate equality follows from the fact that conditioned on x_i's, $E(\sum_{i\in r} x_i) = p\sum_{i\in s} x_i$.

2. **Proof of (9)**: Under model (1),

$$V(\hat{Y}_C) = \frac{N^2}{n^2}\left\{V\left(\sum_{i\in r} x_i^{1/2} e_i\right) + V\left(\beta\sum_{i\in r} x_i + \sum_{i\in s-r} x_i\right)\right\}$$
$$= \frac{N^2}{n^2}\left\{\sigma^2 p\mu_x + \beta^2 V\left(\sum_{i\in r} x_i\right) + V\left(\sum_{i\in s-r} x_i\right) + 2\beta\mathrm{Cov}\left(\sum_{i\in r} x_i, \sum_{i\in s-r} x_i\right)\right\}$$
$$= \frac{N^2}{n}\{\sigma^2 p\mu_x + \beta^2[pv_x + p(1-p)\mu_x^2] + (1-p)(v_x + p\mu_x^2) - 2\beta p(1-p)\mu_x^2\}$$
$$= \frac{N^2}{n}\{\sigma^2 p\mu_x + (\beta^2 p + 1 - p)v_x + (\beta - 1)^2 p(1-p)\mu_x^2\}.$$

3. **Proof of (11)**: Under the assumed conditions on (y_i, x_i) and $(\tilde{y}_i, \tilde{x}_i)$,

$$\mathrm{mse}(\hat{Y}_{C\text{-}R}) = \frac{N^2}{n^2} V\left(\sum_{i\in r} y_i + \sum_{i\in s-r} z_i\right)$$
$$= \frac{N^2}{n^2} V\left(\sum_{i\in r} \epsilon_i + \sum_{i\in s} z_i\right)$$
$$= \frac{N^2}{n^2}\left\{V\left(\sum_{i\in r} \epsilon_i\right) + V\left(\sum_{i\in s} z_i\right) + 2\mathrm{Cov}\left(\sum_{i\in r} \epsilon_i, \sum_{i\in s} z_i\right)\right\}$$
$$= \frac{N^2}{n}\{\sigma^2 p(\mu_x + \gamma_x) + (\beta^2 v_x + \sigma^2\gamma_x) - 2\sigma^2 p\gamma_x\}$$
$$= \frac{N^2}{n}\{\sigma^2 p\mu_x + \beta^2 v_x + \sigma^2(1-p)\gamma_x\}.$$

REFERENCES

BUTANI, S., HARTER, R., and WOLTER, K. (1998). Estimation procedures for the Bureau of Labor Statistics current employment statistics program. *Proceedings of the Section on Survey Research Methods, American Statistical Association.*

COCHRAN, W.G. (1977). *Sampling Techniques.* Third Edition. New York: Wiley.

KALTON, G., and KASPRZYK, D. (1986). The treatment of missing data. *Survey Methodology*, 12, 1-16.

KING, C., and KORNBAU, M. (1994). Inventory of economic area statistical practices. ESMD Report Series 9401, Bureau of the Census, Washington D.C.

KREWSKI, D., and RAO, J.N.K. (1981). Inference from stratified samples: properties of the linearization, jackknife and balanced repeated replication methods. *Annals of Statistics*, 9, 1010-1019.

LEE, H., RANCOURT, E., and SÄRNDAL, C.-E. (1994). Experiments with variance estimation from survey data with imputed values. *Journal of Official Statistics*, 10, 231-243.

RAO, J.N.K., and SHAO, J. (1992). Jackknife variance estimation with survey data under hot deck imputation. *Biometrika*, 79, 811-822.

RUBIN, D.B. (1976). Inference and missing data. *Biometrika*, 63, 581-592.

VALLIANT, R. (1993). Poststratification and conditional variance estimation. *Journal of American Statistical Association*, 88, 89-96.

Survey Methodology, June 2000
Vol. 26, No. 1, pp. 87-98
Statistics Canada

Model-Based Estimation With Link-Tracing Sampling Designs

STEVEN K. THOMPSON and OVE FRANK[1]

ABSTRACT

Samples from hidden and hard-to-access human populations are often obtained by procedures in which social links are followed from one respondent to another. Inference from the sample to the larger population of interest can be affected by the link-tracing design and the type of data it produces. The population with its social network structure can be modeled as a stochastic graph with a joint distribution of node values representing characteristics of individuals and arc indicators representing social relationships between individuals. In this paper maximum likelihood estimators of population graph parameters are described. Predictors of realized population graph quantities are obtained using predictive likelihood. These estimators and predictors are compared with conventional data summaries and illustrated with a numerical example.

KEY WORDS: Snowball samples; Adaptive sampling; Graph sampling; Ignorable designs; Link-tracing designs; Network sampling; Likelihood; Predictive likelihood.

1. INTRODUCTION

In studies of hidden and hard-to-access human populations, link-tracing procedures, in which social links are followed from one respondent to another, are commonly involved in obtaining the sample. For example, in a study of injection drug use in relation to the spread of the HIV infection, initial respondents may he asked to identify drug-injection or sexual partners who are then added to the sample. For such a study, the social links are of inherent importance for understanding the issues of interest while at the same time being useful or essential in building the sample. However, inference from the sample to the larger population or social structure of interest can be affected by the link-tracing procedures and the type of data they produce. In this paper we evaluate this inference problem in relation to the design and describe some inference methods for such studies based on maximum likelihood estimation and prediction.

Human populations with social structure are often modeled as graphs, with the nodes of the graph representing individuals and the edges or arcs of the graph representing social links, relationships, or transactions. The population graph itself can be viewed either as a fixed structure or as a realization of a stochastic graph model. In real studies of human populations, particularly those that are hidden or hard to access, it is seldom possible to obtain data on the whole population or graph structure. Rather, data are obtained from a sample, and the sample may have been obtained by innovative and unconventional means, including methods taking advantage of the arcs or links from one individual to another. The data may contain information about characteristics of sample individuals, social links within the sample, and in some cases information about links between individuals in the sample and individuals outside the sample.

In this paper we use the term "sampling design" to refer to the procedure by which the sample is selected, whether deliberate or happenstance. For many ethnographic and sociological studies of hidden populations, link-tracing designs are considered the only practical way to obtain a sample large enough to study. In other studies, the social structure is itself the object of interest and the link-tracing methods are used in order to obtain meaningfully structured samples to study.

The statistical literature on design and estimation with link-tracing designs includes procedures variously termed snowball sampling, chain-referral sampling, random walks, nexus sampling, network or multiplicity sampling, and adaptive sampling. A type of link-tracing design in which individuals in an initial sample were asked to identify a fixed number of acquaintances, who in turn were asked to identify the same number of acquaintances, and so on for a fixed number of stages or waves, was termed "snowball sampling" by Goodman (1961). A Bernoulli procedure was assumed for the initial sample. Snowball designs were developed in the graph setting with a variety of initial probability sampling designs and any numbers of links and waves by Frank (1971, 1977a,b, 1978a,b, 1979a), who obtained a variety of design and model based methods for estimating graph quantities from the sample data. Snijders (1992) used the same term "snowball sampling" to include designs in which only a subsample of links from each node is traced. The case in which only one of the links from a node is selected at random and followed to another node, and then one of its links selected, and so on, was called a "random walk" by Klovdahl 1989. Link-tracing sampling methods in which there is only one link from each node have been termed "chains" (Erickson 1979). Frank and Snijders (1994) consider model- and design-based

[1] Steven K. Thompson, Department of Statistics, 326 Thomas Building, Pennsylvania State University, University Park, PA 16802 USA; Ove Frank, Department of Statistics, Stockholm University, S-10691 Stockholm, Sweden. This research is part of an ongoing, equal collaboration effort and order of authorship was determined by a coin toss.

estimation of a hidden population size, that is, the number of nodes in the graph, based on snowball samples. Additional practical and statistical issues in sampling from social networks with various types of snowball, chain-referral, and other link-tracing designs are discussed in Granovetter (1976), Morgan and Rytina (1977), Frank (1979b, 1981, 1988), Watters and Biernacki (1989), van Meter (1990), Spreen (1992), Wasserman and Faust (1994), Spreen and Zwaagstra (1994), Karlberg (1997), Jansson (1997), Spreen (1998), and Robins (1998).

Design-based estimation methods were developed additionally for the closely related designs of network or multiplicity sampling, in which social, kinship, and administrative links were traced (Birnbaum and Sirken 1965, Kalton and Anderson 1986, Levy 1977, Levy and Lemeshow 1991, Sirken 1970, 1972a, b, Sirken and Levy 1974, Sudman, Sirken, and Cowan 1988). For example, in a survey of a rare disease, an initial sample of households might be selected at random and data obtained both for residents of the households and for their siblings. The design-based estimation in these strategies is helped by the symmetry of the links and the encompassing of complete connected components in the sample, and unbiased estimators have been obtained for network sampling with many different initial designs.

Another link-tracing procedure for which design-based estimators are available is adaptive cluster sampling (Thompson 1990, 1997, Thompson and Seber 1996), which has been formulated in the graph setting as well as the spatial setting. Following selection of an initial sample of nodes by any of a number of initial designs, the decision on whether to follow links from a node or not depends on the value of a variable of interest observed for the node. For example, in an epidemiological study of a sexually transmitted disease, sexual or social links may be followed only from respondents who have been infected. Design-unbiased estimation methods have been worked out for a wide variety of adaptive cluster sampling strategies.

Design-based methods of inference, such as the design-based estimation methods of network sampling, snowball sampling, and adaptive cluster sampling, have the advantage that properties such as design-unbiasedness or consistency do not depend for their validity on any assumed model for the population. On the other hand, these properties do depend on the sampling design being carried out as specified. The model-based methods described in this paper, on the other hand, do depend on an assumed model for the population or graph. Their practical advantage is that they apply to a wide range of sample selection procedures, and thus allow more leeway in how the sample is actually selected.

In fact many real studies of hidden and hard-to-reach populations use sample selection procedures, including link-tracing, that are not readily analyzed based on idealized design-induced probabilities. For example, in a study to examine the relation of network structure and risk behaviors such as needle sharing among drug injectors in the Bushwick section of Brooklyn, "index" (initial) respondents were used as "auxiliary recruiters" to bring members of their networks into the study (Friedman, Neaigus, Jose, Curtis, Goldstein, Ildefonso, Rothenberg and Des Jarlais 1997, Neaigus, Friedman, Goldstein, Ildefonseo, Curtis and Jose 1995, Neaigus, Friedman, Jose, Goldstein, Curtis, Ildefonso and Des Jarlais1996). Only about 61% of the linked individuals were actually recruited, however. In a long-term study on the heterosexual transmission of HIV infection (Rothenberg, Woodhouse, Potterat, Muth, Darrow and Klovdahl 1995), the target population of interest consisted of commercial sex workers, their paying and nonpaying partners, persons who use injectable drugs, and the sexual partners of drug users in the Colorado Springs area. Persons in the purposively-selected initial sample were interviewed and, in addition to their individual characteristics, identities of their sexual partners were obtained. Persons named by two or more respondents were also located and interviewed. The wide range of link-tracing procedures used in studies such as these has motivated the emphasis in this paper on model-based inference methods.

When we compare the maximum likelihood estimators and predictors obtained in this paper with commonly-used conventional estimates or data summaries such as sample means and proportions of node or link values, we find that in most cases the conventional estimates are not the best estimates. Similarly, estimators that would be appropriate if the data included the whole graph may not be appropriate with data on only a sample from the graph. An implication of these results is that conventional estimates or unadjusted summaries of sample data obtained through link-tracing procedures can be misleading if viewed as pertaining to population or whole-graph characteristics. The interpretations of this discrepancy provided in this paper give some insight into the conditions under which the best estimate would tend to be lower, or higher, than the conventional one.

Notation and basic issues for design and inference in the graph setting are presented in section 2. In section 3, a wide range of link-tracing procedures, all of which can be analyzed using the approach presented in this paper, are described. In section 4, a class of graph models that we use to illustrate the inference methods of the paper is described. Estimative and predictive maximum likelihood methods for graph parameters and realized population values are described in section 5.

2. GRAPH MODELS AND SAMPLING DESIGNS

Consider a graph of N nodes (units) labeled $1, 2, ..., N$. Associated with the u-th node is a variable of interest Y_u. We denote the full set of node labels $U = \{1, 2, ..., N\}$ and the sequence of node variables by $\mathbf{Y} = (Y_1, ..., Y_N)$. For two distinct nodes u and v, the indicator variable X_{uv} equals

one if there is an arc (directional link) from u to v and zero otherwise. The matrix of arc indicators, having X_{uv} as the element in the u-th row and v-th column, is the graph adjacency matrix, denoted **X**. For convenience we will assume the diagonal elements X_{uu} are zero. The ordered pair (u, v) is sometimes referred to as a dyad of type $(Y_u, Y_v; X_{uv}, X_{vu})$. A graph model is given by a joint probability or density $f(\mathbf{y}, \mathbf{x}; \psi)$ for outcomes **y** and **x** of **Y** and **X**, respectively, and it may depend on one or more unknown parameters ψ.

A sample s from the graph is a subset of nodes and a subset of node pairs. We can write the combined sample as $s = (s^{(1)}, s^{(2)})$, where $s^{(1)}$ denotes the subset of nodes selected for observation of the associated y-values and $s^{(2)}$ denotes the subset of node pairs selected for observation of the associated x-values. The data consist of the node and node-pair labels in the combined sample together with the associated node and arc-indicator values, that is $d = \left(u, (v,w), y_u, x_{vw}: u \in s^{(1)}, (v,w) \in s^{(2)}\right)$ or, more simply, $d = (s, \mathbf{y}_{s^{(1)}}, \mathbf{x}_{s^{(2)}})$. Further, it is often convenient to use $\mathbf{y}_s$ to denote the y-values of the nodes in the combined sample and $\mathbf{x}_s$ for the x-values of the node pairs in the combined sample, with $\mathbf{y}_{\bar{s}}$ and $\mathbf{x}_{\bar{s}}$ denoting the values of the unsampled nodes and node pairs. Often the sampling procedure results in a connection between $s^{(1)}$ and $s^{(2)}$. For example, if all relationships from sample nodes to other sample nodes, and no others, are recorded, then $s^{(2)} = s^{(1)} \times s^{(1)}$. In general, however, the nodes on which y-values are recorded and the node pairs on which x-values are recorded may be quite unrelated sets. In particular, the link-tracing procedures considered in this paper often lead to data on links from nodes in $s^{(1)}$ to nodes outside of $s^{(1)}$.

The sampling design is the procedure by which the sample is selected. This selection procedure may be controlled by the investigators, as is the case with a deliberately implemented probability sampling design, or may be beyond the control of the investigators and determined by the circumstances of the situation. If the probability of selecting the sample does not depend on node values y or link values x or parameters ψ involved in the graph model, we refer to the design as "conventional." For a conventional design the probability of selecting sample s can be written $p(s)$ or $p(s; \varphi)$, where φ denotes any unknown parameters involved in the design (but not the model), as in a Bernoulli sampling with unknown inclusion probability φ for each node. The sampling design may depend on one or more auxiliary variables that are known for the whole population, but that dependence will be left implicit in the notation $p(s)$. Conventional designs include the classical probability designs such as simple random, systematic, stratified, multistage, and unequal probability sampling, as well as model-based purposive and balanced designs based on auxiliary variables.

If the probability of selecting the sample depends on any y or x values, we refer to the design as "adaptive," since the selection procedure adapts to the realized configuration of node and link values in the population. In addition, the design can involve unknown parameters ψ. Thus, in general the sampling design in the graph setting has a selection probability that can be written $p(s \mid \mathbf{y}, \mathbf{x}; \psi)$ where **y** denotes the sequence of node values, **x** the matrix of arc indicator values, and ψ any parameters involved.

Likelihood-based inference, such as maximum likelihood estimation or prediction and Bayes methods, is simplified if the design can be ignored at the inference stage. The fact that the sampling design does not affect the value of a Bayes or likelihood-based estimator in survey sampling was noted by Godambe (1966) for designs that do not depend on any values of the variable of interest and by Basu (1969) for designs that do not depend on values of the variable of interest outside the sample. Scott and Smith (1973) showed that the design could become relevant to inference when the data lacked information about the labels of the units in the sample. Rubin (1976) gave exact conditions for a missing data mechanism – of which a sampling design can be viewed as an example – to be relevant in frequentist and likelihood-based inference. For likelihood-based methods such as maximum likelihood and Bayes methods, the design is "ignorable" if the design or mechanism does not depend on values of the variable of interest outside the sample or on any parameters in the distribution of those values. For frequency-based inference such as design- or model-unbiased estimation, however, the design is relevant if it depends on any values of the variable of interest, even in the sample. Scott (1977) showed that the design is relevant to Bayes estimation if auxiliary information used in the design is not available at the inference stage. Sugden and Smith (1984) gave general and detailed results on when the design is relevant in survey sampling situations. Thompson and Seber (1996) described adaptive designs in which the selection procedure deliberately takes advantage of observed values of the variable of interest, and discussed the relevance of the design in inference from a variety of design and model based perspectives. Similar issues of design and inference arise with adaptive experimental designs, such as medical experiments in which ethical considerations motivate adaptive treatment allocation to favor the more promising treatments as the study progresses (*cf.* Flournoy and Rosenberger 1995, Rosenberger 1996, Wei, Smythe, Lin and Park 1990). It is important to underscore that a design that is said to be "ignorable" for likelihood-based inference might not be ignorable for a frequentist-based inference, such as model-unbiased estimation, and that even though a design may be ignorable at the inference stage, in that for example the way an estimator is calculated does not depend on the design used, the design is still relevant a priori to the properties of the estimator.

The sample data $d = (s, \mathbf{y}_s, \mathbf{x}_s)$ are a function of the sample selected and of the graph values **y** and **x**. The likelihood can be written

$$L(\psi, d) = \sum p(s \mid \mathbf{y}, \mathbf{x}; \psi) f(\mathbf{y}, \mathbf{x}; \psi) \qquad (1)$$

where the sum is over outcomes $(\mathbf{y}, \mathbf{x})$ consistent with the data d. Since the y and x values for nodes and node pairs in the sample are fixed by the data, the sum is over all possible values of the unobserved variables $\mathbf{y}_{\bar{s}}$ and $\mathbf{x}_{\bar{s}}$ and it actually represents the marginal probability of the sample s selected and the associated observed variables $\mathbf{y}_s$ and $\mathbf{x}_s$.

Thus, in general the likelihood function depends on both the design and the model. The quantity $\sum_{\mathbf{y}_{\bar{s}}, \mathbf{x}_{\bar{s}}} f(\mathbf{y}, \mathbf{x}; \psi)$, based on the model only without consideration of the design, was termed the "face-value likelihood" by Dawid and Dickey (1977) because inference based on this function alone takes the data at face value without considering how the data were selected.

For any design in which the selection of the sample depends on graph y and x values only through those values $\mathbf{y}_s$ and $\mathbf{x}_s$ included in the data, the design probability can be moved out of the sum and forms a separate factor in the likelihood. If in addition the design and model parameters are distinct and not related, the likelihood can be written

$$L(\varphi, \psi, d) = p(s \mid \mathbf{y}_s, \mathbf{x}_s; \varphi) \sum_{\mathbf{y}_{\bar{s}}, \mathbf{x}_{\bar{s}}} f(\mathbf{y}, \mathbf{x}; \psi) \qquad (2)$$

where φ denotes the design parameters and ψ denotes the model parameters. The design then does not affect the value of estimators or predictors based on direct likelihood methods such as maximum likelihood or Bayes estimators. For any such "ignorable" design, the sum in the above likelihood, over all values of $\mathbf{y}$ and $\mathbf{x}$ leading to the given data value, is simply the marginal probability of the y and x values associated with the sample data. This marginal distribution depends on what sample was selected, but does not depend on how that sample was selected. For likelihood-based inference with a design ignorable in this sense, the face-value likelihood gives the correct inference.

3. SOME LINK-TRACING DESIGNS

A variety of link-tracing designs are described in this section. Each of these designs is ignorable in the likelihood sense provided the initial sample is selected by an ignorable procedure and provided the data include all the values involved in the selection procedure. Since for all the designs described in this section, the node-pair sample $s^{(2)}$ has a deterministic functional relationship to the node sample $s^{(1)}$, the superscript notation will be omitted and the final node sample $s^{(1)}$ will be denoted simply s.

The simple likelihood methods described in this paper apply to a wide range of ignorable link-tracing designs, including those described in this section. Further research is needed on methods for nonignorable designs, including those with nonignorable selection of the initial sample. Methods for dealing with nonsampling errors such as nonresponse and reporting errors with link-tracing designs are also in need of further development (*cf.*, Thompson 1997).

3.1 Single-Wave Design

In a single-wave link-tracing design an initial sample of nodes is selected by any ignorable design from the population of nodes in the graph. For each node in the sample, nodes adjacent from that node are added to the sample. The snowball procedure is assumed to stop after one wave. Thus, node v will be added if for some node u in the initial sample $x_{uv} = 1$.

Let s_0 denote the set of nodes in the initial sample and s_1 denote the added nodes not in the initial sample. The whole sample is $s = s_0 \cup s_1$.

The entire set of labels can be written as the union of three disjoint sets, $U = s_0 \cup s_1 \cup \bar{s}$. The values $\mathbf{y}$ associated with the nodes can be correspondingly ordered as a sequence $(\mathbf{y}_{s_0}, \mathbf{y}_{s_1}, \mathbf{y}_{\bar{s}})$, where $\mathbf{y}_a = (y_u: u \in a)$ is the subsequence of $\mathbf{y}$ restricted to indices in subset $a \subset U$. The adjacency matrix $\mathbf{x}$ is ordered correspondingly and partitioned into submatrices $\mathbf{x}_{s_0 s_0}, \mathbf{x}_{s_0 s_1}, \mathbf{x}_{s_0 \bar{s}}$ and so on, where $\mathbf{x}_{ab} = (x_{uv}: u \in a, v \in b)$. Ordering the adjacency matrix in this way facilitates the specification of factors in the likelihood.

With the design above, the probability of selecting sample s depends only on $\mathbf{x}_{s_0 U}$ and so can be written $p(s \mid \mathbf{x}_{s_0 U})$, where $\mathbf{x}_{s_0 U}$ can also be replaced by its column permutation $(\mathbf{x}_{s_0 s_0}, \mathbf{x}_{s_0 s_1}, \mathbf{x}_{s_0 \bar{s}})$. That is, the probability of selecting the final sample $s = s_0 \cup s_1$ depends on links from the initial sample to other units in the graph, both in s and in $\bar{s}$. The data consist of $(s, \mathbf{y}_s, \mathbf{x}_{s_0 U})$. Since the design does not depend on any x or y values outside the data or on model parameter values, the design is ignorable for likelihood-based inference.

3.2 Multi-Wave Samples

Consider a snowball sample with $k + 1$ waves after the initial sample. The sample will be denoted $s = s_0 \cup s_1$ with $s_0 = s_{00} \cup s_{01} \cup s_{02} \cup \ldots \cup s_{0k}$. An initial sample s_{00} is selected by any design that is ignorable in the likelihood sense. Links are followed and every node with an arc from any node in s_{00} and not already in the sample is added to form the first-wave sample s_{01}. That is, $s_{01} = \{v: x_{uv} = 1 \text{ for some } u \in s_{00}, v \notin s_{00}\}$. Then links are followed in s_{01} to give the second-wave sample $s_{02} = \{v: x_{uv} = 1 \text{ for some } u \in s_{01}, v \notin s_{00} \cup s_{01}\} = \{v: x_{uv} = 1 \text{ for some } u \in s_{00} \cup s_{01}, v \notin s_{00} \cup s_{01}\}$. Finally, the $(k+1)$-wave sample, denoted simply s_1, is added by following links from the k-th wave sample s_{0k}. That is $s_1 = \{v: x_{uv} = 1 \text{ for some } u \in s_0, v \notin s_0\}$. No links from s_1 are followed.

If $s_{0j} = \emptyset$ for any $j < k$ then sampling stops, so that the number of waves added is less than k if at some point there are no links leading out of the current sample to unsampled nodes.

The data consist of sets of node labels in the different waves of the sample and the ordered node pairs from s_0 to U, the sequence of node-values y_s for all nodes in the sample, and the link indicator variables $\mathbf{x}_{s_0 U}$ from s_0 to the set U of nodes in the graph. Thus the data consist of the

subgraph data for s_0, that is $(s_0, \mathbf{y}_{s_0}, \mathbf{x}_{s_0 s_0})$, together with the node values $\mathbf{y}_{s_1}$ for the nodes in the final-wave s_1, the link indicators $\mathbf{x}_{s_0 s_1}$ from s_0 to s_1, and the link indicators $\mathbf{x}_{s_0 \bar{s}}$ from the nodes in s_0 to the nodes not in the sample.

Since the design does not depend on any y or x values outside the data nor on any of the graph model parameters, the design is ignorable and the structure of the data is exactly the same with the $(k+1)$-wave snowball as with the 1-wave snowball design, and with the notation we have used the likelihood and estimation formulas are unchanged with the more general design.

3.3 Completed-Wave Designs

With a completed snowball sample, the procedure of adding waves is continued until no further links lead out of the sample. Then the number of completed waves K is a random variable and $s_{0,K+1} = s_1$ is the first empty set in the sequence $(s_{00}, s_{01}, \ldots)$. The data are $d = (s_0, \mathbf{y}_{s_0}, \mathbf{x}_{s_0 U})$ or equivalently $(s_0, \mathbf{y}_{s_0}, \mathbf{x}_{s_0 s_0}, \mathbf{x}_{s_0 \bar{s}_0})$. Inference can then proceed with the same likelihood and estimation formulas but with the simplication that the data contains no set s_1 for which $\mathbf{y}_{s_1}$ and $\mathbf{x}_{s_0, s_1}$ are known but from which links are unknown.

3.4 Link-Tracing Adaptive on Node Values

Consider a design in which the decision to follow the links from node u depends on the node value y_u. For example, in a study on injection drug use, the initial sample may contain both users ($y_u = 1$) and nonusers ($y_u = 0$). If the investigators choose to follow social links only from users, then the design depends adaptively on the node y-values as well as the links. Similarly, in a study of sexually transmitted diseases, investigators may be instructed to follow sexual or social links more frequently from infected respondents than from noninfected respondents. The design then can be written $p(s \mid \mathbf{y}_s, \mathbf{x}_{s_0 U})$, since the selection procedure depends on both node and link values. If the data contain all values on which the design depends, that is, $d = (s, \mathbf{y}_s, \mathbf{x}_{s_0 U})$, then the design is ignorable and maximum likelihood inference is simplified as described in the following sections.

3.5 Tracing Only a Subsample of Sample Links

The designs described above can be generalized to procedures in which only a sample of the links leading out from node u in s_0 are followed. Examples include the "random walk" design of Klovdahl (1989) and the generalization of snowball designs described in Snijders (1992). In the random walk design, an initial respondent is asked to give the names of several social contacts. One of these contacts is chosen at random to be interviewed and asked in turn to name several contacts, one of which is chosen at random, and so on. In practice, dead ends can occur when a respondent either reports no contacts or reports only contacts who are already in the sample. In such cases investigators either backtrack and try different leads from previous respondents or find a new initial respondent.

With these subsampling link-tracing designs, the procedure for selecting the sample, though complicated from a design-probability point of view , depends only on values in the sample and on links leading from the sample. We again assume that the initial sample is obtained by any ignorable procedure. Let $s_0 = s_{00} \cup s_{01} \cup s_{02} \cup \ldots \cup s_{0k}$ consist of all of the waves from which at least some links are followed. Thus, s_{01} consists of the nodes not previously included obtained by following a subsample of the links from nodes in the initial sample s_{00}, s_{02} consists of the nodes not previously included obtained by following a subsample of the links from nodes in $s_{00} \cup s_{01}$, and so on. No links are followed from the final wave s_1. Allowing for the possibility of dependence on node values, the design can be written $p(s \mid \mathbf{y}_s, \mathbf{x}_{s_0 U})$, so that with data $d = (s, \mathbf{y}_s, \mathbf{x}_{s_0 U})$, the design is ignorable for likelihood-based inference.

3.6 Data from Link-Tracing Designs

With any of the single or multi-wave link-tracing designs described above, it is of considerable practical importance what data are recorded. If the data include only the sample node labels, the y-values for nodes in the sample, and the arc indicators for pairs of units in the sample, that is, $d = (s, \mathbf{y}_s, \mathbf{x}_{ss})$, then the design is nonignorable and must be integrated into the likelihood, which can complicate analysis.

Consider also a study in which social links are used in the design , to find the sample, but only node characteristics (y-values), not relationships are recorded, so that the data are $d = (s, \mathbf{y}_s)$. Then the design is nonignorable.

If on the other hand the data from the link-tracing design include not only the linkages within the sample but the out-linkages (or lack thereof) from all but the last wave to the rest of the graph, that is, $d = (s, \mathbf{y}_s, \mathbf{x}_{s_0 U})$, then the design depends only on graph values in the data and so factors out of the likelihood.

4. A GRAPH MODEL WITH LINKS RELATED TO NODE VALUES

The likelihood-based approach described in section 2 with sample data from link-tracing designs of types described in section 3 will be illustrated using a class of graph models described in this section. This class of models builds on conditional independence between dyads as in the contact models of Frank (1979a) and Wellman, Frank, Espinoza, Lundquist and Wilson (1991). Conditional on the node values, independence is assumed between dyads, with the distribution of links between pairs of nodes depending on node value. Thus, unconditionally these models have dependence between dyads because of the dependence on the node values. In the models of Holland

and Leinhardt (1981), dyads are assumed to be independent but with distributions that depend on fixed node parameters. Wasserman (1980) also assumed independence of dyads in modeling the change in a graph over time as a stochastic process. Bayesian extensions and stochasstic blockmodels of Holland, Laskey, and Leinhardt (1983), Fienberg, Meyer, and Wasserman (1985), Wang and Wong (1987), and Frank (1988) provide generalizations to joint distributions with dependence between node values and graph links. Models by Frank and Harary (1982) for randomly colored graphs exhibit a similar structure. The Markov graph models of Frank and Strauss (1986) provide another approach to dependence among dyads but present difficulties for maximum likelihood estimation. Review of a variety of graph models is found in Wasserman and Faust (1994) and Frank (1997).

The maximum likelihood estimation and prediction methods of this paper apply equally to sample data with graph models other than the class of stochastic block models we have used. With other models, the same conditions for ignorability apply. We have chosen this class of models because it is rich enough to encompass important aspects of realism such as dependence between dyads and between dyads and node values, and it is simple enough to have explicit full-graph maximum likelihood estimators for comparison with the estimators based on samples. With other classes of models such as the Markov graph models, estimation even with full-graph data requires numerical methods.

For practical use of the model based approach it is important to have diagnostic tools for evaluations and comparisons between alternative models. For example, with the two-block model used here the conditional independence of dyads could be tested by counting pairs of dyads of different types within and between the blocks. Within each block there are three types of dyads and six types of pairs of dyads. Between the two blocks there are four types of dyads and ten types of pairs of dyads. A Pearson goodness-of-fit statistic between observed and expected counts of the 22 types of pairs of dyads within and between the blocks is asymptotically chi-square distributed with 12 degress of freedom under the conditional dyad independence assumption. Goodness-of-fit testing for graph models is discussed by Holland and Leinhardt (1981) and Frank and Strauss (1986), and this direction of research needs further development in particular in connection with sample data from link-tracing designs.

In the assumed model the node variables $Y_1, ..., Y_N$ are independent, identically distributed (i.i.d.) Bernoulli random variables with probabilities $P(Y_u = i) = \theta_i$, for $i = 0$, 1, with $\theta_0 + \theta_1 = 1$. Conditional on the node values $Y_1, ..., Y_N$, the dyads (X_{uv}, X_{vu}) are independent, for $1 \le u < v \le N$, with conditional distribution given by $P[(X_{uv}, X_{vu}) = (k, l) \mid Y_u = i, Y_v = j] = \lambda_{ijkl}$ for all combinations of $i = 0, 1, j = 0, 1, k = 0, 1$, and $l = 0, 1$. For all combinations of i and j, the sums over k and l are denoted $\lambda_{ij\bullet\bullet} = \sum_k \sum_l \lambda_{ijkl}$ and equal 1. In order to get graph probabilities not depending on node identities, the following symmetry requirements are needed: $\lambda_{1110} = \lambda_{1101}$, $\lambda_{1011} = \lambda_{0111}$, $\lambda_{1010} = \lambda_{0101}$, $\lambda_{1001} = \lambda_{0110}$, $\lambda_{0010} = \lambda_{0001}$, and $\lambda_{1000} = \lambda_{0100}$. The pattern of these restrictions is illustrated in Table 1.

Table 1

	(x_{uv}, x_{vu})			
(y_u, y_v)	(0,0)	(0,1)	(1,0)	(1,1)
(0,0)	●	●	●	●
(0,1)	●	●	●	●
(1,0)	●	●	●	●
(1,1)	●	●	●	●

With these restrictions, it is convenient to introduce the notation

$$\lambda_{ijkl} = \begin{cases} \gamma'_{i+j,k+l}, & \text{if } (ijkl) = (0110) \text{ or } (1001), \\ \gamma_{i+j,k+l}, & \text{otherwise} \end{cases}$$

where $\gamma_{00} + 2\gamma_{01} + \gamma_{02} = 1$, $\gamma_{10} + \gamma_{11} + \gamma'_{11} + \gamma_{12} = 1$, and $\gamma_{20} + 2\gamma_{2521} + \gamma_{22} = 1$. We can interpret γ'_{11} and γ_{11} as the probabilities of dyads with an arc from an unmarked to a marked node only and from a marked to an unmarked node only, respectively. Moreover, for $(ij) \neq (11)$, γ_{ij} is the probability of a dyad with j arcs on i marked and $2 - i$ unmarked nodes.

It will also be convenient to denote $\lambda_{ij1\bullet} = \sum_l \lambda_{ij1l} = \alpha_{ij}$ and $\lambda_{ij11} = \beta_{i+j}$ for $i = 0$, 1 and $j = 0$, 1. Here α_{ij} is the probability of an arc from a node of value i to a node of value j, and β_k is the probability of mutual arcs between k marked nodes.

Let N_i denote the total number of nodes with value i in the graph, for $i = 0$, 1, so that $N_0 + N_1 = N$. Let further M_{ijkl} denote the total number of dyads of type $(ijkl)$, that is, the total number of ordered node pairs (u, v), with $u < v$, such that $(Y_u, Y_v, X_{uv}, X_{vu}) = (ijkl)$.

The likelihood for the full graph under the model with parameters (θ, λ) is

$$L(\theta, \lambda; \mathbf{y}, \mathbf{x}) = \left(\prod_{i=0}^{1} \theta_i^{N_i} \right) \left(\prod_{i=0}^{1} \prod_{j=0}^{1} \prod_{k=0}^{1} \prod_{l=0}^{1} \lambda_{ijkl}^{M_{ijkl}} \right). \qquad (3)$$

In terms of the γs,

$$\prod_{i=0}^{1} \prod_{j=0}^{1} \prod_{k=0}^{1} \prod_{l=0}^{1} \lambda_{ijkl}^{M_{ijkl}} = \left(\prod_{i=0}^{2} \prod_{j=0}^{2} \gamma_{ij}^{R_{ij}} \right) (\gamma'^{R'_{11}}_{11})$$

where the Rs are dyad counts corresponding to the pattern in Table 1. That is, $R_{00} = M_{0000}$, $R_{01} = M_{0001} + M_{0010}$,

$R_{02} = M_{0011}$, $R_{10} = M_{0100} + M_{1000}$, $R_{11} = M_{0101} + M_{1010}$, $R'_{11} = M_{0110} + M_{1001}$, $R_{12} = M_{0111} + M_{1011}$, $R_{20} = M_{1100}$, $R_{21} = M_{1101} + M_{1110}$, $R_{22} = M_{1111}$. Note that R'_{11} (R_{11}) is the number of dyads with an arc from an unmarked (marked) to a marked (unmarked) node only. Also note that except for $(ij) = (11)$, R_{ij} is the number of dyads on i marked nodes with j arcs.

The maximum likelihood estimators with the whole graph as data are the proportions $\hat{\theta}_i = N_i / N$, $\hat{\gamma}_{ij} = R_{ij} / R_i$, and $\hat{\gamma}'_{11} = R'_{11} / R_1$, where $R_0 = N_0(N_0 - 1)/2$, $R_1 = N_0 N_1$, and $R_2 = N_1(N_1 - 1)/2$. In terms of the λs, this means $\hat{\lambda}_{ijkl} = R'_{11} / R_1$ if $(ijkl) = (0110)$ or (1001) and $\hat{\lambda}_{ijkl} = R_{i+j,k+l} / R_{i+j}$ otherwise.

5. INFERENCE FROM LINK-TRACING DESIGNS

5.1 Estimating Graph Model Parameters

Consider any of the link-tracing designs, for which an initial or multiwave sample is selected and links out from nodes in s_0 are followed to add the set s_1 of nodes not in s_0 that are adjacent after nodes in s_0. The data are $d = (s, \mathbf{y}_s, \mathbf{x}_{s_0 U})$, so that the design depends on y and x values only through those in the data and is thus ignorable.

With the graph model described in the previous section, the likelihood with the sample data given by equation (2) in section 2 is in this case

$$L(\theta, \lambda, d) = P(s \mid \mathbf{y}_s, \mathbf{x}_{s_0 U}) \sum \left(\prod_{u=1}^{N} \theta_{y_u} \right) \left(\prod_{u<v} \lambda_{y_u y_v x_{uv} x_{vu}} \right)$$

where the sum is over all values y_u and x_{uv} that are not fixed by the sample data.

Similar to the notation for population counts in the previous section, let $n_i(a)$ denote the number of nodes $u \in a$ with $y_u = i$ for arbitrary subsets $a \subset U$. Let $m_{ijkl}(a, b)$ be the count of pairs of nodes (u, v) such that $u \in a$, $v \in b$, $(y_u, y_v, x_{uv}, x_{vu}) = (ijkl)$, and $u < v$ if both u and v belong to $a \cap b$. An index replaced by a dot means summation over that index. For instance, according to the link-tracing designs described in section 3, only $m_{ijk\bullet}(s_0, s_1)$ is observed, not $m_{ijkl}(s_0, s_1)$.

With data from any of the link-tracing designs described in section 3, the likelihood function is

$$L(\theta, \lambda; d) = P(s \mid \mathbf{y}_s, \mathbf{x}_{s_0 U}) \left(\prod_i \theta_i^{n_i(s)} \right) \left(\prod_{ijkl} \lambda_{ijkl}^{m_{ijkl}(s_0, s_0)} \right) \times \left(\prod_{ijk} \lambda_{ijk\bullet}^{m_{ijk\bullet}(s_0, s_1)} \right) \prod_{v \in \bar{s}} \left[\sum_j \theta_j \prod_{ik} \lambda_{ijk\bullet}^{m_{i\bullet k\bullet}(s_0, v)} \right]. \quad (4)$$

For the link-tracing designs in which all links, rather than a subsample, from the initial sample are traced, all of the elements in the submatrix $\mathbf{x}_{s_0 \bar{s}}$ are zero and $m_{i\bullet 0\bullet(s_0, v)} = n_i(s_0)$ for $v \in \bar{s}$, which simplifies the likelihood function to

$$L(\theta, \lambda; d) = P(s \mid \mathbf{y}_s, \mathbf{x}_{s_0 U}) \left(\prod_i \theta_i^{n_i(s)} \right) \left(\prod_{ijkl} \lambda_{ijkl}^{m_{ijkl}(s_0, s_0)} \right) \times \left(\prod_{ijk} \lambda_{ijk\bullet}^{m_{ijk\bullet}(s_0, s_1)} \right) \left[\sum_j \theta_j \prod_i \lambda_{ij0\bullet}^{n_i(s_0)} \right]^{n(\bar{s})}. \quad (5)$$

The factor $\prod \theta_i^{n_i(s)}$ gives the probability of the observed node values in the sample. The factor $\prod \lambda_{ijkl}^{m_{ijkl}(s_0, s_0)}$ gives the probability of the observed dyad types within $s_0 \times s_0$ given the node values. The factor $\prod \lambda_{ijk\bullet}^{m_{ijk\bullet}(s_0, s_1)}$ gives the probability of the observed dyad types in $s_0 \times s_1$. Since x_{uv} but not x_{vu} is observed, for $u \in s_0$ and $v \in s_1$, the marginal probability that $x_{uv} = k$ given $y_u = i$ and $y_v = j$ is $\lambda_{ijk\bullet}$.

The final factor of (5), with square brackets, gives the probability that there are no arcs from the initial sample to $\bar{s}$. For a node v of the $n(\bar{s})$ nodes outside the sample, θ_j is the probability that $y_v = j$. From any of the $n_i(s_0)$ sample nodes $u \in s_0$ with $y_u = i$, the conditional probability of no link to v, that is, that $x_{uv} = 0$, $\lambda_{ij0\bullet}$.

More formally, the bracketed term can be obtained by conditioning on the number $n_j(\bar{s})$ of nodes of type j in $\bar{s}$. Conditional on $n_j(\bar{s})$, the probability that all the link indicators from s_0 to $\bar{s}$ are zero is obtained as follows. From the $n_i(s_0)$ nodes of type i in s_0 to the $n_j(\bar{s})$ nodes of type j in $\bar{s}$, the probability that all links are zero is $\lambda_{ij0\bullet}^{n_i(s_0) n_j(\bar{s})}$. Using the binomial distribution of $n_1(\bar{s})$ with the law of total probability, the probability that all the links from s_0 to $\bar{s}$ are zero, given y_s, is

$$\sum_{n_1(\bar{s})=0}^{n(\bar{s})} \binom{n(\bar{s})}{n_1(\bar{s})} \left(\prod_j \theta_j^{n_j(\bar{s})} \right) \left(\prod_i \prod_j \lambda_{ij0\bullet}^{n_i(s_0) n_j(\bar{s})} \right) = \left[\sum_j \theta_j \prod_i \lambda_{ij0\bullet}^{n_i(s_0)} \right]^{n(\bar{s})}. \quad (6)$$

With the completed-wave design, the above likelihood expressions are simplified since the terms $m_{ijk\bullet(s_0, s_1)}$ are all zero, so that the factors involving these terms are all equal to one. We also note that $\lambda_{ij0\bullet} = 1 - \alpha_{ij}$ and $\lambda_{ij1\bullet} = \alpha_{ij}$ can be substituted to simplify the likelihood.

5.1.1 Estimative Likelihood Equations

The maximum likelihood estimators for the parameters θ_1, α_{ij}, and β_k are obtained as the common solutions to the equations

$$\frac{d \log L}{d\theta_1} = \frac{d \log L}{d\alpha_{ij}} = \frac{d \log L}{d\beta_k} = 0 \quad (7)$$

for $i = 0, 1$, $j = 0, 1$, $k = 0, 2$. Differentiating the logarithm of the likelihood (5) with respect to θ_1 and setting equal to zero gives

$$\frac{d \log L}{d\theta_1} = \frac{\partial \log L}{\partial \theta_1} - \frac{\partial \log L}{\partial \theta_0} = 0$$

where the partial derivatives are given by

$$\frac{\partial \log L}{\partial \theta_k} = \frac{n_k(s)}{\theta_k} + n(\bar{s}) \frac{\prod_i \lambda_{ik0\bullet}^{n_i(s_0)}}{\sum_j \theta_j \prod_i \lambda_{ij0\bullet}^{n_i(s_0)}}$$

for $k = 0, 1$.

Moreover,

$$\frac{d \log L}{d\alpha_{ij}} = \frac{\partial \log L}{\partial \lambda_{ij10}} + \frac{\partial \log L}{\partial \lambda_{ji01}} - \frac{\partial \log L}{\partial \lambda_{ij}00} - \frac{\partial \log L}{\partial \lambda_{ji00}} \quad (8)$$

and

$$\frac{d \log L}{d\beta_k} = \sum_{\substack{i,j \\ i+j=k}} \left(\frac{\partial \log L}{\partial \lambda_{ij00}} + \frac{\partial \log L}{\partial \lambda_{ij11}} - \frac{\partial \log L}{\partial \lambda_{ij01}} - \frac{\partial \log L}{\partial \lambda_{ij10}} \right) \quad (9)$$

where the partial derivatives are given by

$$\frac{\partial \log L}{\partial \lambda_{ijkl}} = \frac{m_{ijkl}(s_0, s_0)}{\lambda_{ijkl}} + \frac{m_{ijk\bullet}(s_0, s_1)}{\lambda_{ijk\bullet}} + (1 - k) n(\bar{s}) \frac{\theta_j n_i(s_0) \lambda_{ij0\bullet}^{n_i(s_0)-1}}{\sum_j \theta_j \prod_i \lambda_{ij0\bullet}^{n_i(s_0)}}.$$

It is convenient to write the likelihood equation for θ_1 as

$$\frac{n_1(s)}{\theta_1} - \frac{n_0(s)}{\theta_0} + \frac{n(\bar{s})(\rho - 1)}{\theta_1 \rho + \theta_0} = 0 \quad (10)$$

where

$$\rho = \prod_{i=0}^{1} \left(\frac{\lambda_{i10\bullet}}{\lambda_{i00\bullet}} \right)^{n_i(s_0)} = \prod_{i=0}^{1} \left(\frac{1 - \alpha_{i1}}{1 - \alpha_{i0}} \right)^{n_i(s_0)}.$$

Note that $\rho = \rho_0^{n_0(s_0)} \rho_1^{n_1(s_0)}$, where $\rho_i = (1 - \alpha_{i1})/(1 - \alpha_{i0})$ is the ratio between the probabilities of no arc from an i-node to a positive and a zero node, respectively.

An interpretation of the influence of the graph structure on estimation of θ_1 is provided by considering the graph parameters α – and hence ρ – as fixed. Denote the sample proportion of positive nodes by $\hat{\theta}_c = n_1(s)/n(s)$. This is the conventional or naive estimator of θ_1, using the sample proportion of positive nodes. If $\rho = 1$, then the maximum likelihood estimator $\hat{\theta}_1$ would be $\hat{\theta}_c$. If $\rho < 1$, then the maximum likelihood estimator $\hat{\theta}_1$ would be less than $\hat{\theta}_c$, and if $\rho > 1$, $\hat{\theta}_1 > \hat{\theta}_c$. In particular, $\alpha_{i1} = \alpha_{i0}$ for $i = 0, 1$ implies $\rho = 1$ and the maximum likelihood estimator is $\hat{\theta}_1 = \hat{\theta}_c$.

Consider for instance the case in which for any given value of y_u, a link from node u to node v is more likely when $y_v = 1$ than when $y_v = 0$, so that $\alpha_{i1} > \alpha_{i0}$, for $i = 0, 1$. Then $(1 - \alpha_{i1})/(1 - \alpha_{i0}) < 1$, for $i = 0, 1$, so that $\rho < 1$ and the maximum likelihood estimator $\hat{\theta}_1$ is less than the conventional estimator $\hat{\theta}_c$. One could say that the link-tracing design is leading investigators to an unrepresentatively high proportion of positive nodes, and the maximum likelihood estimator is adjusting for this.

In specific cases some of the λ_{ijkl} might be set to zero and the likelihood equations have to be modified accordingly. Some specific cases will now be illustrated.

5.1.2 A Symmetric Model

Symmetric models have $\lambda_{ijkl} = 0$ for $k \neq l$ so that arcs are always mutual or, equivalently, they can be considered as undirected edges.

The full symmetric model has parameters $\lambda_{ijkk} = \lambda_{jikk}$ for $i, j, k = 0, 1$, with $\lambda_{ij00} + \lambda_{ij11} = 1$. Here $\lambda_{ij11} = \beta_{i+j} = \alpha_{ij} = \alpha_{ji}$ and

$$\rho = \prod_{i=0}^{1} \left(\frac{1 - \beta_{i+1}}{1 - \beta_i} \right)^{n_i(s_0)}.$$

Letting $m_{ijkl}(s_0, s) = r_{i+j,k+l}$, we obtain the maximum likelihood estimators as the solutions to the equations

$$\frac{n_1(s)}{\theta_1} - \frac{n_0(s)}{\theta_0} - \frac{n(\bar{s})(1 - \rho)}{\theta_0 + \rho\theta_1} = 0 \quad (11)$$

$$\frac{r_{02}}{\beta_0} - \frac{r_{00}}{1 - \beta_0} - \frac{n(\bar{s}) n_0(s_0) \theta_0}{(1 - \beta_0)(\theta_0 + \rho\theta_1)} = 0 \quad (12)$$

$$\frac{r_{12}}{\beta_1} - \frac{r_{10}}{1 - \beta_1} - \frac{n(\bar{s})[n_1(s_0)\theta_0 + n_0(s_0)\rho\theta_1]}{(1 - \beta_1)(\theta_0 + \rho\theta_1)} = 0 \quad (13)$$

$$\frac{r_{22}}{\beta_2} - \frac{r_{20}}{1 - \beta_2} - \frac{n(\bar{s}) n_1(s_0) \rho\theta_1}{(1 - \beta_2)(\theta_0 + \rho\theta_1)} = 0. \quad (14)$$

If the symmetric model is further simplified by assuming $\beta_0 = \beta_1 = 0$, there are only the two parameters θ_1 and β_2, and the equations to be solved are

$$\theta_1 \beta_2 = r_{22}/N n_1(s_0)$$

and

$$\frac{N - n_1(s)/\theta_1}{N - n_0(s)/\theta_0} = (1 - \beta_2)^{n_1(s_0)}.$$

For instance suppose the value $y_u = 1$ indicates injection drug use and $x_{uv} = 1$ indicates u and v are injection partners, so that links are only possible between users and tracing these links can only add users to the sample. As an illustration, consider a population of size N = 10, 000 with statistics $n_1(s_0) = 7$, $n_0(s_0) = 43$, $n_1(s) = 47$, and $r_{22} = 42$. The likelihood equations are $\theta_1 \beta_2 = 0.0006$ and $(10000 - 47/\theta_1)/(10000 - 43/\theta_0) = (1 - \beta_2)^7$, leading to the maximum likelihood estimators $\hat{\theta}_1 = 0.12$ and $\hat{\beta}_2 = 0.005$. The naive estimator for θ_1 in this case would be the sample proportion 47/90 = 0.52 and the naive estimator for β_2 would be

$$42 \Big/ \binom{47}{2} = 0.039,$$

the proportion of links between users in the sample out of the number possible.

5.1.3 An Asymmetric Model

A specific asymmetric model has $\lambda_{ijkl} = \lambda_{ijk\bullet}\lambda_{ij\bullet l} = \lambda_{ijk\bullet}\lambda_{jil\bullet}$, so that all arcs are independent. Now $\beta_{i+j} = \alpha_{ij}\alpha_{ji}$ and we obtain the maximum likelihood estimators as the solutions to the equations

$$\rho = \frac{N - n_1(s)/\theta_1}{N - n_0(s)/\theta_0}$$

and

$$\frac{\alpha_{ij}}{1 - \alpha_{ij}} = \frac{m_{ij1}}{m_{ij0} + n_i(s_o)\rho^j\theta_j(N - n_0(s)/\theta_0)}$$

for $i = 0, 1\ j = 0, 1$, where $m_{ijk} = m_{ijk\bullet}(s_0, s)$.

In particular, if we specify this asymmetric model by $\alpha_{ij} = ij\alpha$, so that arcs are possible with probability α only between marked nodes, then the equations to be solved are

$$\frac{N - n_1(s)/\theta_1}{N - n_0(s)/\theta_0} = (1 - \alpha)^{n_1(s_0)}$$

and

$$\frac{\alpha}{1 - \alpha} = \frac{m_{111}}{m_{110} + [N\theta_1 - n_1(s)]n_1(s_0)}.$$

Again, iterative methods are appropriate.

5.2 Predictive Likelihood for the Total of the Unobserved Node Values

For predicting the value of the unobserved random variable $n_1(\bar{s})$ from the observed data, the relevant likelihood is

$$L[\theta, \lambda; d, n_1(\bar{s})] = p(s \mid \mathbf{y}_s, \mathbf{x}_{s_0U})$$

$$\times \left(\prod_i \theta_i^{n_i(s) + n_i(\bar{s})}\right)\binom{n(\bar{s})}{N_1(\bar{s})}\left(\prod_{ijkl}\lambda_{ijkl}^{m_{ijkl}(s_0, s_0)}\right)$$

$$\times \left(P\prod_{ijk}\lambda_{ijk\bullet}^{m_{ijk\bullet}(s_0, s_1)}\right)\left(\prod_{ij}\lambda_{ij0\bullet}^{n_i(s_0)n_j(\bar{s})}\right). \quad (15)$$

Use of the term "prediction" implies only that the object of inference is a random variable rather than a fixed, unknown parameter, and does not necessarily imply forecasting in time.

The estimative likelihood for $n_1(\bar{s})$ is obtained from (15) by substituting the estimates $\hat{\theta}$ and $\hat{\lambda}$ that maximize the (marginal) likelihood (5). The value of $n_1(\bar{s})$ maximizing the estimative likelihood would be the estimative maximum likelihood predictor of $n_1(\bar{s})$. While estimative likelihood methods tend to produce reasonable point predictions in many cases, they are less useful as a basis for prediction intervals, since the estimates of the parameters are in essence treated as the true values (*cf.*, Bjørnstad 1990, 1996, Lejeune and Faulkenberry 1982). For this reason, we emphasize the use of the profile predictive likelihood.

Rather than substituting fixed estimators of the parameters into (15) and maximizing this estimative likelihood with respect to $n_1(\bar{s})$, the likelihood (15) is now simultaneously maximized with respect to both parameters and $n_1(\bar{s})$. This means that for each value of $n_1(\bar{s})$ there are parameter values $\tilde{\theta}_i[n_1(\bar{s})]$ and $\tilde{\lambda}_{ijkl}[(n_1(\bar{s})]$ which maximize (15) with respect to θ and λ. Substitution of these values into (15) defines the profile likelihood $L_p[n_1(\bar{s}); d]$ for $n_1(\bar{s})$. The value of $n_1(\bar{s})$ maximizing the profile likelihood is the profile maximum likelihood predictor of $n_1(\bar{s})$.

For any given value of $n_1(\bar{s})$, the likelihood is maximized where the derivatives with respect to the remaining parameters equal zero. The maximizing values of θ_i are straightforward and are given by

$$\tilde{\theta}_i = \frac{n_i(s) + n_i(\bar{s})}{N}. \quad (16)$$

For the remaining parameters we use $d \log L/d\alpha_{ij}$ and $d \log L/d\beta_k$ from (8) and (9), with the partial derivatives now given by

$$\frac{\partial \log L}{\partial \lambda_{ijkl}} = \frac{m_{ijkl}(s_0, s_0)}{\lambda_{ijkl}} + \frac{m_{ijk\bullet}(s_0, s_1)}{\lambda_{ijk\bullet}} + (1 - k)\frac{n_i(s_0)n_j(\bar{s})}{\lambda_{ij0\bullet}} \quad (17)$$

Note that the $n_j(\bar{s})$ for j = 0, 1 are contained in (15) only in the factors

$$\binom{n(\bar{s})}{n_1(\bar{s})}\prod_j \Lambda_j^{n_j(\bar{s})}$$

where $\Lambda_j = \theta_j\Pi_i\lambda_{ij0\bullet}^{n_i(s_0)}$. Since L is proportional to a binomial probability with parameters $n(\bar{s})$ and $\Lambda_1/(\Lambda_0 + \Lambda_1)$, it follows that the maximum of L over $n_1(\bar{s})$ is obtained for $n_1(\bar{s})$ equal to the integer closest to

$$\frac{n(\bar{s})\Lambda_1}{\Lambda_0 + \Lambda_1} + \frac{\Lambda_1 - \Lambda_0}{2(\Lambda_0 + \Lambda_1)}$$

or either of the integers closest to this number if there are two of them. In fact (see, for instance, Feller 1957, p.140), the mode of a binomial distribution with parameters (n, p) is the integer in the interval $[(n + 1)p - 1, (n + 1)p]$ or either of the endpoints if they are integers. Thus, the mode is the integer or the integers that are closest to the interval midpoint $(n + 1)p - (1/2) = np + (p - q)/2$, where $q = 1 - p$.

If initial values of the parameter estimators are obtained from the solution of (7) and substituted into the Λ_j, then a predicted value $n_1(\bar{s})$ is given as above. If this predicted value is inserted into (16) and (17), then new estimates of the parameters are obtained that can be substituted into the Λ_j to find a new predicted value of $n_1(\bar{s})$, continuing until the

values converge to the solution minimizing (15). Alternatively, the solution can be found by direct computation of the likelihood (15) for different values of $n_1(\bar{s})$, substituting the solutions obtained from (16) and (17) for the parameter values.

5.2.1 Example: Symmetric Model

The predictive likelihood equation (15) for the symmetric model is

$$L[\theta,\beta;d,n_1(\bar{s})] = p(s \mid \mathbf{y}_s, \mathbf{x}_{s_0 U})\left(\prod_i \theta_i^{n_i(s)+n_i(\bar{s})}\right)\binom{n(\bar{s})}{n_1(\bar{s})} \times\left(\prod_{i,j}\beta_{i+j}^{m_{ij11}(s_0,s)}(1-\beta_{i+j})^{m_{ij00}(s_0,s)+n_i(s_0)n_j(\bar{s})}\right). \quad (18)$$

Let $r_{kl} = r_{kl}(s_0, s)$ denote the count of node pairs in $s_0 \times s$ with total node value k and total number of links l. With the symmetric model, l can take only the values 0, indicating no link between the nodes, or 2, indicating a symmetric link. In particular, $r_{02} = m_{0011}(s_0, s)$, $r_{12} = m_{0111}(s_0, s) + m_{1011}(s_0, s)$, and $r_{22} = m_{1111}(s_0, s)$ denote the sample counts of links between nodes of total value k, for $k = 0, 1, 2$, respectively. With this notation the last factor in (18) can be written

$$\prod_{k=0}^{2}\beta_k^{r_{k2}}(1-\beta_k)^{r_{k0}+\sum_{\substack{i,j\\ i+j=k}} n_i(s_0)n_j(\bar{s})}.$$

Denote by $c_k = c_k[n_1(\bar{s})]$ the number of possible node pairs in $s_0 \times U$ having total value k, so that

$$c_k = r_{k\bullet} + \sum_{\substack{i,j\\ i+j=k}} n_i(s_0)n_j(\bar{s}) = \sum_{\substack{i,j\\ i+j=k}} n_i(s_0)[n_j(s) + n_j(\bar{s})].$$

For any given value of $n_1(\bar{s})$, the likelihood is maximized by $\tilde{\theta}_i = [n_i(s) + n_i(\bar{s})]/N$ for $i = 0, 1$ and $\tilde{\beta}_k = r_{k2}/c_k$ for $k = 0, 1, 2$. Note that $\tilde{\theta}$ and the $\tilde{\beta}_k$ are functions of the unobserved variable $n_1(\bar{s})$.

The profile predictive likelihood function for $n_1(\bar{s})$ is obtained by substituting the maximizing values $\tilde{\theta}$ and $\tilde{\beta}_k$ for the parameters in (18), giving

$$L_p[n_1(\bar{s}); d] = p(s \mid \mathbf{y}_s, \mathbf{x}_{s_0 U})\left(\prod_i\left(\frac{n_i(s)+n_i(\bar{s})}{N}\right)^{n_i(s)+n_i(\bar{s})}\right) \times\binom{n(\bar{s})}{n_1(\bar{s})}\left(\prod_k\left(\frac{r_{k2}}{c_k}\right)^{r_{k2}}\left(1-\frac{r_{k2}}{c_k}\right)^{c_k-r_{k2}}\right)$$

which is a function of $n_1(\bar{s})$ alone. The maximum profile likelihood predictor of $n_1(\bar{s})$, easily obtained by straightforward computation, is an integer between 0 and $n(\bar{s})$ giving the largest value of (19).

5.3 On Assessing Accuracy of Estimates

For confidence intervals and other forms of inference, the inverse of the observed Fisher information $\mathbf{I}(\hat{\phi})$ is suggested, where $\hat{\phi}$ is the vector of parameter maximum likelihood estimates and $\mathbf{I}$ is the matrix of negated second derivatives of the log likelihood function evaluated at those estimated values. The use of the observed, as opposed to expected, Fisher information to assess the accuracy of an estimate is described in Efron and Hinkley (1978). More recently, Lindsay and Li (1997) argue that the observed information gives a better assessment of the realized, as opposed to expected, error of the estimate. In developing large-sample approximations to the properties of the estimators of θ and λ it is important to make appropriate assumptions about how λ depends on N so that the graph model and the sample do not degenerate. See for instance the asymptotic results for some simple graph models given by Palmer (1985).

As with the calculation of the maximum likelihood estimates themselves, the calculation of the observed information matrix is not affected by the link-tracing sampling design, since the design is ignorable for likelihood based on inference. This is in contrast to the expected Fisher information, the value of which is affected by the design in addition to the graph model, unless the design is a conventional one not depending on any $\mathbf{y}$ or $\mathbf{x}$ values.

For a $(1 - \varepsilon)$-level prediction interval for a random variable such as $n_1(\bar{s})$, one method would be to use a central region having mass $(1 - \varepsilon)$ of the normalized profile likelihood function for $n_1(\bar{s})$ (*cf.*, Bjørnstad 1990, 1996). For the symmetric model, the $(1 - \varepsilon)$ prediction interval for $n_1(\bar{s})$, is readily obtained by computing (19) for $n_1(\bar{s}) = 0, 1, 2, \ldots$, until the computed values become negligible, normalizing by dividing by the cumulative total $\sum_{n_1(\bar{s})=0}^{n(\bar{s})} L_p$ and using the $\varepsilon/2$ and $1 - \varepsilon/2$ quantiles as the interval endpoints.

ACKNOWLEDGEMENTS

Support for this research was provided by the National Science Foundation (DMS-9626102), the National Institutes of Health, National Institute on Drug Abuse (RO1 DA09872), and the Swedish Council for Research in the Humanities and the Social Sciences (HSFR F 0750/96).

REFERENCES

BASU, D. (1969). Role of the sufficiency and likelihood principles in sample survey theory. *Sankhyā* A 31, 441-454.

BIRNBAUM, Z.W., and SIRKEN, M.G. (1965). Design of sample surveys to estimate the prevalence of rare diseases: Three unbiased estimates. *Vital and Health Statistics*, 2, 11. Washington: Government Printing Office.

BJØRNSTAD, J.F. (1990). Predictive likelihood: A review. *Statistical Science*, 5, 242-265.

BJØRNSTAD, J.F. (1996). On the generalization of the likelihood function and the likelihood principle. *Journal of the American Statistical Association*, 91, 791-806.

DAWID, A.P., and DICKEY, J.M. (1977). Likelihood and Bayesian inference from selectively reported data. *Journal of the American Statistical Association*, 72, 845-850.

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.

ERICKSON, B. (1979). Some problems of inference from chain data. *Sociological Methodology*, 10, 276-302.

FIENBERG, S.E., MEYER, M.M., and WASSERMAN, S.S. (1985). Statistical analysis of multiple sociometric relations. *Journal of the American Statistical Association*, 80, 51-67.

FLOURNOY, N., and ROSENBERGER, W.F., Eds. (1995). *Adaptive Designs*. Hayward, CA: Institute of Mathematical Statistics.

FRANK, O. (1971). *Statistical Inference in Graphs*. Stockholm: Försvarets forskningsanstalt.

FRANK, O. (1977a). Survey sampling in graphs. *Journal of Statistical Planning and Inference*, 1, 235-264.

FRANK, O. (1977b). Estimation of graph totals. *Scandinavian Journal of Statistics*, 4, 81-89.

FRANK, O. (1978a). Estimating the number of connected components in a graph by using a sampled subgraph. *Scandinavian Journal of Statistics*, 5, 177-188.

FRANK, O. (1978b). Sampling and estimation in large social networks. *Social Networks*, 1, 91-101.

FRANK, O. (1979a). Estimation of population totals by use of snowball samples. In *Perspectives on Social Network Research*, (Eds., P.W. Holland and S. Leinhardt). New York: Academic Press, 319-347.

FRANK, O. (1979b). Moment properties of subgraph counts in stochastic graphs. *Annals of the New York Academy of Sciencies*, 319, 207-218.

FRANK, O. (1981). A survey of statistical methods for graph analysis. *Sociological Methodology*, 110-155.

FRANK, O. (1988). Random sampling and social networks: a survey of various approaches. *Mathmatiques, Informatique et Sciences humaines*, 26, 19-33.

FRANK, O. (1997). Composition and structure of social networks. *Mathmatiques, Informatique et Sciences humaines*, 35, 11-23.

FRANK, O., and HARARY, F. (1982). Cluster inference by using transitivity indices in empirical graphs. *Journal of the American Statistical Association*, 77, 835-840.

FRANK, O., and SNIJDERS, T. (1994). Estimating the size of hidden populations using snowball sampling. *Journal of Official Statistics*, 10, 53-67.

FRANK, O., and STRAUSS, D. (1986). Markov graphs. *Journal of the American Statistical Association*, 81, 832-842.

FRIEDMAN, S.R., NEAIGUS, A., JOSE, B., CURTIS, R., GOLDSTEIN, M., ILDEFONSO, G., ROTHENBERG, R.B., and DES JARLAIS, D.C. (1997). Sociometric risk networks and HIV risk. *American Journal of Public Health*. In press.

GODAMBE, V.P. (1966). A new approach to sampling from finite populations. I. *Journal of the Royal Statistical Society* B, 28, 310-319.

GOODMAN, L.A. (1961). Snowball sampling. *Annals of Mathematical Statistics*, 32, 148-170.

GRANOVETTER, M. (1976). Network sampling: some first steps. *Americal Journal of Sociology*, 81, 1287-1303.

HOLLAND, P.W., LASKEY, K.B., and LEINHARDT, S. (1983). Stochastic block-models: First steps. *Social Networks*, 5, 109-137.

HOLLAND, P.W., and LEINHARDT, S. (1981). An exponential family of probability distributions for directed graphs (with discussion). *Journal of the American Statistical Association*, 76, 33-65.

JANSSON, I. (1997). On statistical modeling of social networks. Ph.D. Thesis, Stockholm University.

KALTON, G., and ANDERSON, D.W. (1986). Sampling rare populations. *Journal of the Royal Statistical Society* A, 149, 65-82.

KARLBERG, M. (1997). Triad count estimation and transitivity testing in graphs and digraphs. Ph.D. Thesis, Stockholm University.

KLOVDAHL, A.S. (1989). Urban social networks: Some methodological problems and possibilities. In, *The Small World*, (Ed. M. Kochen). Norwood, NJ: Ablex Publishing, 176-210.

LEJEUNE, M., and FAULKENBERRY, G.D. (1982). A simple predictive density function. *Journal of the American Statistical Association*, 77, 654-657.

LEVY, P.S. (1977). Optimum allocation in stratified random network sampling for estimating the prevalence of attributes in rare populations. *Journal of the American Statistical Association*, 72, 758-763.

LEVY, P.S., and LEMESHOW, S. (1991). *Sampling of Populations; Methods and Applications*. New York: Wiley.

LINDSAY, B.G., and LI., B. (1997). On second-order optimality of the observed Fisher information. *Annals of Statistics*, 25, 2172-2199.

MORGAN, D.L., and RYTINA, S. (1977). Comment on "Network sampling: some first steps" by Mark Granovetter. *American Journal of Sociology*, 83, 722-727.

NEAIGUS, A., FRIEDMAN, S.R., GOLDSTEIN, M.F., ILDEFONSEO, G., CURTIS, R., and JOSE, B. (1995). Using dyadic data for a network analysis of HIV infection and risk behaviors among injection drug users. In (Eds., R.H. Needle, S.G. Genser, and R.T. Trotter II) *Social Networks, Drug Abuse, and HIV Transmission*. NIDA Research Monograph 151. Rockville, MD: National Institute of Drug Abuse, 20-37.

NEAIGUS, A., FRIEDMAN, S.R., JOSE, B., GOLDSTEIN, M.F., CURTIS, R., ILDEFONSO, G., and DES JARLAIS, D.C. (1996). High-risk personal networks and syringe sharing as risk factors for HIV infection among new drug injectors. *Journal of Acquired Immune Deficiency Syndromes and Human Retrovirology*, 11, 499-509.

PALMER, E.M. (1985). *Graphical Evolution*. New York: Wiley.

ROBINS, G.L. (1998). Personal attributes in inter-personal contexts: statistical models for individual characteristics and social relationships. Ph.D. Thesis, University of Melbourne.

ROSENBERGER, W.F. (1996). New directions in adaptive designs. *Statistical Science*, 11, 137-149.

ROTHENBERG, R.B., WOODHOUSE, D.E., POTTERAT, J.J., MUTH, S.Q., DARROW, W.W., and KLOVDAHL, A.S. (1995). Social networks in disease transmission: The Colorado Springs study. In (Eds., R.H. Needle, S.G. Genser, and R.T. Trotter II), *Social Networks, Drug Abuse, and HIV Transmission*. NIDA Research Monograph 151. Rockville, MD: National Institute of Drug Abuse, 3-19.

RUBIN, D.B. (1976). Inference and missing data. *Biometrika*, 63, 581-592.

SCOTT, A.J. (1977). On the problem of randomization in survey sampling. *Sankhyā* C, 39, 1-9.

SCOTT, A.J., and SMITH, T.M.F. (1973). Survey design, symmetry, and posterior distributions. *Journal of the Royal Statistical Society* B, 35, 57-60.

SIRKEN, M.G. (1970). Household surveys with multiplicity. *Journal of the American Statistical Association*, 63, 257-266.

SIRKEN, M.G. (1972a). Stratified sample surveys with multiplicity. *Journal of the American Statistical Association*, 67, 224-227.

SIRKEN, M.G. (1972b). Variance components of multiplicity estimators. *Biometrics*, 28, 869-873.

SIRKEN, M.G., and LEVY, P.S. (1974). Multiplicity estimation of proportions based on ratios of random variables. *Journal of the American Statistical Association*, 69, 68-73.

SNIJDERS, T.A.B. (1992). Estimation on the basis of snowball samples: how to weight. *Bulletin de Methodologie Sociologique*, 36, 59-70.

SPREEN, M. (1992). Rare populations, hidden populations, and link-tracing designs; what and why? *Bulletin de Methodologie Sociologique*, 36, 34-58.

SPREEN, M. (1998). Sampling personal network structures: statistical inference in ego-graphs. Ph.D. Thesis, University of Groningen.

SPREEN, M., and ZWAAGSTRA, R. (1994). Personal network sampling, outdegree analysis and multilevel analysis: introducing the network concept in studies of hidden populations. *International Sociology*, 9, 475-491.

SUDMAN, S., SIRKEN, M.G., and COWAN, C.D. (1988). Sampling rare and elusive populations. *Science*, 240, 991-996.

SUGDEN, R.A., and SMITH, T.M.F. (1984). Ignorable and informative designs in survey sampling inference. *Biometrika*, 71, 495-506.

THOMPSON, S.K. (1990). Adaptive cluster sampling. *Journal of the American Statistical Association*, 85, 1050-1059.

THOMPSON, S.K. (1997). Adaptive sampling in behavioral surveys. In (Eds., L. Harrison, and A. Hughes), *The Validity of Self-Reported Drug Use: Improving the Accuracy of Survey Estimates*. NIDA Research Monograph 167, Rockville, MD: National Institute of Drug Abuse, 296-319.

THOMPSON, S.K., and SEBER, G.A.F. (1996). *Adaptive Sampling*. New York: Wiley.

van METER, K.M. (1990). Methodological and design issues: techniques for assessing the representatives of snowball samples. In (Ed., E.Y. Lambert), *The Collection and Interpretation of Data from Hidden Populations*. NIDA Monograph 98. Rockville, MD: National Institute on Drug Abuse, 31-43.

WANG, Y.J., and WONG, G.Y. (1987). Stochastic blockmodels for directed graphs. *Journal of the American Statistical Association*, 82, 8-19.

WASSERMAN, S. (1980). Analyzing social networks as stochastic processes. *Journal of the American Statistical Association*, 75, 280-294.

WASSERMAN, S., and FAUST, K. (1994). *Social Network Analysis: Methods and Applications*. New York: Cambridge University Press.

WATTERS, J.K., and BIERNACKI, P. (1989). Targeted sampling: Options for the study of hidden populations. *Social Problems*, 36, 416-430.

WEI, L.J., SMYTHE, R.T., LIN, D.Y., and PARK, T.S. (1990). Statistical inference with data-dependent treatment allocation rules. *Journal of the American Statistical Association*, 85, 156-162.

WELLMAN, B., FRANK, O., ESPINOZA, V., LUNDQUIST, S., and WILSON, C. (1991). Integrating individual, relational and structural analysis. *Social Networks*, 13, 223-249.

Survey Methodology, June 2000
Vol. 26, No. 1, pp. 99-107
Statistics Canada

Calibration and Restricted Weights

ALAIN THÉBERGE[1]

ABSTRACT

To better understand the impact of imposing a restricted region on calibration weights, the author reviews the latter's asymptotic behaviour. Necessary and sufficient conditions are provided for the existence of a solution to the calibration equation with weights within given intervals. A more general formulation of the calibration problem leads to a compromise between the need to satisfy the calibration equation and the attempt to obtain weights that are close to Horvitz-Thompson weights. If the requirements for the calibration equation are relaxed, then various estimation methods with restricted weights can be used. The estimators that are introduced usually have the same asymptotic properties as the calibration estimator with no weight restrictions, and some have weights which can be calculated explicitly, without any iterative process. The author shows how these estimators can be adapted to take advantage of a synthetic estimator. An approach similar to that used to restrict weights is applied to outliers.

KEY WORDS: Small domains; Moore-Penrose inverse; Inequality solutions; Asymptotic properties; Outliers.

1. INTRODUCTION

The calibration estimator has good asymptotic properties. However, for samples of small size, or if calibration is done at the domain level and some of the domains involve few observations, the weights of such an estimator can include extreme values. One way of overcoming this problem consists in using the calibration method with distance measurements which restrict the weights of observations to certain intervals about the sampling weights. This approach was developed by Deville and Särndal (1992). Other methods aimed at providing robust estimates satisfying the calibration equation can be found in Duchesne (1999). That paper contains an extensive bibliography on robust estimators. However, there is no guaranteed solution to the calibration equation with restricted weights. Even when such weights exist, the statistician might prefer solving the problem of extreme weights by relaxing somewhat the requirements for the calibration equation, instead of tightening the constraints on the weights by using a distance measurement that is more "restrictive". This paper provides a formulation of the calibration problem which offers more flexibility to the statistician. The problem in fact is one of minimization similar to that encountered in ridge regression. Bardsley and Chambers (1984) encountered this same minimization problem in their search for model-based estimators. This formulation of the calibration problem can be used to restrict weights without the use of special distances between calibrated weights and Horvitz-Thompson weights. Rao and Singh (1997) combined this approach with iterative methods using distance measurements. Other ways of restricting weights will also be reviewed.

In the next section, the calibration method is outlined without applying limits to the values of weights. The calibration problem thus outlined does not assume there is a solution to the calibration equation. The asymptotic properties of calibrated weights are discussed. These properties have a bearing on the asymptotic behaviour of the estimators whose weights are restricted. In section 3, necessary and sufficient conditions are provided for the existence of restricted weights which satisfy the calibration equation. Section 4 discusses how the estimation problem can be formulated by varying the importance attributed to the calibration equation. Section 5 provides various means of restricting weights without recourse to a specific distance. Section 6 introduces an estimator with restricted weights which is useful for small domains. Finally, in section 7, outliers are discussed in terms of a method similar to that used to deal with extreme weights.

2. CALIBRATION

Let $\boldsymbol{Y} \in \mathbb{R}^{N \times d}$ denote a matrix of d variables of interest for a population of size N, and let $\boldsymbol{c} \in \mathbb{R}^N$ denote a vector of known constants; a sample s of size n is drawn, and the subscript s is used to designate the sub-vectors or sub-matrices corresponding to the sample. We wish to estimate $\boldsymbol{Y}'\boldsymbol{c}$ using $\boldsymbol{Y}_s'\boldsymbol{w}_s$, where $\boldsymbol{w}_s \in \mathbb{R}^n$ is a weight vector for the sampled units. For a vector $\boldsymbol{v}$ and a positive diagonal matrix $\boldsymbol{F}$ of identical dimension, we define $\|\boldsymbol{v}\|_{\boldsymbol{F}}^2 = \boldsymbol{v}'\boldsymbol{F}\boldsymbol{v}$. For an auxiliary information matrix $\boldsymbol{X} \in \mathbb{R}^{N \times p}$, $\boldsymbol{A} \in \mathbb{R}^{N \times N}$ the diagonal matrix of sampling weights, given positive diagonal matrices $\boldsymbol{U}_s \in \mathbb{R}^{n \times n}$ and $\boldsymbol{T} \in \mathbb{R}^{p \times p}$, we seek, among the weight vectors $\boldsymbol{w}_s \in \mathbb{R}^n$ which minimize $\|\boldsymbol{X}_s'\boldsymbol{w}_s - \boldsymbol{X}'\boldsymbol{c}\|_{\boldsymbol{T}}^2$, the one which minimizes $D_s(\boldsymbol{w}_s) = \|\boldsymbol{w}_s - \boldsymbol{A}_s\boldsymbol{c}_s\|_{\boldsymbol{U}_s}^2$. This formulation of the problem, which does not assume the existence of weights satisfying the calibration equation, $\boldsymbol{X}_s'\boldsymbol{w}_s = \boldsymbol{X}'\boldsymbol{c}$, can be found in Théberge (1999). The

[1] Alain Théberge, Social Survey Methods Division, Statistics Canada, Ottawa, Ontario, K1A 0T6 Canada.

solution represents the vector of calibrated weights w_{cal}. We have

$$w_{cal} = a_s c_s + U_s^{-1} X_s T^{1/2} (T^{1/2} X_s' U_s^{-1} X_s T^{1/2})^{\dagger} T^{1/2} (X'c - X_s' a_s c_s), \quad (1)$$

where $F^{\dagger}$ denotes the Moore-Penrose inverse of the matrix F.

To better review the asymptotic properties of calibration estimators with restricted weights, let us now examine the behaviour of w_{cal} when $n \to \infty$. We assume there exists an asymptotic setup in which the size of the population and the size of the sample tend towards infinity (see for example Isaki and Fuller (1982)), and for which we have

$$\begin{aligned} Y'c &= O_p(N^{\gamma}) \ (\gamma \geq 0) \\ X'c - X_s' a_s c_s &= O_p(n^{-1/2} N^{\gamma}) \quad (2) \\ T^{1/2} X_s' U_s^{-1} X_s T^{1/2} &= O_p(n). \end{aligned}$$

It follows that $(T^{1/2} X_s' U_s^{-1} X_s T^{1/2})^{\dagger} = O_p(n^{-1})$, since one of the properties of the Moore-Penrose inverse of a matrix F is $F^{\dagger} F F^{\dagger} = F^{\dagger}$. Usually, we can expect to have $\gamma = 1$ when each element of the vector c has a value of 1 (estimate of a total), and $\gamma = 0$ when each element of c has a value of $1/N$ (estimate of a mean). For conditions (2) we therefore have,

$$\begin{aligned} w_{cal} - a_s c_s &= U_s^{-1} X_s T^{1/2} (T^{1/2} X_s' U_s^{-1} X_s T^{1/2})^{\dagger} T^{1/2} (X'c - X_s' a_s c_s) \\ &= O_p(n^{-1}) O_p(n^{-1/2} N^{\gamma}) \\ &= O_p(n^{-3/2} N^{\gamma}). \quad (3) \end{aligned}$$

Thus $w_{cal} - a_s c_s$ converges in probability to $\mathbf{0}$, if

$$\lim_{n, N \to \infty} n^{-3/2} N^{\gamma} = 0.$$

For an asymptotic setup such as that of Brewer (1979) in which the sampling fraction n/N is constant, or any setup for which the sampling fraction converges to a positive number, this condition is verified if $\gamma < 3/2$.

Writing $w_{cal} = a_s c_s + U_s^{-1} X_s T^{1/2} H_s^{\dagger} T^{1/2} (X'c - X_s' a_s c_s)$, where $H_s = T^{1/2} X_s' U_s^{-1} X_s T^{1/2}$, we have

$$\begin{aligned} D_s(w_{cal}) &= (X'c - X_s' a_s c_s)' T^{1/2} H_s^{\dagger} H_s H_s^{\dagger} T^{1/2} (X'c - X_s' a_s c_s) \\ &= (X'c - X_s' a_s c_s)' T^{1/2} H_s^{\dagger} T^{1/2} (X'c - X_s' a_s c_s) \\ &= O_p(n^{-1/2} N^{\gamma}) O_p(n^{-1}) O_p(n^{-1/2} N^{\gamma}) \\ &= O_p(n^{-2} N^{2\gamma}). \quad (4) \end{aligned}$$

Again for an asymptotic setup in which the sampling fraction converges to a positive number, we have $D_s(w_{cal})$ converging in probability to 0, if $\gamma < 1$. Thus there are cases, *e.g.* for the estimate of a total, where $w_{cal} - a_s c_s$ converges in probability to $\mathbf{0}$, but where $D_s(w_{cal}) = \| w_{cal} - a_s c_s \|_{U_s}^2$ does not converge to 0.

3. CALIBRATION EQUATION SOLUTIONS AND RESTRICTED WEIGHTS

Even in the absence of weight restrictions, there might not be a solution to the calibration equation. By applying Graybill (1983, 113) to the calibration problem, we find that the calibration equation $X_s' w_s = X'c$ can be solved if and only if $(X_s^{\dagger} X_s)' X'c = X'c$. If there is a solution, the calibrated weights might be negative or exceptionally large. Deville and Särndal (1992) proposed using various distance measures other than a weighted sum of squares to measure the distance between Horvitz-Thompson weights and calibrated weights, so as to restrict the weights to certain intervals while satisfying the calibration equation. This approach can only work if there are in these intervals weights which satisfy the calibration equation. The main goal of this section is to find necessary and sufficient conditions for the existence of a weight vector w_s within given bounds, such that the estimates of totals for auxiliary variables are also bounded. In other words, we are seeking conditions for the existence of a vector w_s such that $w^{(L)} \leq w_s \leq w^{(H)}$ and $t^{(L)} \leq X_s' w_s \leq t^{(H)}$, where $w^{(L)}$, $w^{(H)}$, $t^{(L)}$ and $t^{(H)}$ are given. In particular, by assuming $t^{(L)} = t^{(H)} = X'c$, we would obtain conditions for the existence of weights restricted to the intervals $w^{(L)} \leq w_s \leq w^{(H)}$, satisfying the calibration equation.

A first step is provided by the following Fan (1956) theorem. It is formulated here for a matrix M of finite dimension, although the proof provided by Fan also applies to a matrix of infinite dimension. The theorem uses the kernel of M', $N(M')$, defined as the set of vectors α such that $M'\alpha = \mathbf{0}$.

Theorem: Let $M \in \mathbb{R}^{m \times n}$ and $l \in \mathbb{R}^m$, $\exists\, w \in \mathbb{R}^n$ such that $Mw \geq l$ if and only if for any $\lambda \geq \mathbf{0}$ in $N(M')$, we have $l'\lambda \leq 0$.

Corollary: Let $M \in \mathbb{R}^{m \times n}$ and $l, h \in \mathbb{R}^m$, $\exists\, w \in \mathbb{R}^n$ such that $l \leq Mw \leq h$ if and only if first $l \leq h$ and secondly $\lambda \in N(M') \Rightarrow -l'\lambda_- \leq h'\lambda_+$, where $\lambda_+ = \max(\lambda, \mathbf{0})$ and $\lambda_- = \min(\lambda, \mathbf{0})$ with the extrema taken elementwise.

The corollary is obtained by using the theorem with

$$M = \begin{pmatrix} M \\ -M \end{pmatrix}, \ l = \begin{pmatrix} l \\ -h \end{pmatrix} \text{ and } \lambda = \begin{pmatrix} -\lambda_- \\ \lambda_+ \end{pmatrix}$$

Let p denote the dimension of $N(M')$. If p is equal to zero, then $\lambda \in N(M')$ implies $\lambda = \mathbf{0}$, and the condition of the theorem (or the similar condition of the corollary) is obviously met. If p is equal to one, then $\lambda \in N(M')$ implies that λ is a multiple of a vector z, and it is sufficient to

check the condition for $\lambda = z$ and $\lambda = -z$. If we use the property $(-\lambda)_- = -(\lambda)_+$, the problem outlined at the beginning of the section can now be resolved if X_s is a vector. The corollary with

$$M = \begin{pmatrix} I_n \\ X_s' \end{pmatrix}, l = \begin{pmatrix} w^{(L)} \\ t^{(L)} \end{pmatrix}, h = \begin{pmatrix} w^{(H)} \\ t^{(H)} \end{pmatrix},$$

and the fact that

$$z = \begin{pmatrix} -X_s \\ 1 \end{pmatrix}$$

spans $N(M')$, provide the necessary and sufficient conditions

$$\begin{aligned} w^{(L)} &\le w^{(H)} \\ t^{(L)} &\le t^{(H)} \\ (X_s)_+' w^{(L)} + (X_s)_-' w^{(H)} &\le t^{(H)} \\ t^{(L)} &\le (X_s)_+' w^{(H)} + (X_s)_-' w^{(L)}. \end{aligned} \tag{5}$$

The third inequality in (5) states that the weighted total of the auxiliary variable must not be greater than $t^{(H)}$, when the smallest possible weight $w^{(L)}$ is given to units for which the auxiliary variable is positive, and when the greatest possible weight $w^{(H)}$ is given to units for which the auxiliary variable is negative. The fourth inequality in (5) states that the weighted total of the auxiliary variable must not be less than $t^{(L)}$, when the largest possible weight is given to units for which the auxiliary variable is positive, and when the smallest possible weight is given to units for which the auxiliary variable is negative.

Even for $p > 1$, it is sufficient to check the condition of the corollary for a finite number of values of λ. Let $V \in \mathbb{R}^{m \times p}$, $2 \le p \le m$ denote a matrix whose columns form a basis for $N(M')$. It is always possible to construct V such that p of its rows, $v_1, v_2, ..., v_m$, are the unit vectors of $\mathbb{R}^p$, and we will assume that V is of this form. It will be shown in Appendix A that it is sufficient to check the condition of the corollary for vectors $\lambda = V\varphi$ and $\lambda = -V\varphi$, where $\varphi = (\varphi_1, ..., \varphi_p)'$ is a non-zero vector satisfying $v_i'\varphi = 0$ for a subset of $(p-1)$ linearly independent vectors v_i. We must therefore check the condition at the most for $\binom{m}{p-1}$ vectors φ, *i.e.* at the most $2\binom{m}{p-1}$ values of λ.

Using the corollary with

$$M = \begin{pmatrix} I_n \\ X_s' \end{pmatrix}, l = \begin{pmatrix} w^{(L)} \\ t^{(L)} \end{pmatrix}, h = \begin{pmatrix} w^{(H)} \\ t^{(H)} \end{pmatrix},$$

and noting that the columns of

$$V = \begin{pmatrix} -X_s \\ I_p \end{pmatrix}$$

form a basis for $N(M')$, we obtain the following necessary and sufficient conditions for the existence of a solution to the problem mentioned at the beginning of this section whenever $X_s \in \mathbb{R}^{n \times p}$ with $p > 1$. We must have $w^{(L)} \le w^{(H)}$, $t^{(L)} \le t^{(H)}$, and for each subset of $(p-1)$ linearly independent rows of

$$V = \begin{pmatrix} -X_s \\ I_p \end{pmatrix}$$

it is necessary that

$$\begin{aligned} (X_s\varphi)_+' w^{(L)} - \varphi_-' t^{(L)} &\le -(X_s\varphi)_-' w^{(H)} + \varphi_+' t^{(H)} \\ -(X_s\varphi)_-' w^{(L)} + \varphi_-' t^{(L)} &\le (X_s\varphi)_+' w^{(H)} - \varphi_-' t^{(H)} \end{aligned} \tag{6}$$

for a non-zero vector $\varphi \in \mathbb{R}^p$ orthogonal to each row of the subset. The second inequality in (6) is obtained from the first by changing the sign of φ.

If $V_{sub} \in \mathbb{R}^{p \times p}$ is a non-singular matrix whose rows are rows of V, then each column of V_{sub}^{-1} is a vector perpendicular to $(p-1)$ linearly independent rows of V. Hence the following result:

There exists a weight vector w_s such that $w^{(L)} \le w_s \le w^{(H)}$ and $t^{(L)} \le X_s' w_s \le t^{(H)}$ if and only if $w^{(L)} \le w^{(H)}$, $t^{(L)} \le t^{(H)}$ and

$$\begin{aligned} (X_s V_{sub}^{-1})_+' w^{(L)} - (V_{sub}^{-1})_-' t^{(L)} &\le -(X_s V_{sub}^{-1})_-' w^{(H)} \\ &\quad + (V_{sub}^{-1})_+' t^{(H)} \\ -(X_s V_{sub}^{-1})_-' w^{(L)} + (V_{sub}^{-1})_+' t^{(L)} &\le (X_s V_{sub}^{-1})_+' w^{(H)} \\ &\quad - (V_{sub}^{-1})_-' t^{(H)} \end{aligned} \tag{7}$$

for all non-singular matrixes $V_{sub} \in \mathbb{R}^{p \times p}$ whose rows are rows of

$$V = \begin{pmatrix} -X_s \\ I_p \end{pmatrix}.$$

These conditions are somewhat redundant. For example, if inequalities (7) are met for $V_{sub} = V_1$, then they are necessarily met for any matrix V_2 obtained from V_1 through a permutation of rows.

Another example is provided by weighting observations in a contingency table. Assuming $\hat{N}_{ij} = n_{ij} w_{ij}$ $(i = 1, 2, ..., R; j = 1, 2, ..., C)$, where n_{ij} is the number of observations in cell (i,j) of a contingency table and w_{ij} is the weight of each of these observations, we wish to know if there are weights w_{ij} such that $\hat{N}_{ij}$ satisfies certain constraints. For example, motivated by the problem of convergence of the raking ratio procedure, Bacharach (1965) provided necessary and sufficient conditions for the existence of weights w_{ij} such that $\hat{N}_{ij} \ge 0$, $\sum_{i=1}^{R} \hat{N}_{ij} = N_{.j}$ $(j = 1, ..., C)$, $\sum_{j=1}^{C} \hat{N}_{ij} = N_{i.}$ $(i = 1, ..., R)$, where the values of $N_{.j}$ and $N_{i.}$ are given. The following result, demonstrated in Appendix B, is more general.

For arbitrary constants $N_{ij}^{(L)}, N_{ij}^{(H)}, N_{.j}^{(L)}, N_{.j}^{(H)}, N_{i.}^{(L)}, N_{i.}^{(H)}, N_{..}^{(L)}$, and $N_{..}^{(H)}$, there are $\hat{N}_{ij}$ such that

$$\begin{aligned} &N_{ij}^{(L)} \le \hat{N}_{ij} \le N_{ij}^{(H)} && i=1,\ldots,R;\ j=1,\ldots,C;\\ &N_{.j}^{(L)} \le \sum_{i=1}^{R} \hat{N}_{ij} \le N_{.j}^{(H)} && j=1,\ldots,C;\\ &N_{i.}^{(L)} \le \sum_{j=1}^{C} \hat{N}_{ij} \le N_{i.}^{(H)} && i=1,\ldots,R;\\ &N_{..}^{(L)} \le \sum_{i=1}^{R}\sum_{j=1}^{C} \hat{N}_{ij} \le N_{..}^{(H)}, \end{aligned} \tag{8}$$

if and only if

$$\begin{aligned} &N_{ij}^{(L)} \le N_{ij}^{(H)} && i=1,\ldots,R;\ j=1,\ldots,C;\\ &N_{.j}^{(L)} \le N_{.j}^{(H)} && j=1,\ldots,C;\\ &N_{i.}^{(L)} \le N_{i.}^{(H)} && i=1,\ldots,R;\\ &N_{..}^{(L)} \le N_{..}^{(H)} \end{aligned}$$

$$\begin{aligned} &\sum_{j\in T}\left(N_{.j}^{(L)} - \sum_{i\notin S} N_{ij}^{(H)}\right) \le \sum_{i\in S}\left(N_{i.}^{(H)} - \sum_{j\notin T} N_{ij}^{(L)}\right)\\ &\sum_{i\in S}\left(N_{i.}^{(L)} - \sum_{j\notin T} N_{ij}^{(H)}\right) \le \sum_{j\in T}\left(N_{.j}^{(H)} - \sum_{i\notin T} N_{.j}^{(L)}\right)\\ &N_{..}^{(L)} + \sum_{j\notin T}\left(N_{.j}^{(H)} - \sum_{i\notin S} N_{ij}^{(H)}\right) \le \sum_{i\in S}\left(N_{i.}^{(H)} - \sum_{j\in T} N_{ij}^{(L)}\right) + \sum_{j=1}^{J} N_{.j}^{(H)}\\ &\sum_{i=1}^{I} N_{i.}^{(L)} + \sum_{j\notin T}\left(N_{.j}^{(L)} - \sum_{i\notin S} N_{ij}^{(H)}\right) \le \sum_{i\in S}\left(N_{i.}^{(L)} - \sum_{j\in T} N_{ij}^{(L)}\right) + N_{..}^{(H)} \end{aligned} \tag{9}$$

for any $S \subseteq \{1,2,\ldots,R\}$, $T \subseteq \{1,2,\ldots,C\}$.

The number of inequalities to be checked can be reduced. For example, instead of checking

$$\sum_{j\in T}\left(N_{.j}^{(L)} - \sum_{i\notin S} N_{ij}^{(H)}\right) \le \sum_{i\in S}\left(N_{i.}^{(H)} - \sum_{j\notin T} N_{ij}^{(L)}\right)$$

for any $S \subseteq \{1,2,\ldots,R\}$, and $T \subseteq \{1,2,\ldots,C\}$, it can be readily shown that an equivalent procedure would be to check that

$$\sum_{j\in T} N_{.j}^{(L)} \le \sum_{i=1}^{R} \min\left(\left(N_{i.}^{(H)} - \sum_{j\notin T} N_{ij}^{(L)}\right), \sum_{j\in T} N_{ij}^{(H)}\right)$$

for any $T \subseteq \{1,2,\ldots,C\}$.

4. MITIGATED CALIBRATION

There may be dissatisfaction with the two-step approach of calibration, where an attempt is first made to find weight vectors that best satisfy the calibration equation, and then from this set of vectors to find the one which comes closest to Horvitz-Thompson weights. For small samples, this method may lead to weights which the statistician will find too far from Horvitz-Thompson weights.

There may be a preference for varying the importance attributed to the calibration equation relative to the norm of $\boldsymbol{w}_s - \boldsymbol{A}_s\boldsymbol{c}_s$. Thus, there may be a desire to find a weight vector $\boldsymbol{w}_s$ which minimizes

$$\left\| \begin{pmatrix} \boldsymbol{w}_s - \boldsymbol{A}_s\boldsymbol{c}_s \\ \boldsymbol{X}_s'\boldsymbol{w}_s - \boldsymbol{X}'\boldsymbol{c} \end{pmatrix} \right\|_{\boldsymbol{V}}^2,$$

where

$$\boldsymbol{V} = \begin{pmatrix} \boldsymbol{U}_s & \boldsymbol{0} \\ \boldsymbol{0} & \alpha\boldsymbol{T} \end{pmatrix}$$

and $\alpha \ge 0$. We then minimize

$$\| \boldsymbol{w}_s - \boldsymbol{A}_s\boldsymbol{c}_s \|_{\boldsymbol{U}_s}^2 + \alpha \| \boldsymbol{X}_s'\boldsymbol{w}_s - \boldsymbol{X}'\boldsymbol{c} \|_{\boldsymbol{T}}^2 = D_s(\boldsymbol{w}_s) + \alpha \| \boldsymbol{X}_s'\boldsymbol{w}_s - \boldsymbol{X}'\boldsymbol{c} \|_{\boldsymbol{T}}^2 .$$

A similar minimization problem is encountered with ridge regression. For $\alpha = 0$ the solution is provided by Horvitz-Thompson weights $\boldsymbol{w}_s = \boldsymbol{A}_s\boldsymbol{c}_s$. For $\alpha > 0$, we seek $\boldsymbol{w}_s(\alpha)$ minimizing $\| \boldsymbol{K}(\boldsymbol{w}_s - \boldsymbol{A}_s\boldsymbol{c}_s) - \boldsymbol{b} \|_{\boldsymbol{V}}^2$, where $\boldsymbol{K} = (\boldsymbol{I}_n, \boldsymbol{X}_s)'$, $\boldsymbol{b} = (\boldsymbol{0}_{1\times n}, (\boldsymbol{X}'\boldsymbol{c} - \boldsymbol{X}_s'\boldsymbol{A}_s\boldsymbol{c}_s)')'$ and $\boldsymbol{0}_{1\times n} \in \mathbb{R}^n$ is a row vector of zeros. Ben-Israel and Greville (1980) yields

$$\boldsymbol{w}_s(\alpha) - \boldsymbol{A}_s\boldsymbol{c}_s = (\boldsymbol{K}'\boldsymbol{V}\boldsymbol{K})^{-1}\boldsymbol{K}'\boldsymbol{V}\boldsymbol{b}. \tag{10}$$

Thus by substituting the values of $\boldsymbol{K}$, $\boldsymbol{V}$, and $\boldsymbol{b}$ we obtain

$$\boldsymbol{w}_s(\alpha) = \boldsymbol{A}_s\boldsymbol{c}_s + \alpha(\boldsymbol{U}_s + \alpha\boldsymbol{X}_s\boldsymbol{T}\boldsymbol{X}_s')^{-1}\boldsymbol{X}_s\boldsymbol{T}(\boldsymbol{X}'\boldsymbol{c} - \boldsymbol{X}_s'\boldsymbol{A}_s\boldsymbol{c}_s). \tag{11}$$

It is easily shown that

$$\alpha(\boldsymbol{U}_s + \alpha\boldsymbol{X}_s\boldsymbol{T}\boldsymbol{X}_s')^{-1}\boldsymbol{X}_s\boldsymbol{T} = \boldsymbol{U}_s^{-1}\boldsymbol{X}_s(\alpha^{-1}\boldsymbol{T}^{-1} + \boldsymbol{X}_s'\boldsymbol{U}_s^{-1}\boldsymbol{X}_s)^{-1},$$

hence

$$\boldsymbol{w}_s(\alpha) = \boldsymbol{A}_s\boldsymbol{c}_s + \boldsymbol{U}_s^{-1}\boldsymbol{X}_s(\alpha^{-1}\boldsymbol{T}^{-1} + \boldsymbol{X}_s'\boldsymbol{U}_s^{-1}\boldsymbol{X}_s)^{-1}(\boldsymbol{X}'\boldsymbol{c} - \boldsymbol{X}_s'\boldsymbol{A}_s\boldsymbol{c}_s). \tag{12}$$

The estimator $Y_s' w_s(\alpha)$ thus becomes $\hat{Y}'c + (Y_s - \hat{Y}_s)' A_s c_s$, where $\hat{Y} = X\hat{\beta}_s(\alpha)$ and

$$\hat{\beta}_s(\alpha) = (X_s' U_s^{-1} X_s + \alpha^{-1} T^{-1})^{-1} X_s' U_s^{-1} Y_s.$$

The vector of regression coefficients, then, is the one obtained with ridge regression. Just as the calibration method, and the generalized regression method described by Särndal, Swensson and Wretman (1992), lead to the same estimators, a similar parallel can be drawn between mitigated calibration and ridge regression.

On the basis of equation (12), we can also use Ben-Israel and Greville (1980), and the fact that $F^{\dagger} = F'(FF')^{\dagger}$ with $F = T^{1/2} X_s' U_s^{-1/2}$, to show that

$$\lim_{\alpha \to \infty} w_s(\alpha) = w_{\text{cal}}.$$

This result was to be expected, since finding the vector $w_s(\alpha)$ which minimizes $D_s(w_s) + \alpha \| X_s' w_s - X'c \|_T^2$ when $\alpha \to \infty$ is equivalent to finding the weight vector which minimizes $D_s(w_s)$ among those which minimize $\| X_s' w_s - X'c \|_T^2$.

Rao and Singh (1997) defined tolerances for each of the p constraints of the calibration equation, and they established a relationship between these tolerances and the matrix αT.

For $\alpha \in [0, \infty[$ the function $w_s(\alpha)$ is represented by a curve in $\mathbb{R}^n$ which links point $A_s c_s$ to point w_{cal}. If $p = 1$, *i.e.* if X is a vector, this curve is a line segment. In fact, in this case the matrix $(\alpha^{-1} T^{-1} + X_s' U_s^{-1} X_s)^{-1}$ and the vector $X'c - X_s' A_s c_s$ are scalars, and the weights $w_s(\alpha)$ given by (12) are therefore equal to Horvitz-Thompson weights plus a multiple of vector $U_s^{-1} X_s$. And again for $p = 1$, we have

$$\lim_{\alpha \to \infty} w_s(\alpha) = w_{\text{cal}} = A_s c_s + [(X'c - X_s' A_s c_s) / (X_s' U_s^{-1} X_s)] U_s^{-1} X_s$$

which leads to the estimator

$$Y_s' w_{\text{cal}} = Y_s' A_s c_s + [(Y_s' U_s^{-1} X_s) / (X_s' U_s^{-1} X_s)](X'c - X_s' A_s c_s)$$

Taking $U = A^{-1} \text{diag}(X)$, we obtain the ratio estimator

$$Y_s' A_s c_s + [(Y_s' A_s 1_{n \times 1}) / (X_s' A_s 1_{n \times 1})](X'c - X_s' A_s c_s),$$

where $1_{a \times b} \in \mathbb{R}^{a \times b}$ is a matrix of ones.

Ben-Israel and Greville (1980, 111, exercise 15) showed that $D_s(w_s(\alpha))$ is an increasing monotonic function of α. Note however that for a unit $k \in s$, $|w_k(\alpha) - a_k c_k|$ is not necessarily a monotonic function of α. As α increases, the weight vector $w_s(\alpha)$ moves away from the Horvitz-Thompson weight vector, but this does not necessarily apply to each coordinate of the vector.

In this article, mitigated calibration is used to restrict weights, *i.e.* when the size of the sample is relatively small. It can easily be shown, however, that for an asymptotic setup satisfying (2) and for which $\hat{\beta}_s(\alpha) - \beta(\alpha) \to 0$ in probability, with

$$\beta(\alpha) = (X' U^{-1} X + \alpha^{-1} T^{-1})^{-1} X' U^{-1} Y,$$

we have $Y_s' w_s(\alpha)$ is an asymptotically unbiased estimator whose asymptotic variance is

$$(Y - Y^*)' \text{diag}(c)(A \Pi A - 1_{N \times N}) \text{diag}(c)(Y - Y^*),$$

where $Y^* = X\beta(\alpha)$, Π is the matrix of inclusion probabilities of order 2, and $\text{diag}(c)$ is the diagonal matrix formed from vector c.

5. ESTIMATION METHODS WITH RESTRICTED WEIGHTS

In order to avoid obtaining weights having extreme values, we may wish to force the weight vector to be within a given region. This restricted region will be assumed to be convex and closed, and $A_s c_s$ will be assumed to be a point in this region. For example, if $w^{(L)} < A_s c_s < w^{(H)}$, we may wish to restrict the weights to region $R_w = \{w_s : w^{(L)} \leq w_s \leq w^{(H)}\}$. We will assume that

$$\lim_{n \to \infty} w^{(L)} - A_s c_s < 0 \quad \text{and} \quad \lim_{n \to \infty} w^{(H)} - A_s c_s > 0.$$

The approach described in section 3 consists in selecting a distance measure between calibrated weights and Horvitz-Thompson weights which will provide weights that satisfy the calibration equation and which lie in the restricted region, should such weights in fact exist. The approach dealt with in this section is to temperate the requirement that the calibration equation be satisfied when the vector of calibration weights w_{cal} is outside the restricted region. Various means to temperate this requirement lead to different weighting methods.

When w_{cal} lies outside the restricted region, we could for example look for those points on the curve $w_s(\alpha)$ parametered by $\alpha \geq 0$ which are on the border of this region. One property of these points is that they solve the minimization problem described in section 4 for corresponding values of α, thus through matrix T, the importance of each calibration equation can be weighted. Using the example of the restricted region provided above, if

$$w_{\text{cal}} = \lim_{\alpha \to \infty} w_s(\alpha)$$

lies within this region, then $w_{\text{res 1}} = w_{\text{cal}}$ can be used as a restricted weight vector, otherwise $w_{\text{res 1}} = w_s(\alpha)$ with

$\alpha < \infty$ can be chosen such that $w_s(\alpha)$ is on the boundary of the restricted region. If the asymptotic setup is such that conditions (2) are met with $\gamma < 3/2$ then for n sufficiently large, the probability that w_{cal} will be within the restricted region is equal to one. In fact, we have $w_{cal} - A_s c_s$ converging in probability to $\mathbf{0}$. The asymptotic properties of the estimator using the restricted weights, $w_{res\,1}$, are therefore identical to those of the calibration estimator. It is worth noting that since $|w_k - a_k c_k|$ is not necessarily a monotonic function of α, it is possible for $w_s(\alpha)$ to be on the boundary of the restricted region for several values of α, even if the restricted region is convex. Finding all these values is not necessarily a simple matter, and a decision has to be taken as to which value to use.

Another option for restricting weights would be to use as a restricted region those weights w_s which satisfy $D_s(w_s) \leq l$ for a bound $l > 0$. Then $w_{res\,2} = w_{cal}$ is taken as a restricted weight vector if w_{cal} lies in the restricted region, otherwise we seek $\alpha > 0$ such that $D_s(w_s(\alpha)) = l$. This value of α is unique and can be found through iteration. Next we calculate the weights $w_{res\,2} = w_s(\alpha)$ which correspond to this value of α using equation (12). If the asymptotic setup is such that conditions (2) are met with $\gamma < 1$, and if l does not vary with n, then for n sufficiently large, the probability that w_{cal} will be within the restricted region is equal to one. In fact, we have $D_s(w_{cal})$ converging in probability to 0. The asymptotic properties of the estimator using restricted weights, $w_{res\,2}$, are then identical to those of the calibration estimator. Unfortunately, when estimating a total, we must expect to have $\gamma = 1$. In order to overcome this snag, we can use $l\sqrt{n}$ as a bound, instead of l. We can justify this bound on the basis that the length of the main diagonal of a hypercube of $\mathbb{R}^n$ is equal to the diameter of the sphere which circumscribes this hypercube, whereas the diameter of the sphere inscribed in this same hypercube is smaller by a factor of $\sqrt{n}$. The fact remains that a statistician might be uncomfortable using an asymptotic setup where the bound increases with the size of the sample. Furthermore, with this approach, the weights of the observations cannot be limited individually. Only the distance between the restricted weight vector and the Horvitz-Thompson weight vector is controlled.

With the methods described above, we look for those points on curve $w_s(\alpha)$ which lie on the boundary of the restricted region. The value of α for which $w_s(\alpha)$ lies on the boundary of the restricted region must often be found iteratively. It would be simpler to replace the curve $w_s(\alpha)$ by the line segment linking $A_s c_s$ to w_{cal}. For the restricted region R_w, the restricted weight vector would be $w_{res\,3} = w_{cal}$ if w_{cal} is in the restricted region, otherwise $w_{res\,3}$ would be equal to the point at which the line segment crosses the boundary of restricted region, *i.e.*

$$w_{res\,3} = A_s c_s + \xi(w_{cal} - A_s c_s),$$

where

$$\xi = \min_k \{\max[(w^{(L)} - A_s c_s)/(w_{cal} - A_s c_s), (w^{(H)} - A_s c_s)/(w_{cal} - A_s c_s)]\},$$

vector division being elementwise, the maximum of the two vectors being elementwise, and $\min_k$ providing the minimum element. We could also consider the weight vector of the restricted region, $w_{res\,4}$, which comes closest to w_{cal}. Again for restricted region R_w, we would have

$$w_{res\,4} = \min\,[\max\,(w_{cal}, w^{(L)}), w^{(H)}].$$

The asymptotic properties of estimators using restricted weights $w_{res\,3}$ or $w_{res\,4}$ are identical to those of the calibration estimator, as long as $w_{cal} - A_s c_s$ converges in probability to $\mathbf{0}$, which is usually the case.

One interesting property of all the approaches discussed in this section is that, no matter what the restricted region, the existence of restricted weights is guaranteed. This is not always the case when using an approach based on distance measures. A simple example will now be introduced to allow comparisons between a few approaches.

We wish to estimate a total on the basis of a simple random sample of size 2 in a population of size 20. In other words, $c = \mathbf{1}_{20\times 1}$ and $a = 10\,(\mathbf{1}_{20\times 1})$. We use the auxiliary information vector $X = (1, 2, 3, \ldots, 20)'$, assume that the selected sample is $s = \{2, 12\}$ and choose U as a diagonal matrix with $u_{kk} = x_k = k$. A rectangular restricted region is provided using points $w^{(L)} = (0,0)'$ and $w^{(H)} = (20, 13)'$. In other words, the weight of the first sample unit must be greater than 0 and less than 20, whereas the weight of the second sample unit must be greater than 0 and less than 13.

Under these conditions, the calibrated weights $w_{cal} = (15, 5)'$ lie outside the restricted region. Since $p = 1$, weights $w_s(\alpha)$ lie on the line segment which links $A_s c_s = (10, 10)'$ to w_{cal}. We therefore have $w_{res\,1} = w_{res\,3}$, which means that the two methods give the same result. In this case, we have $w_{res\,1} = w_{res\,3} = (13, 13)'$. The method which consists in choosing that point in the restricted region which lies closest to the calibrated weights yields $w_{res\,4} = (15, 13)'$. On the other hand, if we look for $w_{res\,5}$, the restricted weights obtained while requiring that the calibration equation be satisfied and while using a distance measurement which assumes an infinite value outside the restricted region, then there is no solution. In fact, for any weight in the restricted region $X_s' w_s \leq 196$, whereas $X'c = 210$. If we had, say, $w^{(H)} = (30, 13)'$, then using $D_s(w_s)$ as a distance measurement within the restricted region we would have $w_{res\,5} = (27, 13)'$. These weights are fairly distant from $w_{cal} = (15, 15)'$ and from $A_s c_s = (10, 10)'$. Such is the price to be paid for insisting on having weights which meet the calibration equation.

6. ESTIMATORS FOR DOMAINS WITH A SYNTHETIC COMPONENT

Restricted weights are used because of the properties of the calibration estimator for small sample sizes. For large sample sizes, we normally have $w_{cal} - A_s c_s$ converging in probability to $\mathbf{0}$, *i.e.* weights that are not problematic. A statistician faced with a problem of extreme weights must therefore in all likelihood also face another problem of small sample sizes, *i.e.* estimation for small domains. This section introduces an estimator whose asymptotic properties are those of the calibration estimator, but which uses restricted weights and takes advantage of a synthetic estimator.

Let $\tilde{Y} = X\tilde{\beta}_s$ denote a synthetic estimate for Y, we have

$$\begin{aligned} \tilde{Y}'w_s &= (X_s\tilde{\beta}_s)'w_s \\ &= \tilde{\beta}_s'X_s'w_s \\ &= \tilde{\beta}_s'X'c \\ &= (X\tilde{\beta}_s)'c \\ &= \tilde{Y}'c \end{aligned} \qquad (13)$$

with equality at the third step if the weights satisfy the calibration equation $X_s'w_s = X'c$. The weights w_{cal} given by (1) minimize $\|X_s'w_s - X'c\|_T^2$. We can therefore estimate $Y'c$ using

$$\hat{\tau} = (Y_s - \tilde{Y}_s)'w_{res} + \tilde{Y}'c. \qquad (14)$$

There will be equality between this estimator and estimator $Y_s'w_{cal}$ once the sample is sufficiently large for the calibration equation to be satisfied and for w_{cal} to lie in the restricted region, *i.e.* once $w_{res} = w_{cal}$. The asymptotic properties of these two estimators are therefore identical under certain conditions discussed in the previous section. The advantage of using estimator $\hat{\tau}$ is that it provides a synthetic estimate when columns of Y_s and $\tilde{Y}_s$ are zero.

7. OUTLIERS

Outliers could be dealt with in much the same way as extreme weights. The strategy is the following: we adopt a restricted region for $Y_s'w_{cal}$, we show that when n is sufficiently large $Y_s'w_{cal}$ lies within the restricted region, and we adopt a more "reasonable" estimator to replace $Y_s'w_{cal}$ in those cases where $Y_s'w_{cal}$ lies outside the restricted region. For a stratified sample, we would normally have one restricted region per stratum.

In section 2, it was shown that under certain conditions for the asymptotic setup, $w_{cal} - A_s c_s = O_p(n^{-3/2}N^{\gamma})$. We thus have $Y_s'w_{cal} - Y_s'A_s c_s = O_p(n^{-1/2}N^{\gamma})$, and if we assume that

$$Y_s'A_s c_s - Y'c = O_p(n^{-1/2}N^{\gamma}), \qquad (15)$$

then $Y_s'w_{cal} - Y'c = O_p(n^{-1/2}N^{\gamma})$. An expert (or a group of experts) could determine on the basis of information gathered independently of survey data that it would not be reasonable to have $Y_s'w_{cal}$ outside a certain region. If $Y'c$ lies within the restricted region (*i.e.* if the expert does not find it unreasonable to have an estimate of the parameter which would be equal to the true value, $Y'c$, of the parameter), if $\gamma = 0$, and if the restricted region does not vary with n or N (or if $\gamma = 1$, and the restricted region varies in proportion to N), then for sufficiently large n, the probability that $Y_s'w_{cal}$ will lie within the restricted region is equal to one. In those cases where $Y_s'w_{cal}$ lies outside the restricted region, we could use as an estimate the point in the restricted region that lies closest to $Y_s'w_{cal}$ or we could assume that the weight of the few observations that are deemed outliers is equal to one, and distribute their original weights (less the number of outliers) among the observations that are not outliers. The asymptotic properties of this modified estimator used to deal with outliers are then identical to those of the unmodified estimator.

In the case of a non-stratified sample, this method is relatively easy to apply. If however the sample is stratified, and if constraints are imposed on estimates for each stratum, then we have two additional problems. First, if the asymptotic setup is such that the number of strata increases in proportion to the size of the sample, then the assumption given in (15) does not hold, since the mean sample size per stratum remains constant as $n \to \infty$. We need to determine whether it is reasonable to adopt an asymptotic setup in which the number of strata is constant (or increases less rapidly than n). Such an asymptotic setup is less plausible if the number of observations per stratum is small. The second problem is linked to the difficulty for the expert to impose constraints on estimates for each of the strata. The greater the number of strata, the greater the risk that $Y'c$ will not lie in the restricted region defined by the expert. In fact, in the case of a stratified sample, it is preferable for the expert to use information that is independent of the survey data, in order to ensure strata homogeneity, prior to finalizing stratification. In other words, it is preferable to use the information available before the survey, in order to prevent outliers, rather than to correct them. If the information has been used in such a way that, before the survey, there is no reason to believe that there is any unrepresentative observation in any stratum, then there is no justification for assuming the opposite once the data have been collected.

8. CONCLUSION

If for large sample sizes the calibrated weights remain within a restricted region, then the asymptotic properties of the estimator with restricted weights are obviously identical to those of the calibration estimator. For a given asymptotic setup, we can usually expect to have $w_{cal} - A_s c_s$ converging in probability to 0, *i.e.* for sufficiently large sample sizes the calibrated weights w_{cal} will remain within the restricted region R_w if $A_s c_s$ lies within R_w. However, we have seen that for the estimate of a total, we do not necessarily have convergence to 0 for $D_s(w_{cal})$. We must therefore avoid having a restricted region defined by $\| w_s - A_s c_s \|_{v_s}^2 \leq l$ at least if we are estimating a total and not a mean.

We have provided necessary and sufficient conditions for the existence of weights restricted to intervals which satisfy the calibration equation. If such weights do not exist, the idea of satisfying the calibration equation exactly must be abandoned. The problem of calibration with restricted weights can be reformulated in such a way that a solution will always be possible. Some of the approaches described in this paper make it possible to obtain a solution without recourse to iterative methods. These are simple methods that are easy to interpret. The asymptotic properties of these estimators are usually identical to those of the calibration estimator without weight restrictions.

The problem of extreme weights is encountered for small sample sizes, thus the problem of estimating for small domains should be considered simultaneously. It is possible to take advantage of synthetic estimators while using an estimator with restricted weights having good asymptotic properties.

It is also possible to modify the calibration estimator, or any other asymptotically consistent estimator, so as to deal with outliers. The conditions under which this modified estimator will have asymptotic properties identical to those of the unmodified estimator are not easily verified, just as it is difficult to verify whether an outlier is in fact unrepresentative. However, such conditions make it possible to identify those factors which allow an estimator that is corrected for outliers to be statistically valid.

ACKNOWLEDGEMENT

The author wishes to thank an associate editor and a referee for constructive comments which have helped improve the paper.

APPENDIX A

We wish to verify that $\Omega(\varphi) = l'(V\varphi)_- - h'(V\varphi)_+$ has a value of zero or less. First, it is easily shown that this is true for a vector φ, if and only if it is true for a vector $k\varphi$ with arbitrary $k > 0$. Only the direction of φ matters. It is therefore sufficient to verify the condition for φ of norm equal to one. For this proof, we will use the l_1-norm of φ, $\|\varphi\|_{l_1} = \sum_{i=1}^{p} |\varphi_i|$. Vectors φ with $\|\varphi\|_{l_1} = 1$ are located in hyperplanes whose intersections lie on points orthogonal to the unit vectors, *i.e.* points at least one of whose coordinates is zero. Function Ω varies linearly except at points φ orthogonal to one or more rows of V. Even when the domain of Ω is restricted to vectors φ with $\|\varphi\|_{l_1} = 1$ that are orthogonal to $0 \leq j < (p-1)$ linearly independent rows of V, function Ω still varies linearly except at points orthogonal to other rows of V or orthogonal to unit vectors (which are likewise rows of V). The maximum of Ω for $\|\varphi\|_{l_1} = 1$ is therefore reached at a point φ orthogonal to $(p-1)$ linearly independent rows of V. It is therefore sufficient to verify the condition for two vectors of opposite direction which are orthogonal to $(p-1)$ linearly independent rows of V, and this for each subset of $(p-1)$ linearly independent rows of V.

APPENDIX B

Let $\mathrm{vec}(F)$ denote the vector obtained by piling successive columns of matrix $F \in \mathbb{R}^{a \times b}$ with the first column on top, and let the Kronecker product of two matrices F and G be defined as

$$F \otimes G = \begin{pmatrix} f_{11}G & \cdots & f_{1n}G \\ \vdots & & \vdots \\ f_{m1}G & \cdots & f_{mn}G \end{pmatrix}. \tag{B1}$$

The result is derived from the corollary in section 3 with

$$M = \begin{pmatrix} I_{RC} \\ I_R \otimes \mathbf{1}_{1\times C} \\ \mathbf{1}_{1\times R} \otimes I_C \\ \mathbf{1}_{1\times RC} \end{pmatrix}, \quad w = \mathrm{vec}((\hat{N}_{ij})'),$$

$$l = \begin{pmatrix} \mathrm{vec}((N_{ij}^{(L)})') \\ N_{1.}^{(L)} \\ \vdots \\ N_{R.}^{(L)} \\ N_{.1}^{(L)} \\ \vdots \\ N_{.C}^{(L)} \\ N_{..}^{(L)} \end{pmatrix}, \quad h = \begin{pmatrix} \mathrm{vec}((N_{ij}^{(H)})') \\ N_{1.}^{(H)} \\ \vdots \\ N_{R.}^{(H)} \\ N_{.1}^{(H)} \\ \vdots \\ N_{.C}^{(H)} \\ N_{..}^{(H)} \end{pmatrix}. \tag{B2}$$

Only a finite set of conditions need be verified, first by noting that the columns of

$$V = \begin{pmatrix} -I_R \otimes 1_{C\times 1} & -1_{R\times 1} \otimes I_C & -1_{RC\times 1} \\ I_R & 0_{R\times C} & 0_{R\times 1} \\ 0_{C\times R} & I_C & 0_{C\times 1} \\ 0_{1\times R} & 0_{1\times C} & 1 \end{pmatrix} \quad \text{(B3)}$$

form a basis for $N(M')$. In other words, $M'V = 0$, the columns of V are linearly independent, and $N(M')$ is of dimension $R+C+1$. Note also that the last $R+C+1$ rows of V are the unit vectors. Finally, we verify the conditions of the corollary for all vectors $\lambda = V\varphi$ and $\lambda = -V\varphi$, where φ is orthogonal to $R+C$ linearly independent rows of V. This last step is described in greater detail in the following paragraph.

An arbitrary subset of $R+C$ linearly independent rows of V which includes the last row of V is denoted L, and the subset of all rows of V which are linear combinations of rows of L is denoted L^+. If L^+ includes row $RC+i$ $(i = 1, ..., R)$ if and only if $i \notin S \subseteq \{1,2, ..., R\}$, and includes row $RC+R+j$ $(j = 1, ..., C)$ if and only if $j \notin T \subseteq \{1,2, ..., C\}$, then we set $\varphi = (\varphi_S', -\varphi_T', 0)'$, where the i-th element of $\varphi_S \in \mathbb{R}^R$ is equal to 1 if $i \in S$ and to 0 otherwise, and the j-th element of $\varphi_T \in \mathbb{R}^C$ is equal to 1 if $j \in T$ and to 0 otherwise. Then

$$V\varphi = ((-\varphi_S \otimes 1_{C\times 1} + 1_{R\times 1} \otimes \varphi_T)', \varphi_S', -\varphi_T', 0)',$$

therefore φ is orthogonal to all rows of L^+, and all the more so φ is orthogonal to all rows of L. Likewise, vector $\varphi^* = (\varphi_S', \varphi_T', -1)'$ is orthogonal to all rows of a subset of $R+C$ linearly independent rows of V which includes row $RC+i$ $(i = 1, ..., R)$ if and only if $i \notin S$, and includes row $RC+R+j$ $(j = 1, ..., C)$ if and only if $j \notin T$, but does not include the last row of V. The condition $-l'\lambda_- \le h'\lambda_+$ with $\lambda = V\varphi$ provides the fifth set of inequalities in (9). Likewise, by assuming λ equal to $-V\varphi$, $V\varphi^*$ and $-V\varphi^*$ we obtain the last three sets of inequalities in (9).

REFERENCES

BACHARACH, M. (1965). Estimating nonnegative matrices from marginal data. *International Economic Review*, 6, 294-310.

BARDSLEY, P., and CHAMBERS, R.L. (1984). Multipurpose estimation from unbalanced samples. *Applied Statistics*, 33, 290-299.

BEN-ISRAEL, A., and GREVILLE, T.N.E. (1980). *Generalized Inverses: Theory and Applications*. Huntington, New York: Robert E. Krieger Publishing Company.

BREWER, K.R.W. (1979). A class of robust sampling designs for large scale surveys. *Journal of the American Statistical Association*, 74, 911-915.

DEVILLE, J.-C., and SÄRNDAL, C.-E. (1992). Calibration estimators in survey sampling. *Journal of the American Statistical Association*, 87, 376-382.

DUCHESNE, P. (1999). Robust calibration estimators. *Survey Methodology*, 25, 43-56.

FAN, K. (1956). On systems of linear inequalities. *Annals of Mathematics Studies*, (Eds. H. W. Kuhn, and A. W. Tucker), 38, 99-156.

GRAYBILL, F.A. (1983). *Matrices with Applications in Statistics*, (Second Edition). Belmont, California: Wadsworth Publishing.

ISAKI, C.T., and FULLER, W.A. (1982). Survey design under the regression superpopulation model. *Journal of the American Statistical Association*, 77, 89-96.

RAO, J.N.K., and SINGH, A.C. (1997). A ridge-shrinkage method for range-restricted weight calibration in survey sampling. *Proceedings of the Section on Survey Research Methods, American Statistical Association.*

SÄRNDAL, C.-E., SWENSSON, B., and WRETMAN, J. (1992). *Model Assisted Survey Sampling*. New York: Springer-Verlag.

THÉBERGE, A. (1999). Extensions of calibration estimators in survey sampling. *Journal of the American Statistical Association*, 94, 635-644.

Survey Methodology, June 2000
Vol. 26, No. 1, pp. 109-111
Statistics Canada

A Cautionary Note on Adjusting Weights for Nonresponse

WILLARD C. LOSINGER, LINDSEY P. GARBER, BRUCE A. WAGNER and GEORGE W. HILL[1]

ABSTRACT

For surveys which involve more than one stage of data collection, one method recommended for adjusting weights for nonresponse (after the first stage of data collection) entails utilizing auxiliary variables (from previous stages of data collection) which are identified as predictors of nonresponse. In the final stage of data collection for the United States National Animal Health Monitoring System's Beef '97 Study, two variables were identified that clearly separated eligible producers by their propensity to respond. However, these variables were noticeably inferior to simple region by herd-size categories as predictors of responses that eligible producers gave for other questions in previous data-collection stages. Therefore, we decided to form weight-adjustment classes by region and herd size, even though other variables were greater predictors of response. When selecting auxiliary variables to adjust weights for nonresponse, we recommend that survey statisticians also evaluate the extent to which these auxiliary variables are related to data which nonrespondents would have provided. Using auxiliary variables which exhibit the greatest variation in response propensity may result in the greatest variation in weight-adjustment factors, but may bias population estimates for parameters unrelated to the chosen auxiliary variables.

KEY WORDS: Nonresponse bias; Response propensity; Logistic regression; National survey.

1. INTRODUCTION

In multistage surveys where some participants fail to respond during the final stage of data collection, one has considerable information about final-stage nonrespondents from previous stages of the survey. Rizzo, Kalton and Brick (1996) presented several methods for selecting auxiliary variables and adjusting weights for nonresponse when a large number of characteristics of the nonrespondents were known. These methods concentrated on identifying and using characteristics that discriminated between respondents and eligible nonrespondents. However, by adjusting weights based on specific variables which demonstrate the greatest difference in response rates, one may potentially introduce bias in the survey estimates if these variables are unrelated to responses that would have been given by nonrespondents during the final stage of data collection. Therefore, one should also utilize data from the previous stages of data collection to determine whether the chosen auxiliary variables are linked to other characteristics of those eligible to participate in the survey.

The Beef '97 Study (of the National Animal Health Monitoring System (NAHMS) of the United States Department of Agriculture (USDA)) took place in 23 states and involved three stages of data collection. In the first stage (December 30, 1996 through February 3, 1997), enumerators from the USDA: National Agricultural Statistics Service collected data on general management practices from 2,713 agricultural operations with one or more beef cows. First-stage respondents who had five or more beef cows on January 1, 1997 were eligible to continue in the second stage of data collection (from March 3 through May 23, 1997), provided they had at least one beef cow and remained in business at the time of the second stage of data collection. A total of 1,190 producers participated in the second stage of data collection, which involved an on-farm visit by a veterinary medical officer or animal health technician and concentrated on the health management of the beef cattle.

All operations that participated in the second stage of data collection were eligible to participate in the third and final stage of data collection (August 1, 1997 through January 31, 1998). A total of 952 (80.0%) eligible operations responded in the final stage. From the first two stages of data collection, a considerable amount of information was available on the 238 nonrespondents for the final stage of data collection. The purpose of this note is to describe the methods that were evaluated for adjusting the sample weights for nonresponse in the final stage of data collection for the NAHMS Beef '97 Study.

In addition to region and herd-size (based on the number of beef cows) categories, 45 variables based on data collected during the first two stages of interviews were evaluated for their impact on final-stage response rates. A stepwise variable selection procedure, with region and herd size forced into a logistic regression model and a significance level of 0.05 for other variables to enter and remain in the model, was used (Table 1). The logistic regression analysis demonstrated that there were some differences in final-stage response by region, but that differences in response by herd size were not significant. Increased nonresponse was associated with having only one breeding

[1] W.C. Losinger, L.P. Garber, B.A. Wagner and G.W. Hill, United States Department of Agriculture, Animal and Plant Health Inspection Service, Veterinary Services, Center for Epidemiology and Animal Health, 555 South Howes Street, Suite 200, Fort Collins, Colorado 80521 U.S.A.

season and not consulting a veterinarian to treat or diagnose disease during 1996. The potential use of the logistic-regression variables as auxiliary variables in creating cells to adjust weights for final-stage nonresponse was examined. Four categorization schemes for nonresponse weight adjustment were proposed:

1. The traditional region by herd size scheme with 15 cells.

2. Region by herd size except in the West, which was subdivided by the number of breeding seasons, for a grand total of 14 cells.

3. Subdividing the cells of option 2 (by either of the auxiliary variables) if the difference in response rate (between the two new subdivisions) was at least ten percent and at least 20 respondents remained in each cell. Two subdivisions occurred, which yielded a total of 16 cells.

4. Continuing the subdivision of categories, based on the greatest difference in response rate, until a minimum number of respondents (no fewer than 20) remained in each cell. This yielded a total of 24 cells.

Table 1
Results of Stepwise Logistic Regression to Identify Variables Associated With Nonresponse to the Final Stage of Data Collection for the National Animal Health Monitoring System's Beef '97 Study. Based on 1,190 Eligible Operations and 238 Nonrespondents

Variable/ Response	Parameter Estimate	*P*
Intercept	0.369	0.181
Region		
Northcentral	0.851	0.000
Southcentral	0.822	0.000
Central	2.062	0.000
Southest	1.164	0.000
West	1.000	
Number of beef cows		
1 - 49	0.299	0.106
50 - 99	0.146	0.151
100 +	1.000	
Number of breeding Seasons		
1	-.370	0.039
>1 or no set season	1.000	
A veterinarian was consulted to treat or diagnose disease in 1996		
Yes	0.441	0.005
No	1.000	

Adjustment factors for weights of final-stage respondents were computed by dividing the sum of second-stage weights for eligible operations by the sum of second-stage weights for final-stage respondents within each cell.

Since the establishment of cells for schemes 2 through 4 was based on variables which demonstrated the greatest differences in response rates, differences in adjustment factors increased for particular subcategories from scheme 1 to scheme 4. For example, for the first scheme, adjustment factors for the Western region were 1.897, 1.504 and 1.579 for the small, medium and large herd size categories respectively. For the second scheme, adjustment factors in the Western region were 1.334 for operations that did not have one defined breeding season, and 1.875 for operations that did have one defined breeding season. For the third scheme, operations in the West that had one defined breeding season were split into two cells based on whether they had used a veterinarian to diagnose or treat disease during 1996: operations that had indicated "yes" received a weight adjustment of 1.548, while operations that had indicated "no" received a weight adjustment of 2.326.

To investigate how well the proposed auxiliary variables might have related to overall management strategies, we selected additional variables from the first two stages of data collection, and, within each region, examined differences in these variables by herd size category, number of breeding seasons, and whether a veterinarian had been consulted to diagnose or treat disease during 1996. Table 2 presents some representative results for the Western region. Some herd-size differences existed in the percent of operations that had one set breeding season and the percent of operations that had consulted a veterinarian during 1996. However, the percent of operations that had consulted a veterinarian was practically identical for operations that had one set breeding season versus operations that did not have one set breeding season, and vice versa. In addition, the percent of operations that vaccinated heifers for brucellosis and the percent of operations that implanted calves with a growth promotant exhibited a wider range by herd size category than by the other two proposed auxiliary variables. Moreover, mean weaning age and mean calf death loss varied more by herd size than by either number of breeding seasons or by whether a veterinarian was consulted. Similar patterns were noticed for other regions.

Although herd size was not a statistically significant predictor of participation in the final stage of data collection for the NAHMS Beef '97 Study (table 1), herd size was found to be more highly related to a number of questionnaire variables than either of the additional proposed auxiliary variables which derived from the logistic regression analysis. Therefore, we utilized the tradional region by herd size category scheme to perform the nonresponse weight adjustment for the final stage of data collection for the NAHMS Beef '97 Study.

Table 2
For 261 Western-Region Operations Eligible to Participate in the Third and Final Phase of Data Collection for the United States National Animal Monitoring System's 1997 Beef '97 Study (August 1 through January 31, 1998), Responses to Selected Variables From the First two Phases of Data Collection by Auxiliary Variables Examined for Weight Adjustment for the Final Stage of Data Collection

	Variables selected from the first two stages of data collection					
Auxiliary variables proposed for weight adjustment for third-stage nonresponse	1	2	3	4	5	6
	Percent				Mean	
Number of beef cows						
1 - 49	69.2	50.8	63.1	15.4	215	6.3
50 - 99	69.2	59.6	80.8	26.9	232	3.9
100+	88.2	70.1	85.4	52.8	223	4.1
Number of breeding seasons						
1	–	62.3	69.8	17.0	223	5.1
>1 or no set season	–	63.5	81.3	43.8	223	4.5
A veterinarian was consulted to treat or diagnose disease in 1996						
Yes	79.2	–	69.8	28.1	222	4.5
No	80.0	–	84.2	44.2	223	4.6

Variables selected from the first two phases of data collection:

1 = Operations with one set breeding season

2 = Operations that consulted a veterinarian to treat or diagnose disease in 1996

3 = Operations that vaccinate any heifers for brucellosis

4 = Operations that implanted any calves with a growth promotant prior to or at weaning during 1996

5 = Average age (in days) of calves at weaning

6 = Percent of calves that died in 1996

Researchers using survey data depend on sample weights to produce population parameter estimates that are approximately unbiased. In the final stage of data collection for the NAHMS Beef '97 Study, a logistic regression analysis identified two variables that were superior to herd size as predictors of nonresponse in the final stage of data collection. However, these variables were generally inferior to herd size in differentiating how producers responded to a number of key questions related to operation management. Using these two variables to establish categories for weight adjustment for nonresponse could have reduced bias in estimates of parameters (from the third stage of data collection) with which they were correlated. However, estimates of parameters not correlated with these variables could have been distorted. Therefore, we chose the traditional approach of performing the nonresponse weight adjustment by region and herd size categories.

Identifying variables that are good predictors of panel nonresponse is a good practice in any multistage survey. Prior to using these variables to adjust weights for unit nonresponse, we recommend that survey statisticians first follow some procedures to determine the extent to which these variables are linked to other characteristics of those eligible to complete the survey. Adjusting the weights based solely on variables that prove to be good predictors of panel nonresponse could potentially result in warped population estimates if these variables are not also good predictors of data that nonrespondents would have provided on the survey instrument.

ACKNOWLEDGEMENTS

The authors are grateful to the National Agricultural Statistics Service who initially selected the sample and the National Agricultural Statistics Service enumerators who made the first on-farm contact; the federal and state veterinarians and animal health technicians who made the subsequent on-farm visits; and all of the eligible beef producers, both respondents and nonrespondents.

REFERENCE

RIZZO, L., KALTON, G., and BRICK, J. M. (1996). A comparison of some weighting adjustment methods for panel nonresponse. *Survey Methodology*, 22, 43-53.

Survey Methodology, June 2000
Vol. 26, No. 1, pp. 113-116
Statistics Canada

Local Unconditional Best Linear Unbiased Estimators: Applications to Survey Sampling

JULIET POPPER SHAFFER[1]

ABSTRACT

Survey statisticians frequently use superpopulation linear regression models. The Gauss-Markov theorem, assuming fixed regressors or conditioning on observed values of regressors, asserts that the standard estimators of regression coefficients are best linear unbiased. Shaffer (1991) showed that the Gauss-Markov theorem doesn't apply when the regressors are random if some aspects of the population distribution of the regressors are known, and introduced an alternative estimator with better properties than the standard estimator under some conditions. This paper derives some generalizations, and notes an optimality property (locally best linear unbiasedness) of the generalized alternative estimator. Implications for estimation in surveys are noted.

KEY WORDS: Regression analysis; Gauss-Markov theorem; Survey sampling; Unbiased estimation; Optimality; Best linear unbiased estimation.

1. INTRODUCTION

In the standard linear regression model for a sample of observations,

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \tag{1}$$

the matrix of regressors, $\boldsymbol{X}$, is assumed to be a known, fixed matrix. Shaffer (1991) showed that when $\boldsymbol{X}$ is assumed to be random, the Gauss-Markov theorem does not hold in general, and described an alternative estimator that is more accurate when $\boldsymbol{\beta}$ is close to zero. Shaffer gave two applications of her results, to estimates of $\boldsymbol{\beta}$ and associated population quantities in multivariate normal superpopulation models and to ratio estimation of population means and totals.

In the present paper, three generalizations of these results are derived.

(a) The results are generalized from a model in which the sample covariance matrix of the errors $\boldsymbol{\epsilon}$ is $\sigma^2\boldsymbol{I}$, where $\boldsymbol{I}$ is the $n \times n$ identity matrix, to the case in which the covariance matrix $\sum$ of $\boldsymbol{\epsilon}$ is $\sigma^2\boldsymbol{B}$, where $\boldsymbol{B}$ is a known, fixed positive-definite matrix, and to some situations in which $\boldsymbol{B}$ is random (since it is the covariance matrix of a randomly-selected sample of regressor values).

(b) A generalized estimator is derived that performs well when the coefficient vector $\boldsymbol{\beta}$ is close to any pre-specified coefficient vector $\boldsymbol{\beta}_0$.

(c) A condition is given for design-unbiasedness of estimators of population means and totals based on the generalized estimator of $\boldsymbol{\beta}$.

Some results under the general model (1) will be given first. Then, modifications that apply to the sample survey situation will be discussed.

Under Model (1) with $\sum = \sigma^2\boldsymbol{I}$, the Gauss-Markov theorem asserts that the sample estimator

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y}, \tag{2}$$

is a best linear unbiased estimator (BLUE) if $\boldsymbol{X}$ is regarded as a fixed matrix. If the rows of $\boldsymbol{X}$ are treated as realizations of random vectors $x_i, i = 1, ..., n$, the Gauss-Markov theorem can be interpreted as an assertion that the estimator in (2) has minimum variance in the class of estimators linear in $\boldsymbol{Y}$ and conditionally unbiased, given these realized values of $\boldsymbol{X}$. However, the use of the term "unbiased" without qualification generally means unconditional unbiasedness. If the requirement of unbiasedness is interpreted to mean unbiased unconditionally, *i.e.*, on the average over random vectors with values in $\boldsymbol{X}$, Shaffer (1991) showed that the Gauss-Markov theorem doesn't apply when $\mathrm{E}(\boldsymbol{X}'\boldsymbol{X})$ is known. In that case, the conditionally biased estimator

$$\hat{\boldsymbol{\beta}}^* = \left[\mathrm{E}(\boldsymbol{X}'\boldsymbol{X})\right]^{-1}(\boldsymbol{X}'\boldsymbol{Y}) \tag{3}$$

is unconditionally unbiased and has smaller variance than $\hat{\boldsymbol{\beta}}$ when $\boldsymbol{\beta}$ is small. In fact, when $\mathrm{E}(\boldsymbol{X}'\boldsymbol{X})$ is known, no BLUE exists.

Comparison of the variances of (2) and (3) under various modeling assumptions, aside from the implications for estimating the coefficients themselves, gives insight into the conditions under which various estimators of other parameters of the populations have desirable properties, both model-based and design-based.

[1] Juliet Popper Shaffer, University of California, Department of Statistics, 367 Evans Hall, #3860, Berkeley, CA 94720-3860, U.S.A.
E-mail: shaffer@berkeley.edu.

2. GENERALIZATION OF THE COVARIANCE MATRIX OF ϵ

If the covariance matrix of ϵ is of the form $\sigma^2 \boldsymbol{B}$, where $\boldsymbol{B}$ is a known, fixed positive-definite matrix, the Gauss-Markov theorem applies to the generalized estimator

$$\hat{\boldsymbol{\beta}} = \left[\boldsymbol{X}'\boldsymbol{B}^{-1}\boldsymbol{X}\right]^{-1}\boldsymbol{X}'\boldsymbol{B}^{-1}\boldsymbol{Y}. \tag{4}$$

The proofs in Shaffer (1991) generalize directly to show that, if $\mathrm{E}(\boldsymbol{X}'\boldsymbol{B}^{-1}\boldsymbol{X})$ is known, the estimator

$$\hat{\boldsymbol{\beta}}^{*} = \left[\mathrm{E}(\boldsymbol{X}'\boldsymbol{B}^{-1}\boldsymbol{X})\right]^{-1}\boldsymbol{X}'\boldsymbol{B}^{-1}\boldsymbol{Y} \tag{5}$$

has smaller variance than (4) when $\boldsymbol{\beta}$ is sufficiently close to zero. The (unconditional) variances of (4) and (5) are

$$\sum_{\hat{\boldsymbol{\beta}}} = \mathrm{E}\left[(\boldsymbol{X}'\boldsymbol{B}^{-1}\boldsymbol{X})^{-1}\right]\sigma^2 \tag{6}$$

and

$$\begin{aligned}\sum_{\hat{\boldsymbol{\beta}}^{*}} &= \left[\mathrm{E}(\boldsymbol{X}'\boldsymbol{B}^{-1}\boldsymbol{X})\right]^{-1}\sigma^2 \\ &\quad + \mathrm{Var.}\left\{\left[\mathrm{E}(\boldsymbol{X}'\boldsymbol{B}^{-1}\boldsymbol{X})\right]^{-1}(\boldsymbol{X}'\boldsymbol{B}^{-1}\boldsymbol{X})\boldsymbol{\beta}\right\}.\end{aligned} \tag{7}$$

When $\boldsymbol{\beta} = 0$, Shaffer shows that (7) is smaller than (6), and therefore, assuming continuity of (7) as a function of $\boldsymbol{\beta}$, it is smaller than (6) when $\boldsymbol{\beta}$ is in a neighborhood of zero.

The results will now be applied in the sample survey context. Let $\boldsymbol{X}_N$ refer to the $N \times p$ matrix, and $\boldsymbol{Y}_N$ to the $N \times 1$ vector, in a finite population. If the N population elements are considered to be a sample from an infinite hypothetical population of potential elements satisfying (1), and if a sample of size n of the finite population is taken, the proofs in Shaffer (1991) generalize directly to show that

$$\hat{\boldsymbol{\beta}}_N^{*} = \left[\mathrm{E}(\boldsymbol{X}_N'\boldsymbol{B}_N^{-1}\boldsymbol{X}_N)\right]^{-1}\boldsymbol{X}_N'\boldsymbol{B}_N^{-1}\boldsymbol{Y}_N \tag{8}$$

and

$$\hat{\boldsymbol{\beta}}_n^{*} = \left[\mathrm{E}(\boldsymbol{X}_n'\boldsymbol{B}_n^{-1}\boldsymbol{X}_n)\right]^{-1}\boldsymbol{X}_n'\boldsymbol{B}_n^{-1}\boldsymbol{Y}_n \tag{9}$$

have variances smaller than those of their corresponding conditional versions $\hat{\boldsymbol{\beta}}_N$ and $\hat{\boldsymbol{\beta}}_n$ respectively, if $\boldsymbol{\beta}$ is close to zero, where the expectation in (8) is over the infinite population of hypothetical elements, and the expectation in (9) is over either the same infinite population or over the finite population of N elements satisfying (1). In order to apply these results, the expectations in (8) and (9) have to be known.

If $\boldsymbol{X}_N$ is to be regarded as fixed, the population model can be written as

$$\boldsymbol{Y}_N = \boldsymbol{X}_N\boldsymbol{\beta} + \epsilon_N, \tag{10}$$

where ϵ_N is a vector of randomly distributed error terms as in (1). Under Model (10), $\hat{\boldsymbol{\beta}}_N$ and $\hat{\boldsymbol{\beta}}_N^{*}$ are identical, but $\hat{\boldsymbol{\beta}}_n$ is still distinct from $\hat{\boldsymbol{\beta}}_n^{*}$. Under Model (10), for a random sample of size n, if

$$\mathrm{E}\left[(\boldsymbol{X}_n'\boldsymbol{B}_n^{-1}\boldsymbol{X}_n)/n\right] = (\boldsymbol{X}_N'\boldsymbol{B}_N^{-1}\boldsymbol{X}_N)/N, \tag{11}$$

the alternative estimator can be written in the form

$$\hat{\boldsymbol{\beta}}_n^{*} = \left[(n/N)(\boldsymbol{X}_N'\boldsymbol{B}_N^{-1}\boldsymbol{X}_N)\right]^{-1}\boldsymbol{X}_n'\boldsymbol{B}_n^{-1}\boldsymbol{Y}_n. \tag{12}$$

In model (10), Equations (11) and (12) will apply if $\boldsymbol{B}_N$ is diagonal and the sampling plan is self-weighting, and under some other conditions and sampling plans, *e.g.*, if $\boldsymbol{B}_N$ is block (cluster) diagonal and complete clusters are sampled. If $\boldsymbol{B}_N$ is diagonal, $\boldsymbol{B}_n$ is not necessarily fixed. For example, suppose a population consists of both men and women, and the variances of the two sexes on the characteristic of interest are known and are different. In that case, if a self-weighting sample is taken, and Model (10) is assumed to hold in both subpopulations, $\boldsymbol{B}_n$ will be diagonal, with entries that are a function of the proportions of the two genders in the sample.

3. LOCALLY BEST LINEAR UNBIASED ESTIMATION

Under the model (1), the estimator (5) is the locally best linear unbiased estimator (LBLUE) when $\boldsymbol{\beta} = 0$; *i.e.*, the estimator, linear in $\boldsymbol{Y}$ and unbiased for $\boldsymbol{\beta}$ with smallest variance in a neighborhood of $\boldsymbol{\beta} = 0$. Furthermore, the generalized linear estimator

$$\hat{\boldsymbol{\beta}}_{(\boldsymbol{\beta}_0)}^{*} = \boldsymbol{\beta}_0 + \left[\mathrm{E}(\boldsymbol{X}'\boldsymbol{B}^{-1}\boldsymbol{X})\right]^{-1}\left[\boldsymbol{X}'\boldsymbol{B}^{-1}(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}_0)\right], \tag{13}$$

allowing for the addition of a constant, is the LBLUE at $\boldsymbol{\beta} = \boldsymbol{\beta}_0$, for an arbitrary vector $\boldsymbol{\beta}_0$. The proof of these results in given in Appendix A. This generalized estimator (13) could be useful in a survey sampling situation in which it was reasonably sure that $\boldsymbol{\beta}$ would be close to some specified value. The variance of (13) is easily shown to equal (7) with $(\boldsymbol{\beta} - \boldsymbol{\beta}_0)$ substituted for $\boldsymbol{\beta}$. (See Appendix A.) Under Model (10) estimators (8), (9), and (12) generalize to

$$\hat{\boldsymbol{\beta}}_{(\boldsymbol{\beta}_0,N)}^{*} = \boldsymbol{\beta}_0 + \left[\mathrm{E}(\boldsymbol{X}_N'\boldsymbol{B}_N^{-1}\boldsymbol{X}_N)\right]^{-1}\left[\boldsymbol{X}_N'\boldsymbol{B}_N^{-1}(\boldsymbol{Y}_N - \boldsymbol{X}_N\boldsymbol{\beta}_0)\right], \tag{14}$$

$$\hat{\boldsymbol{\beta}}_{(\boldsymbol{\beta}_0,n)}^{*} = \boldsymbol{\beta}_0 + \left[\mathrm{E}(\boldsymbol{X}_n'\boldsymbol{B}_n^{-1}\boldsymbol{X}_n)\right]^{-1}\left[\boldsymbol{X}_n'\boldsymbol{B}_n^{-1}(\boldsymbol{Y}_n - \boldsymbol{X}_n\boldsymbol{\beta}_0)\right], \tag{15}$$

and

$$\hat{\boldsymbol{\beta}}_{(\boldsymbol{\beta}_0,n)}^{*} = \boldsymbol{\beta}_0 + \left[(n/N)\boldsymbol{X}_N'\boldsymbol{B}_N^{-1}\boldsymbol{X}_N)\right]^{-1}\left[\boldsymbol{X}_n'\boldsymbol{B}_n^{-1}(\boldsymbol{Y}_n - \boldsymbol{X}_n\boldsymbol{\beta}_0)\right], \tag{16}$$

respectively.

4. CONDITIONS FOR DESIGN UNBIASEDNESS

Assume the Model (10) holds, and that the unconditionally unbiased estimator can be expressed in the form (16). Suppose there exists a $p \times 1$–vector g such that $B_N^{-1} X_N g = \mathbf{1}_N$ and, for every sample of size n, $B_n^{-1} X_n g = \mathbf{1}_n$, where $\mathbf{1}_N$ and $\mathbf{1}_n$ are vectors of ones of length N and n, respectively. Then, given a simple random sample,

(a) the estimator

$$\hat{\bar{Y}}_{(\beta_{0,n})} = \bar{X}_N' \hat{\beta}^*_{(\beta_{0,n})} \tag{17}$$

is a design-unbiased estimator of $\bar{Y}_N$, where $\bar{X}_N' = (1/N)\mathbf{1}_N' X_N$, and

(b) $\hat{\bar{Y}}_{(\beta_{0,n})}$ is a generalized difference estimator of $\bar{Y}_N$.

The proof is given in Appendix B.

Note that a vector g satisfying the conditions of this theorem exists if the model includes an intercept (*i.e.*, X_N includes a column of ones) or if B_N is diagonal and the variance is proportional to the values of one of the regressors. Many applications of regression modeling to sample survey estimation are based on models that incorporate these assumptions. Särndal, Swensson and Wretman (1991, p. 231 and 232) discuss these and more general models, and Chapter, 6, section 4 of that reference has examples of commonly applied models incorporating these assumptions. Chapter 6 as a whole discusses both the general difference estimator of $N\bar{Y}_N$ and the analogous general regression estimator based on $\hat{\beta}_n$. The material in that Chapter also suggests generalizations of these results to more complex estimators and sampling plans.

5. DISCUSSION

To apply the results to estimates of properties of a finite population, it will be assumed that the matrix B is diagonal or has the special block-diagonal form and associated sampling plan discussed above. From the results in section 3, it follows that the estimator (17) of $\bar{Y}_N$ has smaller variance than the estimator

$$\hat{\bar{Y}}_{(\hat{\beta}_n)} = \bar{X}_N \hat{\beta}_n \tag{18}$$

when β is close to β_0. Note that (18) can be written

$$\bar{Y}_{\hat{\beta}_n} = = \frac{1}{N}\left[\sum_{i \in s} X_i' \hat{\beta}_n + \sum_{i \notin s} X_i' \hat{\beta}_n\right], \tag{19}$$

and X_i' is the i-th row of X, and S is the set of elements in the sample. Royall (1970) showed that the best linear model-unbiased estimator of $\bar{Y}_N$ (unbiased conditionally on the obtained sample) is

$$\frac{1}{N}\left[\sum_{i \in s} Y_i + \sum_{i \notin s} X_i' \hat{\beta}_n\right]. \tag{20}$$

In some important cases, the first term in (20) is equal to the first term in (19), in which case (20) and (19) are identical. This will be true, for example, if $B = \sigma^2 I$ and the model (10) contains an intercept, or if $p = 1$ and B is diagonal with diagonal entries proportional to the values of the single regressor. In such cases, (20) and (19) are identical, and the design-unbiased and unconditionally-model-unbiased estimator (17) has a smaller expected squared discrepancy from $\bar{Y}_N$ than the best linear conditionally-model-unbiased estimator (20) when β is close to β_0. Furthermore, if the sampling fraction is negligible, (17) has smaller expected squared discrepancy than (20) when β is close to β_0, even without the requirement that the first terms of (20) and (19) be equal.

If $\hat{\beta}$ is replaced by $\hat{\beta}^*$ in (20), the resulting estimator is no longer unconditionally unbiased. It can be shown, however, using concepts of dependence (Lehmann, 1966) that under the conditions on B noted at the beginning of this section, the resulting estimator will have smaller expected squared discrepancy from $\bar{Y}_N$ than (20) and (19) even without the further restrictions noted in the previous paragraph.

6. CONCLUSION

Since the conditions under which the estimator (5) of β is more efficient than the estimator (4) are very restrictive, and the estimators of population characteristics based on (5) can be derived in other ways, the results given here may be of more theoretical than practical interest. The results do give additional insight into some situations in which simple estimators like the sample mean and the generalized difference estimator are more efficient in estimating the population mean than are ratio estimators, poststratified estimators, regression estimators and other complex estimators. The equations (6) and (7) for comparative variances of (4) and (5) provide an alternative method of comparing respective variances under different regression models and different values of β. Many of these results hold under very simple sampling plans, but it should be possible to generalize them to more complex, unequal probability sampling plans.

ACKNOWLEDGEMENTS

The author is grateful to the late Erik N. Torgersen, who suggested the generalized estimator with optimal properties, to Phillip S. Kott, whose suggestion led to the derivation of the design unbiasedness condition, and to anonymous referees for many valuable comments.

APPENDIX A

Proof that $\hat{\beta}^*_{\beta_0}$ is LBLUE at β_0

Assume model (1), with $\mathrm{Var}(Y|X) = \sigma^2 B$. (The general proof given here applies directly to the model (10) as well.) Consider the sample estimator

$$\hat{\beta}^*_{(\beta_0)} = \beta_0 + \left[\mathrm{E}(X'B^{-1}X)\right]^{-1} X'B^{-1}(Y - X\beta_0).$$

Let $\tau = \beta - \beta_0$ and $Z = Y - X\beta_0$. Then $\mathrm{E}(Z|X) = X\tau$, $\mathrm{Var}(Z|X) = \sigma^2 B$, and $\hat{\tau}^* = \left[\mathrm{E}(X'B^{-1}X)\right]^{-1} X'B^{-1}Z = \hat{\beta}^*_{(\beta_0)} - \beta_0$.

Thus, the properties of $\hat{\beta}^*_{(\beta_0)}$ at $\beta = \beta_0$ are the same as those of $\hat{\beta}^* = \hat{\beta}^*_{(0)}$ at $\beta = 0$, so without loss of generality it will be shown that $\hat{\beta}^*_{(0)}$ is LBLUE at $\beta = 0$. Also without loss of generality, it will be assumed that $B = I$.

Let $C'(X)Y$ be an arbitrary unconditionally-unbiased estimator of β, where $C(X)$ is a matrix of functions of X, of the same dimensions as X. The requirement of unconditional unbiasedness necessitates the restriction $\mathrm{E}[C'(X)X] = I$ (Shaffer 1991). Conditioning first on X and then using the expression for unconditional variance, the variance of $C'(X)Y$ is $\mathrm{E}[C'(X)C(X)]\sigma^2 + \mathrm{Var}(C'(X)X\beta)$. Since we are considering variance at $\beta = 0$, only the first term is nonzero. Letting $C'(X) = \left[\mathrm{E}(X'X)\right]^{-1}X'$, the variance of $\hat{\beta}^*$ is $\left[\mathrm{E}(X'X)\right]^{-1}\sigma^2$.

Let $\tilde{\beta}$ be an arbitrary unconditionally-unbiased estimator of the form $C'(X)Y$. Then $\mathrm{Var}(\tilde{\beta}) = \mathrm{Var}(\hat{\beta}^*) + \mathrm{Var}(\tilde{\beta} - \hat{\beta}^*) + 2\mathrm{Cov}(\hat{\beta}^*, \tilde{\beta} - \hat{\beta}^*)$, so $\mathrm{Var}(\hat{\beta}^*) \le \mathrm{Var}(\tilde{\beta})$ if $\mathrm{Cov}(\hat{\beta}^*, \tilde{\beta} - \hat{\beta}^*) \ge 0$, or if $\mathrm{Cov}(\hat{\beta}^*, \tilde{\beta}) \ge \mathrm{Var}(\hat{\beta}^*)$. An easy calculation, using the restriction $\mathrm{E}[C'(X)X] = I$, shows that $\mathrm{Cov}(\hat{\beta}^*, \tilde{\beta}) = \mathrm{Var}(\hat{\beta}^*)$, which proves that $\hat{\beta}^*_{(\beta_0)}$ is LBLUE at β_0.

APPENDIX B

Proof of the Result in Section 4

$$\begin{aligned}
\bar{X}'_N \hat{\beta}^*_{(\beta_0)} &= \bar{X}'_N \beta_0 + (1/N)\, 1'_N X_N \left[n(1/N)(X'_N B_N^{-1} X_N)^{-1}\right] \\
&\quad X'_n B_n^{-1}(Y_n - X_n\beta_0) \\
&= \bar{X}'_N \beta_0 + (1/n) g' X'_N B_N^{-1} X_N (X'_N B_N^{-1} X_N)^{-1} \\
&\quad X'_n B_n^{-1}(Y_n - X_n\beta_0) \\
&= \bar{X}'_N \beta_0 + (1/n) 1'_n (Y_n - X_n\beta_0) \\
&= \bar{X}'_N \beta_0 + \bar{Y}_n - \bar{X}'_n \beta_0. \qquad \text{(B.1)}
\end{aligned}$$

where B_N and B_n are the appropriate population and sample matrices, respectively. The final expression in (B.1) is the generalized difference estimator based on a value β_0 chosen independently of the sample. This proves part (b) of the result; since the difference estimator is unbiased for $\bar{Y}$ in a self-weighting sample, the result in (a) follows.

REFERENCES

LEHMANN, E.L. (1966). Some concepts of dependence. *Annals of Mathematical Statistics*, 37, 1137-1153.

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

SÄRNDAL, C.-E., SWENSSON, B., and WRETMAN, J. (1991). *Model Assisted Survey Sampling*. New York: Springer.

SHAFFER, J.P. (1991). The Gauss-Markov theorem and random regressors. *The American Statistician*, 45, 269-273.

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 4, 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.

The Canadian Journal of Statistics **La Revue Canadienne de Statistique**

CONTENTS TABLE DES MATIÈRES

Volume 27, No. 4, December/décembre 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 x 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.