C3

SURVEY METHODOLOGY

STATISTICS CANADA STATISTIQUE CANADA
JUL JUIL 7 1994
LIBRARY BIBLIOTHÈQUE

Catalogue 12-001

A JOURNAL
PUBLISHED BY
STATISTICS CANADA

June 1994
•
VOLUME 20
•
NUMBER 1

Statistics Canada Statistique Canada

Canada

SURVEY METHODOLOGY

A JOURNAL
PUBLISHED BY
STATISTICS CANADA

JUNE 1994 • VOLUME 20 • NUMBER 1

Published by authority of the Minister
responsible for Statistics Canada

© Minister of Industry,
Science and Technology, 1994

All rights reserved. No part of this publication may be reproduced,
stored in a retrieval system or transmitted in any form or by any
means, electronic, mechanical, photocopying, recording or otherwise
without prior written permission from Licence Services,
Marketing Division, Statistics Canada,
Ottawa, Ontario, Canada K1A 0T6.

June 1994

Price: Canada: $45.00
United States: US$50.00
Other Countries: US$55.00

Catalogue No. 12-001

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 and Statistical Theory and Methods Abstracts and is referenced in the Current Index to Statistics, and Journal Contents in Qualitative Methods.

MANAGEMENT BOARD

Chairman G.J. Brackstone

Members
B.N. Chinnappa
G.J.C. Hole
F. Mayda (Production Manager)
R. Platek (Past Chairman)
C. Patrick
D. Roy
M.P. Singh

EDITORIAL BOARD

Editor M.P. Singh, *Statistics Canada*

Associate Editors

D.R. Bellhouse, *University of Western Ontario*
D. Binder, *Statistics Canada*
M.J. Colledge, *Statistics Canada*
J.-C. Deville, *INSEE*
J.D. Drew, *Statistics Canada*
W.A. Fuller, *Iowa State University*
M. Gonzalez, *U.S. Office of Management and Budget*
R.M. Groves, *U.S. Bureau of the Census*
D. Holt, *University of Southampton*
G. Kalton, *University of Michigan*
A. Mason, *East-West Center*
D. Pfeffermann, *Hebrew University*
J.N.K. Rao, *Carleton University*
L.-P. Rivest, *Université Laval*
I. Sande, *Bell Communications Research, U.S.A.*
C.-E. Särndal, *Université de Montréal*
W.L. Schaible, *U.S. Bureau of Labor Statistics*
F.J. Scheuren, *U.S. Internal Revenue Service*
J. Sedransk, *State University of New York*
J. Waksberg, *Westat Inc.*
K.M. Wolter, *National Opinion Research Center*
A. Zaslavsky, *Harvard University*

Assistant Editors N. Laniel, M. Latouche, L. Mach and H. Mantel, *Statistics Canada*

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 of the two official languages, English or French to the Editor, Dr. M.P. Singh, Social 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) is $45 per year in Canada, US $50 in the United States, and US $55 per year for other countries. Subscription order should be sent to Publication Sales, Statistics Canada, Ottawa, Ontario, Canada K1A 0T6. A reduced price is available to members of the American Statistical Association, the International Association of Survey Statisticians, and the Statistical Society of Canada.

SURVEY METHODOLOGY

A Journal Published by Statistics Canada
Volume 20, Number 1, June 1994

CONTENTS

In This Issue 1

Small Area Estimation

M.P. SINGH, J. GAMBINO and H.J. MANTEL
Issues and Strategies for Small Area Data 3
Comment: W.A. FULLER 15
G. KALTON 18
Response from the Authors 21

D. HOLT and D.J. HOLMES
Small Domain Estimation for Unequal Probability Survey Designs 23

A.C. SINGH, H.J. MANTEL and B.W. THOMAS
Time Series EBLUPs for Small Areas Using Survey Data 33

J.G. KOVAR and E.J. CHEN
Jackknife Variance Estimation of Imputed Survey Data 45

D.S. TRACY and S.S. OSAHAN
Estimation in Overlapping Clusters with Unknown Population Size 53

N.G.N. PRASAD and J.E. GRAHAM
PPS Sampling over Two Occasions 59

R.R. SITTER and C.J. SKINNER
Multi-way Stratification by Linear Programming 65

W.A. FULLER, M.M. LOUGHIN and H.D. BAKER
Regression Weighting in the Presence of Nonresponse with Application to the 1987-1988 Nationwide Food Consumption Survey 75

E.A. STASNY, B.G. TOOMEY and R.J. FIRST
Estimating the Rate of Rural Homelessness: A Study of Nonurban Ohio 87

In This Issue

This issue of *Survey Methodology* opens with a special section on **Small Area Estimation**. The three papers in this special section consider the problem of domain estimation from a variety of perspectives. I would like to give special thanks to Jon Rao for coordinating the editorial work for this special section. One or two other papers on this topic, which were not yet ready for publication, may also appear in a later issue.

The first paper in the special section, by Singh, Gambino and Mantel, considers the problem of small area statistics from the perspective of survey design. They discuss the role of sample design features such as stratification, clustering and sample allocation in the production of small area statistics for both planned and unplanned domains. A short overview of current approaches to small area estimation is also included. The paper is followed by insightful comments by Fuller and Kalton and a response from the authors.

The paper by Holt and Holmes presents a model based approach to small area estimation that does not "borrow strength" from other domains, and which may be used when auxiliary totals and means are not available. Estimates of model parameters are combined with design based estimates of means or totals of covariates. Using an example from market research it is shown that the method can lead to significant gains in efficiency of estimates for small domains.

The last paper in the special section, by Singh, Mantel and Thomas, presents an empirical comparison of several different small area estimators using simulated sampling from a population of farms. It is shown that, in the context of repeated surveys, estimators based on time series models can perform better, with respect to both bias and mean squared error, than those based on models for a single time point.

Kovar and Chen present results of a simulation study in which they investigated statistical properties of the jackknife approach to variance estimation of imputed data sets. Under this approach, the variance due to imputation is incorporated in the variance estimator. Real data sets, four different imputation methods, simple random sampling and a uniform nonresponse mechanism were used. Performance under a stratified multistage design and a non-uniform nonresponse mechanism was also studied.

Tracy and Osahan propose ratio estimators associated with two sampling strategies for estimation of a population mean in overlapping clusters with unknown population size. While much work by several researchers is available on non-overlapping clusters in the literature, there are many practical sampling situations where one gets overlapping clusters. The first sampling strategy is an equal probability with replacement sampling scheme while the second strategy is an unequal probability sampling scheme.

Prasad and Graham extend the "Random Group Method" for sampling with probability proportional to size (PPS) to sampling over two occasions. They use for this purpose the information on a study variate observed on the first occasion to select the matched portion of the sample on the second occasion.

Sitter and Skinner show how linear programming may be used to find an optimal sample design in the context of a multi-way stratification. Their approach is compared to existing methods both by illustrating the sampling schemes generated for specific examples and by evaluating mean squared errors. Variance estimation is also considered.

Fuller, Loughin and Baker consider regression weighting in the presence of non-response. They exhibit conditions under which the regression estimator remains consistent in the presence of non-response, and discuss implications for the choice of regressor variables. The ideas are illustrated by application to the 1987-88 Nationwide Food Consumption Survey conducted by the Human Nutrition Information Service of the U.S. Department of Agriculture.

The paper by Stasny, Toomey and First gives a description of a survey conducted in 1990 to estimate the rate of rural homelessness in Ohio. The possible magnitude of the bias of the estimator is investigated by simulating sampling from a variety of synthetic populations. It is found that the bias is likely to be small compared to the standard deviation.

The Editor

Survey Methodology, June 1994
Vol. 20, No. 1, pp. 3–22
Statistics Canada

Issues and Strategies for Small Area Data

M.P. SINGH, J. GAMBINO and H.J. MANTEL[1]

ABSTRACT

This paper identifies some technical issues in the provision of small area data derived from censuses, administrative records and surveys. Although the issues are of a general nature, they are discussed in the context of programs at Statistics Canada. For survey-based estimates, the need for developing an overall strategy is stressed and salient features of survey design that have an impact on small area data are highlighted in the context of redesigning a household survey. A brief review of estimation methods with their strengths and weaknesses is also presented.

KEY WORDS: Sample design strategy; Design estimates; Model estimates.

1. INTRODUCTION

For decades, administrative records and censuses were the main sources of data used for policy and planning for both large and small areas. These are still the richest source of statistical data at small area levels in most countries. During the forties and fifties, however, as the reliance on sample surveys increased, survey based estimates complemented the traditional sources because they provide more timely and cost efficient statistical data in a variety of subject matter fields. Although designed to provide reliable estimates primarily at larger area levels such as national and provincial, increasingly such surveys are being used to meet the growing demands for more timely estimates for various types and sizes of domains. No technical problem arises as long as these domains are large enough (*e.g.*, age-sex groups, larger cities and sub-provincial regions) to yield estimates of acceptable reliability. If data are needed for small domains, however, particularly if such domains cut across design strata, special estimation problems arise and several methods have recently been proposed to deal with such problems.

The main message of this paper is to emphasize the need to look at the problem of small area data in its entirety. Small area needs should be recognized at the early stages of planning for large scale surveys. The sampling design should include special features that enable production of reliable small area data using design or model estimators. The handling of this growing challenge to statistical agencies at the estimation stage should be viewed as a last resort.

In section 2, we discuss data needs and the three main sources of socio-economic data in the Canadian context, namely, the census, administrative records and surveys. Section 3 identifies some technical issues regarding the three sources of data and highlights the problems of quality measures and their interpretation. Then a need for developing an overall strategy that includes the planning, designing and estimation stages in the survey process is highlighted in section 4. Two aspects of the design, namely, clustering in a multi-stage sample design and sample allocation are discussed. In section 5, we present some sample design options being incorporated during the current redesign of the Canadian Labour Force Survey, the largest monthly household survey conducted by Statistics Canada, with a view to enhancing the survey capacity to provide better quality small area data. The purpose of section 6 is to review the many different approaches to estimation for small areas. We also suggest some new estimators and provide comments on the strengths and weaknesses of various domain estimators. A cautious approach towards the use of model estimators is stressed.

2. INFORMATION NEEDS AND DATA SOURCES

As the country's national statistical agency, Statistics Canada plays an integral role in the functioning of Canadian society. While guaranteeing the confidentiality of individual respondents' data, the agency's information describes the economic and social conditions of the country and its people. Its economic, demographic, social and institutional statistics programs produce reliable data on many aspects of life at the national, provincial, and sub-provincial levels for use by federal and provincial governments, private institutions, academics and the media. With increases in the planning, administration and monitoring of social and fiscal programs at local levels, there has been increasing demand for more and better-quality data at these levels. Three major sources of social, socio-economic and demographic data with emphasis on small area statistics are briefly discussed below.

[1] M.P. Singh, J. Gambino and H.J. Mantel, Statistics Canada, 16th Floor, R.H. Coats Building, Tunney's Pasture, Ottawa, Ontario, Canada K1A 0T6.

Census of Population: The quinquennial census of population provides benchmark data and serves as the richest source of information, available every five years, for small areas and for various characteristics/domains/ target groups of policy interest such as ethnic minorities, disabled persons, youth and aboriginal peoples.

Administrative Records: Administrative records are an increasingly important source of statistical data. These are extensively used in the demographic field by statistical agencies to produce local area estimates (Schmidt 1952, Verma and Basavarajappa 1987). In certain areas, such as vital statistics, administrative records are the only source of information for production of statistics at various levels of aggregation. In others, the relative merits of administrative records compared to censuses or surveys as data sources in terms of timeliness and quality of data determine the manner and the extent to which these data sources are used. In addition to direct tabulations, administrative records are used in a number of programs as a source of supplementary information for use in improving the quality of survey-based estimates. They are also being used in the construction of sampling frames for conducting surveys. Examples at Statistics Canada include the Business Register and the Address Register of residential dwellings.

Like the census of population, administrative records are very rich in geographical detail, making them a useful source of information for small area statistics. They are available more frequently and, due to recent technological advances, they are becoming a more cost-effective data source. However, administrative data are based on definitions made for programmatic rather than statistical purposes and their content is limited. Details of a Statistics Canada program for integration and development of an administrative records system to produce statistical outputs are given by Brackstone (1987a, 1987b). Experiences in the use of administrative records in other countries are included in the conference proceedings edited by Coombs and Singh (1987).

Household Surveys Program: Household surveys have long been an important source of economic and social statistics at Statistics Canada. Surveys under this program may be placed in three groups, namely, (i) the Labour Force Survey, (ii) Special Surveys and Supplementary Survey Programs and (iii) Longitudinal/Cyclical Surveys. These surveys are briefly introduced below indicating the scope for small area statistics in general.

Starting as a quarterly survey in 1945, the Canadian Labour Force Survey (LFS) became a monthly survey in 1952. The information provided by the survey has expanded considerably over the years and currently it provides a rich and detailed picture of the Canadian labour market. In addition to providing national and provincial estimates the survey regularly releases estimates for subprovincial areas. Regular estimates of standard labour market indicators are also in great demand for small areas such as Federal Electoral Districts, Census Divisions and Canada Employment Centres. These estimates are used by both federal and provincial governments in monitoring programs and allocating funds and other resources among various political and administrative jurisdictions.

Because of cost considerations, the LFS is heavily used as a vehicle for conducting *ad hoc* and periodic surveys at the national and provincial levels in the form of supplementary or special surveys. In the case of supplements, the LFS respondents themselves are asked additional questions, whereas for special surveys a different set of households is selected using the LFS frame. Both special and supplementary surveys are usually sponsored by other government departments and are conducted on a cost-recovery basis. For these surveys, the demands for small area statistics differ greatly from survey to survey, and generally the demands seem to be less pressing than those from the LFS itself.

Statistics Canada conducts a General Social Survey (GSS) annually to serve, in a modest way, the growing data needs on topics of current social policy interest. The GSS program (Norris and Paton 1991) consists of five survey cycles, each covering a different core topic, repeated every five years. Because of the limited size of sample (10,000 households nationally) the focus of the GSS is on estimates at the national level and on analytical statistics.

Longitudinal/panel surveys are new in the Canadian context. Statistics Canada has started two longitudinal surveys that will enrich the household survey program greatly, namely, the Survey on Labour and Income Dynamics and the National Population Health Survey. Both are large scale panel surveys and they are already creating expectations for data at sub-provincial and local area levels.

3. ISSUES IN DOMAIN ESTIMATION

There are numerous policy and technical issues that need to be addressed in the provision of small area statistics. The seriousness of these issues may vary from agency to agency and from one application to the next within the same agency depending on data quality and release policies. These issues are relevant for national and provincial estimates, but they assume higher significance in the context of small area statistics. As Brackstone (1987a) notes "on the issue of small area data evaluation, it is worth noting that error in small area estimates may be more apparent to users than error in national aggregates... at a local area level, there will be critics quick to point out deficiencies... it is true that for small areas, where estimation is more difficult, scrutiny of estimates is also more intensive". Several research and developmental studies on small area estimation are described in two volumes, one edited by Platek *et al.* (1987), and the other by Platek and Singh (1986). For a recent overview of small area estimation techniques currently being used in United States federal statistical programs see U.S. Statistical Policy Office (1993).

Use of Administrative Records: Federal and provincial government policies are the prime factors that influence the supply as well as the demand for small area data in most situations. On the supply side, government program driven administrative records contain a wealth of statistical information that can be used to produce local area data. Examples of files being used in the Canadian context are: Family Allowance, Unemployment Insurance, Income Tax, Health, Education, Old Age Security. Income-related statistics are produced at the local area level on a regular basis. Any **change** in government policy and associated programs can have immediate impact, for better or worse, on the coverage, availability, timeliness or quality of statistics derived from the corresponding administrative records. On the demand side, as noted earlier, governments need local area data for planning, implementing and monitoring their policies.

Conceptual issues: Quite frequently, conceptual and definitional issues in a data series are confounded with sampling and estimation problems. For example, consider the Unemployment Insurance (UI) system in Canada. UI regulations stipulate different qualification and requalification periods depending on the unemployment rate in a given region such that regions with higher unemployment rates require shorter qualifying periods of continuous employment. The estimates of regional unemployment rates derived from the LFS are used in determining the eligibility for an individual to receive benefits. These local area estimates are thus continually under close scrutiny by the public and the media. Such scrutiny however refers more often to **conceptual** issues rather than estimation issues per se; aspects such as the treatment in the survey questionnaire of discouraged workers, lay-offs and job search methods are questioned.

Use of Models and Related Quality Measures: Domain estimates are produced for virtually all large scale surveys, and as long as design estimators, *i.e.*, approximately design-unbiased estimators are of acceptable quality, no problem arises. We consider two classes of design estimators. Following Schaible (1992), **direct** estimators refer to estimators which use values of the study variable only for the time period of interest and only from units in the domain (*e.g.*, the regression estimator with slope estimated using only data from the domain). Such estimators may, and often do, use information on one or more auxiliary variables from other domains or other time periods, and are design unbiased or approximately so. The second class of design estimators, **modified** direct estimators, may use information from other domains on both the auxiliary and the study variable but still retain the property of design unbiasedness or approximate unbiasedness (*e.g.*, the regression estimator with slope estimated using the whole sample). There is a growing literature on **indirect** (or **model**) **estimators**, that is, estimators which use information on both the study and auxiliary variables from outside the domain and/or the time period of interest without any reference to their design unbiasedness properties.

Most producers and users of survey data are accustomed to design estimators and the corresponding design-based inferences. They interpret the data in the context of repeated samples selected using a given probability sampling design, and use estimated design-based **cvs** (coefficients of variation-square root of design variance divided by the design estimate) as the measures of data quality. For situations where either domains are too small or the sampling design did not foresee production of small area estimates, the design estimates may lead to large design **cvs** and model estimates may be the only choice if the survey-based estimates have to be provided for individual domains. A major challenge for statisticians is how to estimate, compare and explain to the users the relative precision of estimates from a survey that produces a large number of estimates at the national, subnational and large and small domain levels, most using design estimators but a few using model estimators. The model-based **cvs** (square root of design variance of model estimate divided by the model estimate) may convey a completely different message and may be several times lower than the corresponding design-based **cvs** for the same small area and in many cases, lower than the design-based **cvs** for much larger areas.

For model estimators, it is usually straightforward to derive expressions for the corresponding mean square errors (*i.e.*, design variance + square of the design bias). Estimation of these expressions, with an adequate degree of reliability, is a different matter. If we follow the argument that the data (*e.g.*, sample size) for such domains are inadequate for producing design estimates, it is unlikely that they would be adequate for producing design estimates of the corresponding variances and biases. As the estimation of bias is relatively more difficult, some authors seek design consistent model estimators, implying perhaps that bias can be ignored. However, if the sample size within the domain is sufficiently large to make the model estimator consistent, then the design estimator itself should give reliable estimates for the domain. For model estimators, suggestions have been made to use estimates of average mean square error computed over all domains. As the need for estimates for different domains usually arises because these domains are thought to be different from each other, a challenging task is to explain why estimates from all such domains are given the same degree of reliability. Another possibility is to construct indirect model-based estimates of the variance and bias of the model estimators for **individual** domains. Finding suitable methods of estimating mean square error for individual domains should be a research priority. Another serious concern for survey practitioners is how to guard against model failures. This suggests a need for research into model validation for complex survey situations. Further, for model estimators that use data on study variables for periods other than the time period of interest, estimates of **change** over different time periods would be of questionable quality; see Schaible (1992). Also, model estimators that borrow strength from other domains in the larger area will suffer a similar drawback when comparing differences in the two domains within the large area.

Issue of Privacy: In order to construct rich data bases for providing small area statistics, it is sometimes necessary to combine census, survey and/or administrative records. This necessitates linkage of records obtained from different sources. However, given the public's concern about privacy, record linkages should be carried out only after careful examination of all their implications. Under the Statistics Act, Statistics Canada may have access to administrative records of other departments for statistical purposes. But even for statistical purposes, as Fellegi (1987) notes, "we should have rigorous and auditable review procedures to ensure that we only carry out record linkage where the resulting privacy invasion is clearly outweighed by the public good from the new statistical information".

4. NEED FOR AN OVERALL STRATEGY

Even though large scale surveys are designed primarily for national and provincial estimates, it is rare that the estimates from such surveys relate only to the national/provincial populations as a whole. That is, invariably, such surveys are used to produce estimates for various **cross**-classified domains and in some cases for **areal** domains (*e.g.*, subprovincial) as well. In many cases, no special attention is paid to achieving a desired level of precision at the domain level either at the design or the estimation stage as long as the reliability is (believed to be) within reasonable limits. Problems arise when the cross-classified domain refers to a rare subpopulation or when the areal domain refers to a small area in which case either no estimates are possible/available or the estimates are of questionable quality. In a number of cases, this may happen simply because not enough attention was paid to these needs at the start of the survey planning process. If small area data needs are to be served using survey data then there is a need to develop an overall strategy that involves careful attention to meeting these needs at the planning, sample design and estimation stages of the survey process. For discussion of the design and estimation aspects, we will classify domains into the following two types:

Planned domains: In sampling terms these are individual strata or groups of strata for which desired samples have been planned. In the Canadian context these are typically subprovincial regions, such as Economic Regions, Unemployment Insurance Regions, and Health Planning Regions. In other cases, such domains could be larger counties, districts or similar subprovincial regions.

Unplanned domains: These are areas that were not identified at the time of design and thus may cut across design strata. Such domains can be of any size and they may create special estimation problems.

Planning: As noted earlier, the data demands from continuing periodic surveys such as the LFS are relatively much higher than from *ad hoc* surveys. In the case of periodic surveys that are redesigned every five or ten years, a suitable strategy can be developed during survey redesigns, since, in such cases, statistical agencies are usually in a much better position to project future small area data needs based on past demands. For *ad hoc* surveys, designers should include the establishment of such needs as an integral part of objective setting for the survey. Thus, in both cases, survey designers should establish the desired degree of precision, not only for national and provincial level estimates, but also for the domains of interest.

The first step of a strategy, in terms of the provision of small area data, will depend on the extent to which domains are identified in advance so that they can be treated as planned domains at the time of the design (or redesign) of the survey. If budgetary considerations do not permit reliable estimates for certain very small domains, then the option of either collapsing domains, pooling estimates over different surveys or not providing the estimates at all should be given serious consideration by survey designers in discussions with the survey sponsors. Some domains cannot be determined in advance. These unplanned domains should be handled through special estimation methods.

Sample design: In practice, it is rare that a design is optimal either for the national or provincial levels or for a single subject matter of interest. Usually varying degrees of compromise are introduced at different stages of sampling and the data collection process to satisfy theoretical and operational constraints. Depending on the data needs, estimates for domains should also form an integral part of this compromise. We will discuss two ways of taking small area data needs into account at the design stage, namely, sample allocation and the degree of clustering of the sample.

Allocation Strategy: In general, an optimum allocation strategy for national level estimates allocates samples to provinces approximately in proportion to their population. The reliability of estimates for smaller provinces in such cases suffers. Therefore a compromise allocation is usually preferred. There are different ways in which this compromise can be achieved depending on the emphasis placed on subnational estimates. Small reductions in sample sizes for larger provinces usually have little effect or the reliability of data for such provinces (or the national level data) but the corresponding sample increase in smaller provinces has significant impact on the reliability of their data.

The same principle holds for planned domains within the provinces. This is because optimum allocations in most situations are flat and the designers can exploit this feature by reallocating sample from the larger areas to planned domains that are smaller in size.

Clustering: Large scale household surveys usually involve stratified multistage designs with relatively large primary sampling units in order to make the design cost-efficient for national and provincial statistics. Such designs are thus highly clustered and, therefore, detrimental to the production of statistics for unplanned areal domains in the sense that, due to chance, some domains may be sample-rich while others may have no sample at all. Given the importance of domain estimates, attempts should be made to minimize the clustering in the sample. The following factors are important in this context: choice of frame, choice of sampling units and their sizes, number and size of strata and stages of sampling. The goal should be to make the design effects as low as possible given the operational constraints.

Estimation: No matter how much attention is paid to domain estimates at the early stages of planning and designing a particular survey, there will always be some smaller domains for which special estimation methods will be required for producing adequate estimates. Recently, synthetic estimators, which borrow strength from domains that resemble the domain of interest, have attracted a good deal of attention. However, since synthetic estimators are very sensitive to the assumption that domains resemble each other, even a small departure from the assumption can make the design bias high and put their use in question. Probability samplers, conscious of design bias, have suggested combinations of direct and synthetic estimators, with a view to addressing the design bias problem while trying to retain the strengths of the synthetic estimator. Empirical Bayes and similar techniques have been used to assign a weight to each component in the combined estimators. A brief review of these developments is given in section 6 on estimation.

5. SAMPLE DESIGN CONSIDERATIONS

5.1 Introduction

The small area problem is usually thought of as one to be dealt with via estimation. However, as was noted in the previous section, there are opportunities to be exploited at the survey design stage. This section uses the Canadian Labour Force Survey (LFS) to illustrate this.

The current LFS design: The Canadian Labour Force Survey is a monthly survey of 59,000 households which are selected in several stages using various methods. The ultimate sampling unit, the household, remains in the sample for six months once it is selected and is then replaced. Higher stage units (primary sampling units (PSU), clusters) also rotate periodically. Each of Canada's ten provinces is divided into economic regions (ER) which the LFS further divides into self-representing areas (medium and large cities) and non-self-representing areas (the rest of the ER). Stratification and sample selection take place within these areas, and the number of stages of sampling as well as the units of sampling differ between these two types of area. For example, in areas outside cities, there are three stages of sampling, whereas there are only two in the cities. For a detailed description of the current LFS design, refer to Singh *et al.* (1990).

5.2 Sampling Stages and Sampling Units

Area frames are usually associated with clustered sampling, *i.e.*, the first-stage units of selection are typically land areas containing a number of second-stage units. If a list of the second-stage units becomes available, then sampling directly from the list becomes possible, leading to a less clustered sample. This will result not only in improved estimates (due to lower design effects) but also in better small area estimates for unplanned domains. The latter holds since, by spreading the sample more evenly, it is more likely that an unplanned areal domain will contain some selected units. In contrast, in a clustered design we are often faced with a situation where one domain has sufficient sample because it happens to contain sampled clusters while a similar domain happens to have too few or no sampled clusters to produce good estimates.

To reduce clustering in the LFS we investigated two options: (i) the possibility of replacing the area frame (with its two stage design) in the larger cities with a list frame using the Address Register and (ii) reducing the sampling stages in rural areas and smaller urban centres. The Address Register, created to improve the coverage of the 1991 Canadian census (Swain, Drew, Lafrance and Lance 1992), consists of a list of addresses, telephone numbers and geographical information for dwellings by census enumeration area (EA). One option involved the selection of a stratified simple random sample of dwellings from the Address Register frame. This sample could then be supplemented with a sample selected from a growth frame which comprises a set of dwellings that are not in the post-censal address register. Handling of growth became the major stumbling block in pursuing option (i) as no cost-efficient method could be devised and tested in time for the current redesign. However, an updating strategy for the post-censal Address Register is still being investigated for future censuses and surveys.

With regard to option (ii), in keeping with the idea that less clustering is better for small area estimates, changes in the units and reduction in the stages of sampling were investigated for the areas outside the cities. Due to the changes that have taken place in data collection techniques, namely, from face-to-face interviewing to telephone and computer assisted interviewing, the cost-variance analyses from the past are no longer relevant. More than 80 percent of LFS interviews are now conducted by telephone. With the increase in telephone interviewing and the resulting decrease in travel, it became feasible in almost all cases to eliminate the current PSU stage and to sample EAs directly.

5.3 Stratification

One approach to stratification, similar in spirit to the above discussion on PSU size, is to replace large strata by many small ones. The hope is that a redefined domain or an unplanned domain will contain mostly complete strata. This will make the sample size in the domain more stable.

There may be several overlapping areas for which estimates are required. For example, each Canadian province is partitioned into both Economic Regions (ER) and Unemployment Insurance regions (UIR). One way to deal with this situation is to treat all the areas created by the intersections of the partitions as strata. In the Canadian case, for example, the 71 ERs and 61 UIRs yield 133 intersections, a manageable number. In some cases, however, the number of intersections may be too large to handle effectively. In addition, some of the intersections may have very small populations, making them unusable as strata.

By combining decreased clustering with smaller strata, we hope to have a design which is better able to meet small area needs. For example, the design should provide more flexibility in satisfying both ER and UIR requirements efficiently and in dealing with future changes in the definition of regions.

5.4 Allocation

If the definitions of small areas are known in advance, we may be able to treat them as planned domains and take them into account when designing the survey. The survey designer may endeavour to allocate sufficient sample in each small area to make the production of reliable estimates feasible. For large surveys such as the Canadian Labour Force Survey, this approach can, at least in theory, make the production of a great many small area estimates feasible. With a monthly sample of 59,000 households, and assuming that, say, 100 households per month are needed to produce reliable quarterly estimates, the country can be divided into about 600 non-overlapping areas, each guaranteed to have sufficient sample. Unions of such areas will also have enough sample to produce reliable monthly estimates.

Various sample allocation strategies are possible. In a top-down approach, once a provincial sample size is determined, the sample is allocated among the sub-provincial regions. However, it may turn out that it is not possible to satisfy the requirements for the reliability of sub-provincial estimates for the given provincial sample size. In a bottom-up strategy, the sample would be allocated to sub-provincial regions first in such a way that reliability objectives for each region are satisfied. As a result, we would expect comparable sample sizes in each sub-provincial region. This approach may result in a provincial sample size that is bigger than the one specified in the top-down approach. Regardless of which of the two strategies is used, adjustments to the initial allocations will usually be required. The resulting allocation will likely resemble a compromise between proportional allocation and equal allocation. In practice, the survey designer must perform a complex juggling act among provincial reliability requirements, sub-provincial requirements for one or more sets of regions, total survey costs and in-the-field details.

The approach taken in the current LFS redesign may be useful in other surveys as well. The sample was allocated in two steps: first, a core sample of 42,000 households was allocated to produce good estimates at the national and provincial levels; then the remaining sample was allocated to produce the best possible sub-provincial estimates. The resulting compromise allocation will produce reliable estimates for almost all planned domains. The compromise resulted in only minor losses at the provincial level and substantial gains at the subprovincial level. For example, the expected CVs for 'unemployed' for Ontario and Quebec are 3.2 and 3.0 per cent, respectively, instead of 2.8 and 2.6. The corresponding figures for Canada are 1.51 and 1.36. Optimizing for the provincial level yields CVs as high as 17.7 per cent for UI regions. With the compromise allocation, the worst case is 9.4 per cent.

Sample redistribution: There is usually some scope for moving sample from one area to another. For example, reducing the sample size by 1,000 households in a large province and making a corresponding increase in a small province will cause a marginal deterioration in the quality of provincial estimates in the former but will improve the estimates in the latter significantly. Similar movements of sample can be attempted within province.

5.5 Other Considerations

Change in definitions of small areas: Survey designers are faced with the fact that the definitions of planned domains may change during the life of a design and they may then have to treat the new domains as unplanned domains. For example, it is quite possible that the definitions of Unemployment Insurance Regions will change two or three years after the new LFS design is introduced in 1995. To deal with this at the design stage, the best that the survey designer can do is to choose as building blocks areas which are standard (*e.g.*, census-defined areas whose definitions are fairly stable) and hope that the redefined regions are unions of these standard areas. This is the approach that was taken in the current LFS redesign.

An alternative is to adopt an update strategy. This entails a reselection of units, doing it in such a way that the overlap between the originally selected units and the newly selected ones is maximized. By taking this approach, the number of new units that have to be listed is minimized. This also minimizes other field disruptions such as the need to hire new interviewers.

6. ESTIMATION

The purpose of this section is to review some of the different approaches to estimation of totals for small areas. No attempt is made to provide an exhaustive review; the discussion indicates the trend of developments in small area estimation research. For a detailed review, see the recent paper by Ghosh and Rao (1993). To facilitate this review we will classify small area estimation methods into two types. This is just one of many possible classification schemes. The first class of estimators we call design estimators, *i.e.*, (approximately) design unbiased estimators, which includes direct and modified direct estimators. As noted earlier, design estimators are often unsatisfactory, having a large variance due to small sample sizes (or even no sample at all) in the small areas. The second class we call indirect (or model) estimators, and it includes synthetic and combined estimators. Some of these estimators are compared empirically in an earlier version of this paper by Singh, Gambino and Mantel (1992).

6.1 Design Estimators

Direct Estimators: Direct small area estimators are based on survey data from only the small area, perhaps making use of some auxiliary data from census or administrative sources in addition to the survey data. The simplest direct estimator of a total is the expansion estimator,

$$\hat{Y}_{e,a} = \sum_{i \epsilon s_a} w_i y_i, \tag{6.1}$$

where s_a is the part of the sample in small area a and w_i is the survey weight for unit i. This estimator is unbiased; however, it may have high variability due to the random sample size in area a.

If the population size N_a is known then a post-stratified estimator,

$$\hat{Y}_{pst,a} = N_a \sum_{i \epsilon s_a} w_i y_i \Big/ \sum_{i \epsilon s_a} w_i = N_a \hat{Y}_{e,a} / \hat{N}_{e,a} = N_a \bar{y}_{e,a}, \tag{6.2}$$

may be used. This estimator is more stable than the expansion estimator; however, there may be some ratio estimation bias in complex surveys.

If the sampling scheme is stratified and the $N_{h,a}$ are known, where $N_{h,a}$ is the population size in stratum h and small area a, an alternative post-stratified estimator is $\hat{Y}_{st,pst,a} = \sum_h (N_{h,a} \sum_{i \epsilon s_{h,a}} w_i y_i / \sum_{i \epsilon s_{h,a}} w_i) = \sum_h N_{h,a} \hat{Y}_{h,e,a} / \hat{N}_{h,e,a} = \sum_h N_{h,a} \bar{y}_{h,a}$. The strata may also be post-strata instead of design strata.

Ratio estimation is similar to post-stratified estimation, the difference being that another auxiliary variable is used in place of the population counts N_a and $N_{h,a}$. For example, if x is a covariate for which the small area totals, X_a, or the stratum small area totals, $X_{h,a}$, are known then we may define the ratio estimators

$$\hat{Y}_{r,a} = X_a \hat{R}_a \quad \text{and} \quad \hat{Y}_{st,r,a} = \sum_h X_{h,a} \hat{R}_{h,a}, \tag{6.3}$$

where $\hat{R}_a = \hat{Y}_{e,a} / \hat{X}_{e,a}$ is an estimate of the ratio Y_a / X_a and $\hat{R}_{h,a} = \hat{Y}_{h,e,a} / \hat{X}_{h,e,a}$.

A regression estimator attempts to account for differences between small area subpopulation and subsample values of the covariates via an estimated regression relationship between the variate of interest, y, and the covariates, x. An advantage of regression type estimation is that it is easily extended to vector covariates. The estimator is given by

$$\hat{Y}_{reg,a} = \hat{Y}_a + \hat{\beta}_a (X_a - \hat{X}_a), \tag{6.4}$$

where $\hat{Y}_a$ may be an expansion or post-stratified estimator, $\hat{X}_a$ must be calculated in the same way as $\hat{Y}_a$, and $\hat{\beta}_a = \sum_{i \epsilon s_a} v_i^{-1} w_i y_i x_i' \{ \sum_{i \epsilon s_a} v_i^{-1} w_i x_i x_i' \}^{-1}$ where v_i are given weights for the regression. Note that $\hat{\beta}_a = \hat{R}_a$ when x is scalar and $v_i = x_i$. When $\hat{Y}_a$ and $\hat{X}_a$ are expansion estimators this estimator is also called the generalized regression estimator. Approximate design unbiasedness of this estimator follows from that of $\hat{Y}_a$ and $\hat{X}_a$.

As with the ratio type estimators, regression type estimation may also be applied within design strata or post-strata.

Modified Direct Estimators: Modified direct estimators may use survey data from outside the domain; however, they remain approximately design unbiased. By a modified direct estimator we mean a direct estimator with a synthetic adjustment for model bias; since the adjustment would have approximately zero expectation with respect to the design, the modified estimator is approximately design unbiased if the direct estimator is. An example is obtained by replacing $\hat{\beta}_a$ in (6.4) by a synthetic estimator $\hat{\beta} = \sum_{i \epsilon s} v_i^{-1} w_i y_i x_i' \{ \sum_{i \epsilon s} v_i^{-1} w_i x_i x_i' \}^{-1}$; we will denote this estimator by $\hat{Y}_{sreg,a}$. $\hat{\beta}$ would generally be more stable than $\hat{\beta}_a$; the choice between them would depend on the size of the variance of $\hat{\beta}_a$ relative to the variation in the $\beta_a s$ over areas a. A compromise is to take a weighted average $\lambda_a \hat{\beta}_a + (1 - \lambda_a) \hat{\beta}$ where λ_a is suitably chosen;

options for the choice of λ_a are discussed under combined estimators in Section 6.2. A second example is obtained by replacing $\hat{\beta}_a$ in (6.4) by $\hat{R} = \hat{Y}_e/\hat{X}_e$; note that $\hat{R}$ is a special case of $\hat{\beta}$ where x is scalar and $v_i = x_i$.

6.2 Indirect Estimators

Synthetic Estimators: Synthetic estimation methods are based on an assumption that the small area is similar in some sense to another area, often a larger area which contains it. Estimates for the other area would generally be more reliable than those for the small area. The resulting synthetic estimator would then have small variance, though it may be badly biased if the underlying assumption is violated.

One of the simplest synthetic estimators arises from the assumption that the small area mean is equal to the overall mean. This leads to the mean synthetic estimator

$$\hat{Y}_{syn,m,a} = N_a \sum_{i \epsilon s} w_i y_i \Big/ \sum_{i \epsilon s} w_i = N_a \bar{y}. \tag{6.5}$$

A more common synthetic estimator is based on stratification or post-stratification,

$$\hat{Y}_{syn,st,m,a} = \sum_h N_{h,a} \sum_{i \epsilon s_h} w_i y_i \Big/ \sum_{i \epsilon s_h} w_i = \sum_h N_{h,a} \bar{y}_h .$$

As with direct estimators, ratio synthetic estimation may be based on other auxiliary data besides the population counts N_a or $N_{h,a}$. For example, the common ratio synthetic estimators based on a covariate x are defined as

$$\hat{Y}_{syn,r,a} = X_a \hat{Y}_e / \hat{X}_e \quad \text{and} \quad \hat{Y}_{syn,st,r,a} = \sum_h X_{h,a} \hat{Y}_{h,e} / \hat{X}_{h,e}, \tag{6.6}$$

where $\hat{Y}_e = \sum_{i \epsilon s} w_i y_i$ is the expansion estimator of the population total for y and $\hat{Y}_{h,e} = \sum_{i \epsilon s_h} w_i y_i$. $\hat{X}_e$ and $\hat{X}_{h,e}$ are similarly defined. These estimators have been studied by Gonzalez (1973), Gonzalez and Waksberg (1973) and Ghangurde and Singh (1977, 1978), among others.

Singh and Tessier (1976) suggested an alternative ratio synthetic estimator, using X instead of $\hat{X}_e$, defined as

$$\tilde{Y}_{syn,r,a} = X_a \hat{Y}_e / X. \tag{6.7}$$

Both $\hat{Y}_{syn,r,a}$ and $\tilde{Y}_{syn,r,a}$ have the same synthetic bias and the ratio bias in $\hat{Y}_{syn,r,a}$ will be negligible for large samples. The choice between these two estimators depends on ρ, the correlation of $\hat{Y}_e$ and $\hat{X}_e$. It can be shown that for large samples $V(\hat{Y}_{syn,r,a}) \leq V(\tilde{Y}_{syn,r,a})$ if $\rho \geq 0.5 c_x / c_y$, where c_x and c_y are the coefficients of variation of $\hat{X}_e$ and $\hat{Y}_e$, respectively. In most cases, when ρ is high or the population is skewed, $\hat{Y}_{syn,r,a}$ would be preferred; however, when c_x is high and the correlation is only moderate, $\bar{Y}_{syn,r,a}$ may be the better choice.

In some situations information on a second auxiliary variable (z) in addition to x may be available. Then a bivariate ratio synthetic estimator may be constructed:

$$\hat{Y}^{(2)}_{syn,r,a} = \gamma_a X_a \hat{Y}_e / \hat{X}_e + (1 - \gamma_a) Z_a \hat{Y}_e / \hat{Z}_e, \tag{6.8}$$

where γ_a is suitably chosen. Extensions to a multivariate ratio synthetic estimator may be considered following Olkin (1958).

Regression synthetic estimation is similar to ratio synthetic,

$$\hat{Y}_{syn,reg,a} = \hat{\beta} X_a ,$$

$$\hat{\beta} = \sum_{i \epsilon s} v_i^{-1} w_i y_i x_i' \left\{ \sum_{i \epsilon s} v_i^{-1} w_i x_i x_i' \right\}^{-1}. \tag{6.9}$$

Again, regression synthetic estimation may also be applied within design strata or post-strata. Royall (1979) suggested a slight variation, $\hat{Y}_{syn,Roy,a} = \sum_{i \epsilon s_a} y_i + \hat{\beta}(X_a - \sum_{i \epsilon s_a} x_i)$, where the sum of y-values for only units not included in the sample is estimated synthetically.

> **Remark**: The examples of modified direct estimators presented in Section 6.1 can also be considered to be ratio or regression synthetic estimators with a design-based adjustment to correct for bias. For example, we may write $\hat{Y}_{sreg,a} = \hat{Y}_{syn,reg,a} + (\hat{Y}_a - \hat{\beta}\hat{X}_a)$ where $\hat{Y}_a - \hat{\beta}\hat{X}_a$ is an estimate of the bias of $\hat{Y}_{syn,reg,a}$. Similarly, $\hat{Y}_{sreg,a}$ can also be written as the Royall estimator, $\hat{Y}_{syn,Roy,a}$, with a design-based adjustment for bias.

Purcell and Kish (1980) discuss another type of synthetic estimation which they call SPREE (structure preserving estimation) for small area estimation of frequency data. Detailed historical counts, perhaps from a census, are combined with less detailed current survey estimates to produce detailed estimates of current counts. The assumption here is that certain relationships among the detailed counts are stable over time.

Combined Estimators: By a combined estimator we mean a weighted average of a design estimator and a synthetic estimator,

$$\hat{Y}_{com,a} = \lambda_a \hat{Y}_{des,a} + (1 - \lambda_a) \hat{Y}_{syn,a}, \tag{6.10}$$

where λ_a is suitably chosen. The aim here is to balance the potential bias of the synthetic estimator against the instability of the design estimator. There are three broad approaches which may be used to define the weights λ_a in (6.10); they may be fixed in advance, sample size dependent, or data dependent.

The first and simplest approach to weighting is to fix the weights in advance, for example, to take a simple average. However, this does not make any allowance for the actual observed reliability of the design estimator. For some realized samples the design estimator for small area a is more reliable than for other realized samples. The weight given to the design estimator should reflect this.

The second general approach to weighting of the design and synthetic parts is called sample size dependent, in which the weights are functions of the ratio $\hat{N}_{e,a}/N_a$. Another possibility, not considered here, is to base the weights on the realized sample values of a covariate x; for example, the weight could be a function of $\hat{X}_{des,a}/X_a$ or of $S^2_{x,a}/\sigma^2_{x,a}$ where $S^2_{x,a}$ is the realized variance of $\hat{X}_{des,a}$, conditional on $\hat{N}_{e,a}$ or some other relevant aspect of the realized sample, and $\sigma^2_{x,a}$ is the unconditional variance of $\hat{X}_{des,a}$.

Some specific estimators in this class have been proposed earlier. Drew, Singh, and Choudhry (1982) proposed the sample size dependent estimator

$$\hat{Y}_{ssd,r,a} = \lambda_a \hat{Y}_{r,a} + (1 - \lambda_a) \hat{Y}_{syn,r,a}, \qquad (6.11a)$$

where

$$\lambda_a = \begin{cases} 1 & \text{if} \quad \hat{N}_{e,a} \geq \delta N_a \\ \hat{N}_{e,a}/\delta N_a & \text{otherwise} \end{cases} \qquad (6.11b)$$

and δ is subjectively chosen to control the contribution of the synthetic component. Särndal (1984) suggested

$$\hat{Y}_{ssd,reg,a} = \lambda_a \hat{Y}_{sreg,a} + (1 - \lambda_a) \hat{Y}_{syn,reg,a}, \qquad (6.12)$$

where $\lambda_a = \hat{N}_{e,a}/N_a$. Rao (1986) suggested a modification to this in which λ_a would be taken to be 1 whenever $\hat{N}_{e,a} \geq N_a$. Särndal and Hidiroglou (1989) refined Rao's suggestion by taking $\lambda_a = (\hat{N}_{e,a}/N_a)^{h-1}$ when $\hat{N}_{e,a} < N_a$, where h is chosen judgementally to control the contribution of the synthetic component.

It is the bias of the synthetic component that is of concern when using these sample size dependent estimators in practice. The weight associated with the synthetic component should be such that the bias is kept within reasonable limits. For example, the sample size dependent estimator of Drew, Singh and Choudhry (1982), with generalized regression estimation replacing the ratio estimation and $\delta = 2/3$, is currently used in the Canadian Labour Force Survey to produce domain estimates. For a majority of domains the weight attached to the synthetic component is zero as the direct estimator itself provides the required degree of reliability. For other domains the weight attached to the synthetic component is about 10% on average and never exceeds 20%. Depending on the risk of bias that one is willing to take, δ may lie in the range [2/3,3/2] for most practical situations.

The third approach to weighting we call data dependent. The optimal weights for combining two estimators generally depend on the mean squared errors of the estimators and their covariance. These quantities would generally be unknown but may be estimated from the data. For our combined estimators this would usually require some modelling of the bias of the synthetic part. An early and well known example of this approach is due to Fay and Herriot (1979). They model the biases of the synthetic estimators for the small areas as independent random effects with an unknown but fixed variance. To be more specific, if $\hat{Y}_{des,a}$ is the design estimator then they consider the model $Y_a = X_a \beta + \alpha_a$ and $\hat{Y}_{des,a} = Y_a + \epsilon_a$ where $\alpha_a \sim (0,\sigma^2)$, $\epsilon_a \sim (0,\nu_a^2)$, and α_a and ϵ_a are independent and uncorrelated over a, σ^2 is unknown and ν_a^2 are assumed known (in practice they would need to be estimated). For a given value of σ^2 the optimal weights for combining $\hat{Y}_{des,a}$ and $X_a \hat{\beta}$ can be calculated. An estimate of σ^2 is obtained by the method of fitting constants and substituted into the optimal weights. Some protection against model mis-specification is obtained by truncating the resulting estimate if it deviates from the direct estimate by more than a specified multiple of ν_a. Schaible (1979) and Battese and Fuller (1981) also consider empirically estimated optimal weights λ_a in (6.12) based on similar random effects models for the small area totals.

Prasad and Rao (1990) provide an estimator of the mean square error of the Fay-Herriot estimator which makes allowance for the estimation of the variance components. Kott (1989) proposes a design consistent estimator of the mean square error, but finds it to be very unstable.

Another alternative is to use historical data to calculate the weights; this has the advantage that the weights may be more stable than if they are estimated from current survey data; however, there is an underlying assumption that the optimal weights are stable over time.

Remark: In sample size dependent estimation the weights are allowed to depend on the observed size of the subsample s_a, but not on the values of the variate of interest. This non-dependence of the weights on the variate of interest has advantages and disadvantages. An advantage is that the same weights would be used for estimation of totals for all variates of interest; they need to be calculated only once. More importantly, the estimate of the sum of two variables is the sum of the estimates of the two variables. A disadvantage is that the weights do not directly take account of either the reliability of the design estimator for the variate of interest or the likely magnitude of the bias of the synthetic estimator.

Combining data over time: For repeated surveys pooling of data over survey occasions to increase the reliability of estimates is a common practice. Depending on the rotation pattern used for such surveys, significant gains in reliability can be achieved. This pooling or averaging over time is thus of particular interest in the context of domain estimation where reliability is usually low. For domain estimation in the Canadian Labour Force Survey it is normal practice to use a sample size dependent estimator based on three month average estimates of employed and unemployed. Due to the six month rotation scheme used, as noted earlier, averaging over three months increases the sample size by one third. If samples completely overlap between periods then averaging does not result in any gain in efficiency. For other rotation patterns the sample size for domain estimates could be more than doubled through this process. There is, however, a conceptual problem with pooled estimates, in that such estimates refer to an average of the parameter of interest (*e.g.*, unemployment) over a period of, say, three months.

In composite estimation the current design estimator is combined with the composite estimator for the previous period, updated by an estimate of change based on the common sample. This idea was used, though not in the context of small area estimation, by Jessen (1942), and Patterson (1950), among others. Binder and Hidiroglou (1988) provide a review. The weights for the combination are typically estimates of the optimal weights under the assumption that these weights are time stationary. These data dependent weights have the disadvantage that they lead to inconsistency of estimates for different characteristics and their sums.

A recent development in small area estimation techniques is the use of time series methods for periodic surveys. The relationship between parameters of interest for different time periods is modelled and this model is exploited to improve the efficiency of the estimates for the current occasion. In most cases some allowance must also be made, through modelling or otherwise, for the non-independence of samples for different survey occasions due to the sample rotation scheme. Some references for this time series approach are Choudhry and Rao (1989), Pfeffermann and Burck (1990), Singh, Mantel and Thomas (1994) and Singh and Mantel (1991). All of these are generalizations of the Fay-Herriot model which allow the regression parameters, small area effects, and survey errors to evolve over time according to various time series models. The vector of small area estimates that results from this approach can be written as a weighted average of the vector of design estimates and a vector of synthetic estimates which are based on past data and the current values of covariates; however, the matrix of weights would not generally be diagonal so that the estimator for any single small area would generally depend also on the design estimates and synthetic estimates for other small areas.

7. CONCLUSION

To produce adequate survey-based domain estimates that are timely and up to date, sample designers must face several challenging tasks. The first is to convince the sponsors/program managers that some small area data needs cannot be met as a by-product of a system designed optimally for national/sub-national estimates. Significant gains, which may vary from survey to survey, can be achieved at the domain level at a marginal reduction in reliability at higher levels. There is a need to develop an overall strategy that incorporates desired reliability for the planned domains as well as for higher levels through compromise allocations, and reduced clustering to help improve estimates for unplanned domains. It should be noted that many of the planned domains at design time may become unplanned (revised) over time in the context of continuous surveys.

The overall strategy should also include consideration of both design estimators for larger domains and model estimators for small domains. A model estimator should be preferred over a design estimator only if its mean square error (design variance + bias^2) is estimable and it is sufficiently smaller than the corresponding variance of the design estimator. We should have estimates of mean square error for each of the individual domains. An option that statistical agencies can exercise is to pool similar domains or pool estimates over different time periods for the same domain. They may even suppress estimates for some domains on account of data reliability or privacy concerns.

The second challenging task for statisticians is to explain to users the different types of measures of reliability for different sets of estimates from the same survey. It is hoped that with more research on model validation and better estimates of mean square errors, designers will get more confidence in using model estimators for small domains. In the meantime model estimators should be used with caution even if they have significantly smaller coefficients of variation.

Censuses, supplemented by data from administrative records, are likely to remain the primary source of small area socio-economic data, especially for countries having a quinquennial census of population and housing. Also, concerns about problems with conceptual issues in the context of data for administrative records are likely to continue until statistical agencies are given an opportunity to influence the development of the forms used to collect such data. Until then, this immensely rich data source cannot be fully exploited for statistical purposes and more so for domain estimation.

ACKNOWLEDGMENT

We are grateful to Jon Rao for handling this paper as an Associate Editor and to the referees for many constructive suggestions.

REFERENCES

BATTESE, G.E., and FULLER, W.A. (1981). Prediction of county crop areas using survey and satellite data. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 500-505.

BINDER, D., and HIDIROGLOU, M.A. (1988). Sampling in time. *Handbook of Statistics*, (Eds. P.R. Krishnaiah and C.R. Rao). New York: Elsevier Science, 187-211.

BRACKSTONE, G.J. (1987a). Small area data: Policy issues and technical challenges. *Small Area Statistics*, (Eds. R. Platek, J.N.K. Rao, C.-E. Särndal and M.P. Singh). New York: Wiley and Sons, 3-20.

BRACKSTONE, G.J. (1987b). Statistical uses of administrative data: Issues and challenges. *Proceedings: Symposium on Statistical Uses of Administrative Data*, (Eds. J.W. Coombs and M.P. Singh), Statistics Canada, 5-16.

CHOUDHRY, G.H., and RAO, J.N.K. (1989). Small area estimation using models that combine time series and cross-sectional data. *Proceedings: Symposium on Analysis of Data in Time*, (Eds. A.C. Singh and P. Whitridge), Statistics Canada, 67-74.

COOMBS, J.W., and SINGH, M.P. (Eds.) (1987). *Proceedings: Symposium on Statistical Uses of Administrative Data*, Statistics Canada.

DREW, J.D., SINGH, M.P., and CHOUDHRY, G.H. (1982). Evaluation of small area estimation techniques for the Canadian Labor Force Survey. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 545-550.

FAY, R.E., and HERRIOT, R.A. (1979). Estimates of income for small places: An application of James-Stein procedures to census data. *Journal of the American Statistical Association*, 74, 269-277.

FELLEGI, I.P. (1987). Opening Remarks. *Proceedings: Symposium on Statistical Uses of Administrative Data*, (Eds. J.W. Coombs and M.P. Singh), Statistics Canada, 1-2.

GHOSH, M., and RAO, J.N.K. (1993). Small area estimation: An appraisal. To appear in *Statistical Science*.

GHANGURDE, P.D., and SINGH, M.P. (1977). Synthetic estimates in periodic household surveys. *Survey Methodology*, 3, 152-181.

GHANGURDE, P.D., and SINGH, M.P. (1978). Evaluation of efficiency of synthetic estimates. *Proceedings of the Social Statistics Section, American Statistical Association*, 53-61.

GONZALEZ, M.E. (1973). Use and evaluation of synthetic estimators. *Proceedings of the Social Statistics Section, American Statistical Association*, 33-36.

GONZALEZ, M.E., and WAKSBERG, J. (1973). Estimation of the error of synthetic estimates. Presented at the first meeting of the International Association of Survey Statisticians, Vienna, Austria.

JESSEN, R.J. (1942). Statistical investigation of a farm survey for obtaining farm facts. *Iowa Agricultural Station Research Bulletin*, 304, 54-59.

NORRIS, D., and PATON, D. (1991). Canada's General Social Survey: Five years of experience. *Survey Methodology*, 17, 227-240.

OLKIN, I. (1958). Multivariate ratio estimation for finite populations. *Biometrika*, 45, 154-165.

PATTERSON, H.D. (1950). Sampling on successive occasions with partial replacement of units. *Journal of the Royal Statistical Society*, Series B, 12, 241-255.

PFEFFERMANN, D., and BURCK, L. (1990). Robust small area estimation combining time series and cross-sectional data. *Survey Methodology*, 16, 217-237.

PLATEK, R., RAO, J.N.K., SÄRNDAL, C.-E., and SINGH, M.P. (1987). *Small Area Statistics*. Invited Presentations. New York: Wiley.

PLATEK, R., and SINGH, M.P. (1986). Small Area Statistics, An International Symposium' 85 (contributed papers), Technical Report Series of the Laboratory for Research in Statistics and Probability, Carleton University, University of Ottawa, Canada.

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.

PURCELL, N.J., and KISH, L. (1980). Postcensal estimates for local areas (or Domains). *International Statistical Review*, 48, 3-18.

RAO, J.N.K. (1986). Synthetic estimates, SPREE and best model based predictors. *Proceedings of the Conference on Survey Research Methodology in Agriculture*, American Statistical Association and National Agricultural Statistics Service, USDA, 1-6.

ROYALL, R.M. (1979). Prediction models in small area estimation. In *Synthetic Estimates for Small Area*, NIDA Research Monograph Series 24, U.S. Department of Health, Education and Welfare.

SÄRNDAL, C.-E. (1984). Design-consistent versus model-dependent estimators for small domains. *Journal of the American Statistical Association*, 79, 624-631.

SÄRNDAL, C.-E., and HIDIROGLOU, M.A. (1989). Small domain estimation: A conditional analysis. *Journal of the American Statistical Association*, 84, 266-275.

SCHAIBLE, W.L. (1979). A composite estimator for small area statistics. *Synthetic Estimates for Small Area*, NIDA Research Monograph Series 24, U.S. Department of Health, Education and Welfare, Library of Congress catalogue number 79-600067, 36-53.

SCHAIBLE, W.L. (1992). Use of small area estimators in U.S. federal programs. *Small Area Statistics and Survey Designs*, (Vol. 1), (Eds. G. Kalton, J. Kordos and R. Platek). Warsaw: Central Statistical Office, 95-114.

SCHMIDT, R.C. (1952). Short-cut methods for estimating county populations. *Journal of the American Statistical Association*, 47, 232-238.

SINGH, A.C., and MANTEL, H.J. (1991). State space composite estimation for small areas. *Proceedings: Symposium 91, Spatial Issues in Statistics*, Statistics Canada, 17-25.

SINGH, A.C., MANTEL, H.J., and THOMAS, B.W. (1994). Time series EBLUPs for small areas using survey data. *Survey Methodology*, 20, 33-43.

SINGH, M.P., DREW, J.D., GAMBINO, J.G., and MAYDA, F. (1990). *Methodology of the Canadian Labour Force Survey*. Catalogue No. 71-526, Statistics Canada.

SINGH, M.P., GAMBINO, J.G., and MANTEL, H. (1992). Issues and options in the provision of small area statistics. *Small Area Statistics and Survey Designs*, (Vol. 1), (Eds. G. Kalton, J. Kordos and R. Platek). Warsaw: Central Statistical Office, 37-75.

SINGH, M.P., and TESSIER, R. (1976). Some estimators for domain totals. *Journal of the American Statistical Association*, 71, 322-325.

SWAIN, L., DREW, J.D., LAFRANCE, B., and LANCE, K. (1992). The creation of a residential address register for coverage improvement in the 1991 Canadian Census. *Survey Methodology*, 18, 127-142.

U.S. STATISTICAL POLICY OFFICE (1993). Indirect Estimators in Federal Programs. Statistical Policy Working Paper 21. Prepared by the subcommittee on Small Area Estimation, Federal Committee on Statistical Methodology.

VERMA, R.B.P., and BASAVARAJAPPA, J.G. (1987). Recent developments in the regression method for estimation of population for small areas in Canada. *Small Area Statistics*, (Eds. R. Platek, J.N.K. Rao, C.-E. Särndal and M.P. Singh). New York: Wiley and Sons, 46-61.

COMMENT

W.A. FULLER[1]

The authors are to be congratulated on an excellent description of the design and estimation considerations associated with domains. The authors discuss estimation for planned domains, particularly situations in which domain membership can be identified in the frame, and estimation for unplanned domains including domains for which the domain membership cannot be determined from the frame. This is a fine contribution to the growing literature on domain estimation.

The authors give a particularly good description of the planning, data collection, and processing activities associated with surveys conducted by Statistics Canada. Included are the traditional design problems of balancing needs for domain estimation with desire for efficiency at higher levels, the importance of confidentiality in using administrative records in constructing domain estimates, and the importance of definitional compatibility in attempting to combine information from different sources.

The importance of considering domain estimation at the design stage is very well taken and is a point often ignored by authors concentrating on small area estimation. As the authors emphasize, careful design can often enable one to construct estimates for domains in a direct and design consistent manner. I am sure that those actually designing surveys have considered the importance of clustering when designing surveys that will be used for domain estimation, but it is pleasant to see an explicit discussion.

The authors describe several types of estimators for domains. Their classification emphasizes the number of alternatives available to the practitioner. It is possible to use the theoretical mean square errors to provide information on the relative merits of the estimators. As an example of such a comparison, assume a simple random sample of size n selected from a population divided into K domains. Assume that the domain sizes and the domain means of an auxiliary variable, X, are available. Consider the three regression estimators of the domain mean,

$$\hat{\mu}_{(1)yi} = \bar{y}_{i.} + (\mu_{xi} - \bar{x}_{i.})b_i,$$

$$\hat{\mu}_{(2)yi} = \bar{y}_{i.} + (\mu_{xi} - \bar{x}_{i.})b_{.}$$

and

$$\hat{\mu}_{(3)yi} = \bar{y}_{..} + (\mu_{xi} - \bar{x}_{..})b_{.},$$

where

$$(\bar{x}_{..},\bar{y}_{..}) = \sum_{i=1}^{k} N^{-1}N_i(\bar{x}_{i.},\bar{y}_{i.}),$$

$$(\bar{x}_{i.},\bar{y}_{i.}) = n_i^{-1}\sum_{j=1}^{n_i}(X_{ij},Y_{ij}),$$

$$b_i = \left[\sum_{j=1}^{n_i}(X_{ij} - \bar{x}_{i.})^2\right]^{-1} \times \sum_{j=1}^{n_i}(X_{ij} - \bar{x}_{i.})(Y_{ij} - \bar{y}_{i.}),$$

$$b_{.} = \left[\sum_{i=1}^{k} N^{-1}N_i n_i^{-1}\sum_{j=1}^{n_i}(X_{ij} - \bar{x}_{i.})^2\right]^{-1} \times \sum_{i=1}^{k} N^{-1}N_i n_i^{-1}\sum_{j=1}^{n_i}(X_{ij} - \bar{x}_{i.})(Y_{ij} - \bar{y}_{i.}),$$

n_i is the number of observations in domain i, N_i is the population size of domain i, μ_{xi} is the population mean of X for domain i, and $\mu_{x.}$ is the grand population mean of X. In the authors' terminology, the first estimator is a direct regression estimator, the second is a modified direct estimator, and the third is a synthetic estimator. We have

$$\text{MSE}\{\hat{\mu}_{(1)yi}|\,n_i\} = n_i^{-1}(1 + n_i^{-1})V\{Y_{\ell j} - \beta_\ell X_{\ell j}|\,\ell = i\} + O(n_i^{-2}),$$

$$\text{MSE}\{\hat{\mu}_{(2)yi}|\,n_i\} = n_i^{-1}(1 + n^{-1})V\{Y_{\ell j} - \beta X_{\ell j}|\,\ell = i\} + O(n^{-2}),$$

$$\text{MSE}\{\hat{\mu}_{(3)yi}|\,n_i\} = (1 + n^{-1}) \times \sum_{i=1}^{k} N^{-2}N_i^2 n_i^{-1}V\{Y_{\ell j} - \beta X_{\ell j}|\,\ell = i\} + (\mu_{xi} - \mu_{x.})^2 V\{b_{.}\} + [\mu_{yi} - \mu_{y.} - \beta(\mu_{xi} - \mu_{x.})]^2 + O(n^{-2}),$$

where $V\{b_{.}\} = E\{(b_{.} - \beta)^2\}$, $V\{a_\ell|\,\ell = i\}$ is the variance of the variable a for domain i,

$$\beta_i = [V\{X_{\ell j}|\,\ell = i\}]^{-1}C\{Y_{\ell j},X_{\ell j}|\,\ell = i\}$$

[1] W.A. Fuller, Distinguished Professor, Statistical Laboratory and Department of Statistics, Iowa State University, Snedecor Hall, Ames, Iowa.

and

$$\beta = \left[\sum_{i=1}^{k} N^{-1} N_i V\{ X_{\ell j} \mid \ell = i \} \right]^{-1} \times \sum_{i=1}^{k} N^{-1} N_i C\{ Y_{\ell j}, X_{\ell j} \mid \ell = i \}.$$

The estimator $\hat{\mu}_{(1)yi}$ uses only information in the sample of n_i observations. Hence, all properties of the estimator are functions of n_i and of the domain parameters. The regression bias is order n_i^{-1} and the variance is order n_i^{-1}. The estimator $\hat{\mu}_{(2)yi}$ uses the domain means, but the entire sample to estimate the regression coefficient. Hence, the basic variance remains order n_i^{-1} and will be larger than the basic variance of $\hat{\mu}_{(1)yi}$ in those situations where $\beta_i \neq \beta$. However, the second order contribution to the variance is order $n_i^{-1}\ n^{-1}$ for $\hat{\mu}_{(2)yi}$ and is order n_i^{-2} for $\hat{\mu}_{(1)yi}$. Also, the regression bias for $\hat{\mu}_{(2)yi}$ is order n^{-1}. If the domains were strata, $\hat{\mu}_{(1)yi}$ might be called the separate regression estimator and $\hat{\mu}_{(2)yi}$ might be called the combined regression estimator.

The estimator $\hat{\mu}_{(3)yi}$ is a synthetic estimator and has a variance of order n^{-1} instead of the order n_i^{-1} variance of the first two estimators. The cost of this reduction in variance is that the bias is order one. Only if the regression line is the same for the domain as for the entire population will the bias be zero.

The average mean square error of the three estimators for any subset of small areas can be estimated. If the n_i are small, the estimated variances will provide only limited information for discriminating among estimators. Likewise, there is only one degree of freedom for bias squared for one particular domain. However, a large domain deviation, relative to the standard error, will lead one to reconsider the synthetic estimator.

In their discussion of models, the authors stress the importance of providing estimators of the reliability for small area estimators. They allude to the fact that the principal estimators of mean square error for model based procedures are estimators of an average mean square error. While this is true, it seems worth mentioning that components-of-variance procedures do not assume the mean square errors to be the same in each domain. Also, for the typical survey situation, the estimators of mean square error need not be constant over domains. For example, one of the terms in the mean square error estimator of the components of variance procedure is the estimator of the variance of the direct estimator. The estimated variance of the direct estimator will be a function of the domain sample size and can also be a function of the direct estimated variance of the direct estimator for that domain. See Battese, Harter, and Fuller (1988), Harville (1976), Prasad and Rao (1990), and Ghosh and Rao (1993).

In their discussion of designs, the authors explain that the variance function is often relatively flat in the vicinity of the optimum allocation to strata. A slight reallocation of sample among strata can markedly increase the efficiency of domain estimators for a relatively small decrease in the efficiency of the overall estimates. The same is true with respect to the combination of direct and synthetic estimators. Thus, if one has a relatively good idea of the variance component associated with small areas, either from a previous study on the same population or from a study on a similar population, and if one is under pressure to produce estimates in a brief time span, then it is reasonable to assign fixed weights to form the linear combination. The loss in efficiency is apt to be modest and the programming required for estimation construction considerably reduced. One estimator in this class, and the one adopted by many practitioners, is the synthetic estimator.

The authors briefly raise the question of internal consistency associated with the construction of small area estimates. As they say, if one uses a data dependent procedure, such as variance components, for each dependent variable, then one produces estimates that are not internally consistent. One option is to use multivariate procedures. See, for example, Fuller and Harter (1987) and Fay (1987). Another procedure suggested by Fuller (1990) is to construct components of variance estimators for a limited subset of variables and then use these estimates as control variables in a regression procedure. The regression procedure produces weights for the individual observations. Once the weights are constructed, any number of output tables can be constructed and all estimates are internally consistent.

It is my observation that the gains made in most practical domain estimation problems come primarily from the wise use of auxiliary information. Thus, effort directed towards obtaining quality auxiliary information is effort well spent. If we are able to find a variable x that is highly correlated with the variable y, then there is less variability remaining to be allocated between area to area variance and sampling variance.

ACKNOWLEDGEMENTS

I thank Jay Breidt for comments.

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.

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.

FULLER, W.A. (1990). Analysis of repeated surveys. *Survey Methodology*, 16, 167-180.

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.

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

GHOSH, M., and RAO, J.N.K. (1993). Small area estimation: An appraisal. Unpublished manuscript. Carleton University, Ottawa, Ontario, Canada.

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.

COMMENT

GRAHAM KALTON[1]

As Singh, Gambino and Mantel (SGM) indicate, there is a growing demand for surveys to provide domain estimates for domains of various sizes and types. This demand is being experienced in many countries throughout the world. In part it may simply reflect a natural growth in the sophistication of survey analysts, who once were content with national estimates and estimates for a few major domains, but who now want to compare and contrast estimates for many different types of domain. In part it results from the needs of policy makers, who require domain information in order to examine how current policies affect different domains, to predict what effects changes in policies might have, and for policy implementation. Information on administrative area domains (*e.g.*, provinces or states, counties, and school districts) is of particular interest for policy purposes (*e.g.*, for identifying low income areas for government support).

In some circumstances the need for domain estimates of adequate precision can be satisfied within the design-based inference framework that is standardly used in the analysis of survey data. This holds for large domains for which the sample sizes are adequate to give the precision required. It can also hold for small domains provided that they are identified in advance, and the sample design is constructed in a way that provides adequate sample sizes. Thus, for example, in the United States, the National Health and Nutrition Examination Survey and the Continuing Survey of Food Intakes by Individuals use differential sampling fractions by age, sex and race/ethnicity and by age/sex and low income status, respectively, in order to provide adequate samples for the domains created by the cross-classifications of these variables. The U.S. Current Population Survey employs differential sampling fractions across the states in order to be able to produce state-level employment estimates. The limitation of this approach is evident when there is a large number of small domains, in which case the sum of the required sample sizes for each domain produces an extremely large overall sample size. This situation occurs often with small administrative districts, such as counties, school districts, and local employment exchanges. In such cases, it may be necessary to discard the standard design-based inference approach in favor of a model-dependent approach that employs a statistical model in the estimation process to borrow strength from data other than that collected in the survey for the given small area. The model-dependent approach may also be required for unplanned small domains, where the need for oversampling had not been foreseen at the design stage.

In response to the demand for small area estimates, a sizeable literature has developed on model-dependent small area estimation methods. Little has, however, been written on the broader issues of small area estimation discussed in the SGM paper, issues that need more attention. Like the authors, I believe that a cautious approach should be adopted to the use of model-dependent small area estimators. I therefore welcome their discussion of methods to make small area estimates within the design-based framework.

From my perspective, the first approach to making small area estimates is to see whether estimates can be produced with adequate precision within the design-based framework. If the domains have been identified in advance, consideration should be given to designing the sample to meet the needs for small area estimates. This may involve ensuring that the small areas do not overlap strata, and ensuring a sufficient sample size for each small area. Another approach suggested by SGM is to minimize the amount of clustering. The smaller the amount of clustering, the less the sample size in each small area is subject to the vagaries of chance. In this regard I see the benefits of less clustering as mainly directed at providing the ability to produce estimates for small areas that were not identified at the design stage. When small areas for which estimates are planned are made into separate strata, the sample size in each small area should be under adequate control even with a clustered sample (provided that the measures of size used in the PPES sampling are reasonable). However, even with planned estimates, there will often be an issue of how to compute variance estimates for a small area from a clustered design, since the number of PSUs sampled in each small area is likely to be small. A variance estimate based on the PSUs within the small area will then be imprecise, with few degrees of freedom, and a generalized variance function approach may be preferred (*e.g.*, assuming that the national design effect applies for each small area). In other words, although the estimate itself may be a design-based estimate, the estimate of its variance may be an indirect one, borrowing strength from other areas. This consideration favors as unclustered a design as possible even for planned small area estimates. The need to model variances is, however, of lesser concern than the need to model the estimates themselves.

An integral part of the design-based framework is a recognition that auxiliary information available for the population may be used at the design stage, at the analysis stage, or at both stages. When information on auxiliary

[1] Graham Kalton, Westat, 1650 Research Blvd., Rockville, Maryland, U.S.A. 20850.

variables that are closely related to the survey variable is available, substantial gains in precision can accrue. The use of auxiliary information at the analysis stage, through such techniques as post-stratification and ratio, regression and difference estimation, has a special appeal for small area estimation. It should be emphasized that ratio and regression estimators may be motivated by assumptions about the model relating the survey variable (Y) and the auxiliary variables (X), but that the resultant estimators are design-consistent irrespective of the appropriateness of the model. The use of an appropriate model produces the greatest gains in precision, but the estimates are approximately unbiased whatever model is chosen. This may be seen in a simple case where variables $X_1, X_2, \dots, X_p$ are known for every element in the population, and the linear combination $\tilde{Y}_i = B_0 + B_1X_{1i} + \dots + B_pX_{pi}$ is used to estimate Y_i, the value of the Y-variable for population element i. Assume, for simplicity that the B's are determined from external data, not dependent on the sample. With $Y_i = \tilde{Y}_i + e_i$, the domain total is $Y_a = \sum_{i \epsilon a} \tilde{Y}_i + \sum_{i \epsilon a} e_i = \tilde{Y}_a + E_a$. Since $\tilde{Y}_a$ is known, the estimation problem is one of estimating E_a. From a sample of elements in domain a, E_a may be estimated by $\hat{E}_a = \sum_{j \epsilon s_a} e_j / \pi_j$, where π_j is the selection probability for element j in the sample. The estimator $\hat{E}_a$ is unbiased, independent of the validity of the model employed. The estimation procedure in fact translates the estimation problem from one of estimating Y_a directly to one of estimating E_a and adding on a known constant $\tilde{Y}_a$. To be effective, the procedure requires the domain variance of the e_i to be smaller than that of the Y_i. There is no requirement that $E_a = 0$. The general logic remains the same in the more usual situation where the B's are estimated from the sample. In this case, the estimate of Y_a is design-consistent, irrespective of the model adopted (Särndal 1984). Moreover, the B's may be estimated from the sample data only for the domain of interest, producing what SGM term a direct estimator, or from the total sample, producing a modified direct estimator. A key consideration in the choice between the direct and modified direct estimators in this case is whether the overall B's also apply for the domain. If not, interaction terms between the X's and the domain indicators are called for in the total sample model. With a full set of these interaction terms, the modified direct estimator in effect then reduces to the direct estimator.

The need for a model-dependent approach occurs when the design-based estimate lacks sufficient precision even after the auxiliary data available have been used in as effective a manner as possible. Indeed, in some cases the computation of a direct estimate may be impossible because there are no sample cases in the small area. In such situations, it becomes necessary to use a statistical model to borrow strength from other data, often data from other areas. Such models are built upon assumptions (*e.g.*, $E_a = 0$ in the above example), and the quality of the resultant small area estimates depends on the suitability of the assumptions made. The assumptions are inevitably incorrect to some degree, leading to biases in the small area estimates. Since indirect estimates are biased, the design-based mean square error (MSE) is widely used as the measure of their quality, where MSE $= V' + B^2$ and V' is the variance and B is the bias of the estimate.

The common way to compare the quality of a direct and an indirect estimate is to compare the variance, V, of the former with the MSE of the latter. However, reading the paper caused me to question whether the MSE is the appropriate measure of quality of an indirect estimator. In a practical setting the variance V of the direct estimate can be estimated whereas the design-based MSE of the indirect estimate cannot. In view of this situation, if $V =$ MSE, then the direct estimator would be clearly preferred. In fact, the direct estimator may tend to be preferred if the direct estimator has adequate precision, irrespective of the likely relative magnitudes of V and MSE. In other cases, if B is the expected bias, then the direct estimator may be preferred to the indirect estimator unless $V > V' + kB^2$, where k is a multiplier greater than 1 that allows for the fact that the unknown bias may be larger than expected.

The same argument can be applied to combined (or composite) estimators that employ a weighted average of a direct and an indirect estimator. Often the principle for choosing the weights is taken to be to minimize the mean square error of the combined estimator, leading to weights for the direct and indirect estimators that are inversely proportional to V and MSE, respectively. However, following the above argument, an alternative procedure would be to minimize the weight of the indirect estimator, subject to the condition that the combined estimator is sufficiently accurate. Alternatively, the weights could be determined on some maximum likely value of the MSE, rather than the expected MSE, to reduce the risk of serious bias in the combined estimator.

I do not follow the rationale for the sample size dependent estimators described by SGM in equation (6.11) and (6.12) in general, but under certain assumptions they may be seen to fit in to the logic given above. With an equal probability sample design and $\delta = 1$, these estimators reduce to the direct estimator when the achieved sample size is greater than, or equal to, the expected sample size. If one assumes that the expected sample size gives adequate precision for the small area, this outcome accords with the above reasoning. If the achieved sample size is smaller than expected, the sample size dependent estimator takes a weighted average of a direct and an indirect estimator. If one assumes that the expected sample size is the minimum sample size to give the required precision, this outcome also accords with the above reasoning. If this indeed is the basis of the sample size dependent estimators, then it would seem useful to generalize them to situations where the expected sample size is not the sample size that just gives the level of precision required.

As has been noted, auxiliary information plays an important role in the production of accurate small area estimates. Such information may be used for improving the precision of design-based estimates or it may be used in the models employed with the model-dependent approach. Ideally auxiliary information that is highly related to the survey variables involved in the estimates is required. The regular compilation of up-to-date auxiliary data for small areas from administrative and other sources can provide a valuable resource for a small area statistics program.

Although the paper mentions the more general problem of small domains, it focuses predominantly on small areas. This is in line with the general literature and the application of indirect estimation procedures. In part, this may be because the number of socio-economic and other small domains of interest (*e.g.*, age/sex domains) is usually relatively small, compared with the numbers of small areas, so that socio-economic domains can be handled by designing the sample to provide direct estimates of adequate precision for each of them. In part, it may be because the definitions of socio-economic and demographic domains are often chosen in the light of the feasibility of producing design-based estimates of adequate precision for them (*e.g.*, using wider age groupings for some domains); in the case of areal domains, however, the areas are predefined, and no collapsing of areas is acceptable. In part, it may be because there is a lack of auxiliary data to use in the statistical models for such domains. In part, it may also be because the analysis of socio-economic domains is often conducted to make comparisons between the domains. Such comparisons are distorted when the estimate for one domain borrows strength from other domains (see, for example, Schaible 1992). This issue brings out the general point that indirect estimates should not be uncritically used for all purposes.

In conclusion, I should like to express my support for the general approach of this paper. Where possible, samples should be designed to produce direct small area estimates of adequate precision, and sample designs should be fashioned with this in mind. Auxiliary data should be used, where possible, to improve the precision of direct small area estimates. When indirect estimates are called for, a cautious approach should be used. Models should be developed carefully, estimators that are robust to failures in the model assumptions should be sought, and evaluation studies should be conducted to assess the adequacy of the indirect estimates. Lacking good measures of quality for individual indirect estimates, such estimates need to be clearly distinguished from design-based estimators. Since indirect estimates are not universally valid for all purposes, users need to carefully assess whether the given form of indirect estimate will satisfy their particular needs.

REFERENCES

SÄRNDAL, C.-E. (1984). Design-consistent versus model-dependent estimators for small domains. *Journal of the American Statistical Association*, 79, 624-631.

SCHAIBLE, W.L. (1992). Use of small area estimators in U.S. Federal programs. *Proceedings of the International Scientific Conference on Small Area Statistics and Survey Design*, (Vol. 1), 95-114, Central Statistical Office of Poland, Warsaw.

RESPONSE FROM THE AUTHORS

We would like to thank Wayne Fuller and Graham Kalton for their stimulating comments, which we find to be quite complementary to the position developed in our paper. In many cases their comments make certain points clearer and strengthen the arguments presented. Encouraged with this kind of endorsement we would like to carry some of the points about survey design further, while responding to the main points made by the discussants.

There is no doubt that survey designers try to optimize the design under operational constraints to meet the stated objectives of a survey. There are usually several objectives to be met by major surveys and it is quite likely that designers have limited influence in the setting of priorities among the various competing objectives. Nevertheless, it is at this stage of priority setting that the case for small area needs should be made strongly, particularly for major continuing surveys.

During the sixties and seventies emphasis in most countries was placed on sub-national (state/provincial) estimates and certain compromises were made to the earlier designs that optimized national estimates. For example, different sampling fractions were used to ensure a minimum sample size for smaller states/provinces. With the demands for data at the sub-state/province level, such as, county, district and municipality, more compromises to the national optimum allocation become necessary, requiring differing sampling fractions among the administrative areas within states/provinces. For example, if the aim is to produce sub-provincial estimates of comparable quality, then provinces will likely receive sample roughly proportional to the number of subprovincial regions they contain. Such an allocation may not be the same as one using the relative population sizes of the provinces. As we discussed in section 5.4, the allocation approach should put more emphasis on a bottom-up strategy. Losses at higher levels and gains at lower levels would differ from survey to survey but it is likely that in many cases a minor loss in CV at the national level will lead to appreciable gains at small area levels.

Kalton stresses the importance of reduced clustering for variance estimation; it is advantageous to increase the degrees of freedom by having a large number of smaller clusters rather than a small number of larger clusters. We would like to emphasize that clustering has another drawback for estimation, and especially small area estimation, namely, a highly clustered design will lead to high design effects, even for planned small domains. The usual reason for resorting to clustered designs is to reduce survey costs. In light of the changes that continue to occur in the data collection process, such as decreased reliance on at-home interviews and increased use of computer assisted interviewing, a periodic review of the cost-variance models that underlie clustering decisions is necessary.

One other issue not addressed in our paper is the impact of sample rotation in continuous surveys. For a given time point, there may be insufficient sample in some small domains to produce reliable estimates. But, as units rotate out of the sample and are replaced, the accumulated or effective sample in the domains increases and may allow the computation of reliable, albeit time-biased, domain estimates. By judicious choice of rotation schemes, survey designers can maximize the cumulative sample size over some time period. For example, for quarterly estimates in a monthly survey, the optimal rotation pattern is $[1(2)]^k$, *i.e.*, repeat the sequence "one month in sample, two months out" k times. This thinking is in the same spirit as Leslie Kish's ideas on cumulation of samples over time.

Kalton clarifies and elaborates the cautious approach to the use of indirect estimators by suggesting a weighted mean squared error, which attaches a weight greater than 1 to the bias term, to allow for the fact that the bias of the indirect estimator may be larger than expected. There are two distinct reasons why the bias may be larger than what is expected from the model for small area effects: random variation within the model, and model breakdown. It is worth recalling here the suggestion of Fay and Herriot (1979) to constrain a combined estimate to be within one standard error of a design estimate; this approach makes allowance for the possibility of large bias in the model estimator for whatever reason. Kalton also reiterates our position that if a direct estimator is of acceptable quality, then in practice, one may decide to use this direct estimator even though its estimated mean squared error exceeds that of model-based competitors. Because there is always the possibility of model failure lurking in the background, this "better safe than sorry" approach is desirable, at least until some experience with particular indirect estimators in specific situations has been gained. This does not contradict the view that there arise situations in which it is necessary to throw caution to the wind.

In his remarks on the sample size dependent estimator, Kalton's comments imply that there is a risk in the strategy which gives the synthetic component zero weight if the observed sample size in the small domain exceeds the expected sample size there since the latter may be too small to yield adequate direct estimates. One option is to use a value $n_{\min}$ which is the size that produces direct estimates that are just barely acceptable. Note, however, that $n_{\min}$ as defined here is characteristic-dependent.

In his comments, Fuller briefly describes an approach to small area estimation that takes advantage of a variance components model and yet has fixed weights for internal consistency among estimators for different characteristics. Besides internal consistency of small area estimates for different characteristics, a second type of consistency that is sometimes required is that estimates of totals for the set of small areas within a larger area should add up to the published direct estimate for the larger area. One way to achieve this is to benchmark the small area estimates to the direct estimate for the larger area using, for example, a simple ratio adjustment; however, if the ratio adjustment factors depend on the characteristic then this would destroy the first type of consistency. Both types of consistency could be achieved simultaneously if the direct estimators for the larger area are generalized regression estimators, $\hat{Y}_e + (X - \hat{X}_e)\hat{\beta}$, and the modified direct (Section 6.1 in the paper) estimators $\hat{Y}_{sreg,a} = \hat{Y}_{e,a} + (X_a - \hat{X}_{e,a})\hat{\beta}$ are used for small areas.

As Fuller notes, the average squared bias of an estimator for any subset of small areas can be estimated. Here we would like to stress again that the average bias over a set of small areas is not directly relevant for any particular small area. It is for this reason that we prefer to use, whenever possible, estimators that are approximately design unbiased. When use of a model estimator is unavoidable, serious attempts should be made to find appropriate covariates for which reliable auxiliary information is available in order to minimize the residual bias of the model estimator.

Perhaps due to the obvious timeliness problems associated with census data, neither of the discussants commented on censuses as a source of data for smaller domains. In this context it is worth mentioning that some form of ongoing major post-censal survey replacing or supplementing the decennial census long-form may be considered. Such a strategy, called rolling samples, is described by Kish (1990); a similar approach, called continuous measurement, is described by Alexander (1994). This approach provides a number of options which are worth investigating as potentially cost effective means of producing timely statistics for smaller domains.

Lastly, we would like to stress that the emphasis we put on keeping domain estimation in mind at the design stage, particularly for medium size domains, in no way undermines the important role of models in estimating for very small domains.

We hope that the general direction of the strategy proposed in the paper, supplemented by the fine points brought out by the discussants, particularly the support and cautions summarized by Kalton in his concluding paragraph, will be helpful to survey designers and researchers in finding solutions appropriate to the particular problems they are dealing with.

ADDITIONAL REFERENCES

ALEXANDER, C.H. (1994). A prototype continuous measurement system for the U.S. Census of Population and Housing. Document for presentation at the annual meeting of the Population Association of America, Miami, Florida, May 5, 1994.

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

Survey Methodology, June 1994
Vol. 20, No. 1, pp. 23–31
Statistics Canada

Small Domain Estimation for Unequal Probability Survey Designs

D. HOLT and D.J. HOLMES[1]

ABSTRACT

The problem of estimating domain totals and means from sample survey data is common. When the domain is large, the observed sample is generally large enough that direct, design-based estimators are sufficiently accurate. But when the domain is small, the observed sample size is small and direct estimators are inadequate. Small area estimation is a particular case in point and alternative methods such as synthetic estimation or model-based estimators have been developed. The two usual facets of such methods are that information is 'borrowed' from other small domains (or areas) so as to obtain more precise estimators of certain parameters and these are then combined with auxiliary information, such as population means or totals, from each small area in turn to obtain a more precise estimate of the domain (or area) mean or total. This paper describes a case involving unequal probability sampling in which no auxiliary population means or totals are available and borrowing strength from other domains is not allowed and yet simple model-based estimators are developed which appear to offer substantial efficiency gains. The approach is motivated by an application to market research but the methods are more widely applicable.

KEY WORDS: Synthetic estimation; Design-based estimation; Small area estimation; Model-based estimation; Market shares.

1. INTRODUCTION

This paper is concerned with the common problem of estimating domain totals and means from a disproportionately allocated sample survey. Some domains may be large, in which case the achieved sample size may be large too and design-based (or direct) estimators will be satisfactory. Some domains may be small, in which case the achieved sample size may be small too and design-based (or direct) estimators will be too imprecise for practical use. The methods proposed will be motivated through the example of estimating sales, market shares and market penetrations for products in a market research survey. The domains are particular auto manufacturers or models. However, the general approach is applicable to other disproportionately allocated surveys of businesses or institutions.

The problem is analogous to that of using synthetic estimation for small area estimation (Gonzales 1973; Gonzales and Hoza 1978; Platek *et al.* 1987). Synthetic estimation usually depends on two factors: (i) the use of auxiliary variables in conjunction with population means or totals for each small area (or domain) to improve estimates through poststratification or regression estimation, and (ii) the improvement of estimates by pooling data across the small areas (or domains). In our situation no auxiliary population means or totals are available and, since the essential objective is to compare domains (*i.e.*, manufacturers and particular products), the idea of borrowing strength between these is inadmissible. A class of synthetic estimators is proposed which uses neither of these two approaches and yet is preferred to the direct survey estimators. The proposed estimators have a simple structure, an interesting interpretation and can be justified under a set of model assumptions which are testable under the general assumption of non-informative survey design.

2. THE MARKET RESEARCH EXAMPLE

Market researchers often estimate the total volume of sales and market shares for each manufacturer of a particular product. We consider the case of autos purchased for company fleet use in a single year. Estimates of totals and market shares are required for each auto manufacturer and for specific models which are widely purchased for fleet use.

The terms 'fleet' and 'company' are each interpreted widely. A fleet car is taken to mean any auto purchased on a commercial as opposed to a private basis, and used in conjunction with a business in the broadest sense. This includes autos purchased for sales representatives which may be purchased in large numbers. It also includes single purchases of luxury cars for company directors and other senior staff of large companies, as well as purchases by small 'companies' such as groups of doctors, or self-employed people such as shop owners. Thus the population of purchasing companies – termed consumers – includes a large number of small companies that purchase only one or two autos every few years.

[1] D. Holt and D.J. Holmes, Department of Social Statistics, University of Southampton, Highfield, Southampton, UK, SO95NH.

In the reference period of one year we define Y_{ki} to be the number of autos of product type k purchased by consumer i. The product type k (the domain) may refer to a specific model of a particular manufacturer, or to all models produced by a manufacturer. Thus, $Y_k = \sum_i Y_{ki}$ is the total number of autos of type k purchased by all consumers. Let Z_i be the total number of autos of any kind purchased by consumer i, and $Z = \sum_i Z_i$ be the total number of auto sales. The market share for product type k is defined as $R_k = Y_k/Z$.

We further define

$$\begin{aligned} Y'_{ki} &= 1 \quad \text{if} \quad Y_{ki} > 0 \\ &= 0 \quad \text{if} \quad Y_{ki} = 0 \end{aligned}$$

and

$$\begin{aligned} Z'_i &= 1 \quad \text{if} \quad Z_i > 0 \\ &= 0 \quad \text{if} \quad Z_i = 0. \end{aligned}$$

Thus, Y'_{ki} and Z'_i are indicator variables for consumers who purchase product type k and at least one auto of any kind, respectively, in the reference period. The number of consumers that purchase product k is thus given by $Y'_k = \sum_i Y'_{ki}$ and the total number of consumers purchasing at least one auto of any kind is given by $Z' = \sum_i Z'_i$. The market penetration for product k, in terms of the proportion of consumers buying a car of any type in the reference period who buy type k, is given by $R'_k = Y'_k/Z'$.

The four parameters Y_k, R_k, Y'_k and R'_k are all legitimate targets of inference in market research and are defined as finite population parameters; namely, domain totals or ratios of domain totals.

3. THE SURVEY DESIGN AND DIRECT ESTIMATORS

The survey design was based upon two mutually exclusive frames and may be regarded as a simple stratified design with ten strata. The first frame was a register (Dun and Bradstreet) of 35,000 companies, stratified into eight strata on the basis of the number of employees and whether the company was classified as 'manufacturing' or 'distributing'. The second frame was a large register of 1.4 million British Telecom business subscribers, stratified into 'private' and 'commercial' numbers. Note that both private and commercial numbers were business subscribers but commercial numbers were allocated if separate commercial premises were occupied.

Using previous survey data the sample was optimally allocated using Neyman allocation to minimize the variance of the estimator of the total number of autos purchased (Z). Data on auto purchases were collected immediately after the end of the reference year. The strata sizes $\{N_h\}$ and sample allocations $\{n_h\}$ for strata $h = 1, \ldots, 10$ are given in Table 1.

Table 1

Sampling Frame: Sample Size and Weight by Stratum

Stratum (h)	Stratum Size N_h	Sample Size n_h	Weight $\pi_h^{-1} = N_h/n_h$
British Telecom:			
Private	389,445	1,150	338.65
Commercial	1,007,399	7,406	136.02
Dun and Bradstreet:			
Manufacturing			
50-99 employees	6,646	235	28.28
100-499	6,826	1,113	6.13
500-999	992	520	1.91
1,000+	1,110	849	1.31
Distributing			
50-99 employees	8,703	472	18.44
100-499	7,625	1,437	5.31
500-999	1,133	484	2.34
1,000+	1,523	1,117	1.36
Overall	1,431,402	14,783	96.83

The sample is a simple, disproportionately allocated stratified design and the direct estimators and their variances are well known. The stratification results in large differences in sampling weights (1.31 to 338.65) and is useful but far from ideal. Many consumers do not purchase any autos at all in the reference year so that each stratum contains a mixture of zero and non-zero responses. For any particular product k the proportion of zero responses in each stratum is obviously larger.

Table 2 contains the direct survey estimates, estimated standard errors (see Holt and Holmes (1993) for derivation), and coefficients of variation for a selection of products from different auto manufacturers. Products A and B represent all models for two major auto manufacturers. Product C is a single model with a substantial share of the fleet market from manufacturer A. The remaining products have small market shares. Products F and G cater for the executive part of the fleet market. The list is incomplete so that the market shares do not sum to one. Also note that the product categories are not mutually exclusive. In general the survey was judged to perform satisfactorily but it was observed over a period of years that estimates for manufacturers or models with small market shares were unstable. This is best seen in terms of the coefficient of variation which is greater than 0.1 for products with small market shares and can be greater than 0.15 or 0.2 in some cases. This instability also affects the estimates of variance as well as the estimates of total sales or market shares of the products.

Table 2

Direct Survey Estimates, Standard Errors and Coefficients of Variation for Selected Products

Product	Estimating Consumers		Estimating Autos	
(k)	Total $\hat{Y}'_k$	Penetration $\hat{R}'_k$	Total $\hat{Y}_k$	Share $\hat{R}_k$
A	59,890	.3843	270,051	.3781
	(2,651)	(.0144)	(35,704)	(.0315)
	(.044)	(.037)	(.132)	(.083)
B	34,282	.2200	153,518	.2149
	(1,960)	(.0117)	(8,653)	(.0131)
	(.057)	(.053)	(.056)	(.061)
C	23,363	.1499	81,381	.1139
	(1,602)	(.0098)	(17,559)	(.0194)
	(.069)	(.065)	(.216)	(.170)
D	13,857	.0889	25,312	.0354
	(1,311)	(.0081)	(2,906)	(.0039)
	(.095)	(.091)	(.115)	(.110)
E	9,025	.0579	24,370	.0341
	(1,146)	(.0072)	(7,336)	(.0101)
	(.127)	(.124)	(.301)	(.296)
F	5,125	.0329	13,724	.0192
	(676)	(.0043)	(2,369)	(.0030)
	(.132)	(.131)	(.173)	(.156)
G	7,518	.0482	11,031	.0154
	(1,015)	(.0064)	(1,456)	(.0022)
	(.135)	(.133)	(.132)	(.143)

Row 1: estimate Row 2: s.e. Row 3: c.v.

4. A MODEL-BASED APPROACH

Given the sample design there is no prospect of improving the efficiency of the direct survey estimators within the conventional sample survey framework. The usual approaches are through the use of auxiliary information for poststratification, ratio or regression estimation but all of these require knowledge of population means or totals. No such information is available. We turn instead to a model-based approach to provide alternative estimators for the whole range of products.

4.1 Estimating Y'_k: the Number of Consumers Purchasing Product Type k

We consider, initially, the number of consumers who buy product type k. We extend the notation from Y'_{ki} to Y'_{khi} in the obvious way to define the indicator random variable of purchase for product k for consumer i in stratum h. We treat each consumer's decision as the outcome of a Bernoulli trial. Let $P_{k|h}$ be the probability that a consumer in stratum h buys an auto of type k $[P_{k|h} = \text{Prob}(Y'_{khi} = 1)]$. We define the model-based equivalent of Y'_k, the total number of consumers of product k, as

$$\Theta'_k = \sum_h N_h P_{k|h}. \tag{1}$$

Assuming that each consumer's decision is independent the likelihood may be written as the usual product of binomial terms. The maximum likelihood estimators are given by $\hat{P}_{k|h} = n_{kh}/n_h$, and the maximum likelihood estimator of Θ'_k is the familiar stratified sampling estimator

$$\hat{\Theta}'_k(1) = \sum_h \frac{N_h}{n_h} n_{kh} = \sum_h N_h \bar{y}'_{kh}, \tag{2}$$

where n_{kh} is the sample count of consumers in stratum h that buy product k, n_h is the stratum sample size and $\bar{y}'_{kh} = n_{kh}/n_h$ is the sample mean for consumers in stratum h (*i.e.*, the sample proportion of consumers in stratum h who buy product k). This estimator is generally unsatisfactory when the sample size for product k is too small.

Suppose we introduce an additional conditioning factor such that every consumer may be categorized into one of its categories f, $f = 1, \ldots, F$, and further extend the definition of the indicator random variable to Y'_{khfi}. These categories f will cut across the strata h and the idea is to define f so that, within any particular category, whether a consumer buys product type k or not is independent of the stratum membership h. In the case of fleet purchases we define a categorization based on the total number of autos owned and operated by each consumer (*i.e.*, the fleet size). A more detailed discussion of the choice of f is given in Section 5.

If N_{hf}, the population counts of consumers in stratum h and fleet size category f, are known then (1) may be extended in the obvious way and the target parameter can now be expressed as

$$\Theta'_k = \sum_h \sum_f N_{hf} P_{k|hf}. \tag{3}$$

Equation (3) is the case of poststratification if $\{N_{hf}\}$ are known, and in this case the additional information will lead to a gain in efficiency (Holt and Smith 1979). When $\{N_{hf}\}$ are unknown we may rewrite the model in terms of two sets of probabilities:

$Q_{f|h}$ = Prob {consumer has fleet size f | stratum h},

$P_{k|hf}$ = Prob {consumer buys product type k | stratum h and fleet size f}.

The target parameter may now be expressed as

$$\Theta'_k = \sum_h \sum_f N_h Q_{f|h} P_{k|hf}. \tag{4}$$

To obtain an alternative model-based estimator we make further assumptions about the model parameters. Suppose now that

$$P_{k|hf} = P_{k|f} \quad \text{for all} \quad h. \tag{5}$$

This implies that conditional on the categorization f (the size of the fleet operated by a consumer), the probability of buying product type k is *independent* of the original stratum membership h. Algebraically, the assumption is analogous to that used in synthetic estimation for small area estimation but in that case information is pooled across areas. That form of the assumption is inadmissible in our case. We choose instead pooling across strata within the domain of study. The idea is to choose a conditioning variable which accounts for the marginal association between choice of product and stratum membership.

Using assumption (5) and with the obvious extension of the notation ($n_{kf} = \sum_h n_{khf}$, *etc.*) it may be shown that

$$\hat{Q}_{f|h} = \frac{n_{hf}}{n_h}, \qquad \hat{P}_{k|f} = \frac{n_{kf}}{n_f}$$

and the maximum likelihood estimator of Θ_k' becomes

$$\hat{\Theta}_k'(2) = \sum_h \sum_f N_h \frac{n_{hf}}{n_h} \frac{n_{kf}}{n_f} = \sum_f \hat{N}_f \frac{n_{kf}}{n_f}$$

$$= \sum_h \hat{N}_f \bar{y}_{kf}', \tag{6}$$

where $\hat{N}_f = \sum_h N_h\, n_{hf}/n_h$, and $\bar{y}_{kf}' = n_{kf}/n_f$ is the unweighted sample mean for consumers in category f (*i.e.* the sample proportion of consumers in category f who buy product k).

Thus (6) has the form of a stratified estimator based on the categorization f but with the population sizes in each stratum $\{N_f\}$ unknown. Note that an estimator of this form, but with known $\{N_f\}$, would arise naturally if a stratified sample based on f had been selected. In fact this is **not** so: the sample members of category f are **not** selected with equal probability. However, the parameter assumptions lead to treating the sample in each category f as if it was an equal probability sample since under assumption (5) the sample weights are uninformative and simply lead to efficiency loss when estimating $P_{k|f}$. Hence, although the sampling fractions n_h/N_h are used to estimate $\{N_f\}$ they are not used explicitly in $\hat{P}_{k|f} = n_{kf}/n_f = \bar{y}_{kf}'$. Note that the estimator pools information across strata h, within domain k but **not** between domains (*i.e.* products).

Note that if n_h/N_h is constant, equation (6) reduces to the usual expansion estimator given by (2), and assumption (5) has not yielded a new estimator. If the sample is disproportionately allocated the assumption leads to the use of the sampling weights for $\hat{N}_f$ (where they are needed) but not for estimating $P_{k|f}$ (where they are uninformative given f and assumption (5)).

Equation (5) is a strong set of assumptions, requiring $P_{k|hf}$ to be exactly equal to a common value $P_{k|f}$ for all h. In practice, random assumptions such as $P_{k|hf} = P_{k|f} + \epsilon_{k|hf}$ may be introduced, where $E[\epsilon_{k|hf}] = 0$ and $V[\epsilon_{k|hf}] = \sigma_\epsilon^2$. These assumptions will lead to hierarchical Bayes or empirical Bayes analysis as described in Ghosh and Rao (1994) or Fay and Herriot (1979). These methods are not developed here since the simple form of the model-based estimator would be lost, together with the insight that this provides. In a similar vein the approach of Särndal and Hidiriglou (1989) or Drew, Singh and Choudhry (1982) may be applied to yield sample size dependent estimators without violating the requirement that no information is pooled across domains (products).

We can compare the estimators in (2) and (6) when assumption (5) holds since it may be shown that

$$V_\xi(\hat{\Theta}_k'(1)) = \sum_h \frac{N_h^2}{n_h} P_{k|h}(1 - P_{k|h})$$

$$= \sum_h \sum_f \frac{N_h^2}{n_h} Q_{f|h} P_{k|f}$$

$$- \sum_h \sum_f \sum_{f'} \frac{N_h^2}{n_h} Q_{f|h} Q_{f'|h} P_{k|f} P_{k|f'}, \tag{7}$$

where the notation $V_\xi(\cdot)$ is used to emphasize that the variance is evaluated with respect to the model-based distribution.

It may also be shown that under assumption (5)

$$V_\xi(\hat{\Theta}_k'(2)) = \sum_h \sum_f \frac{N_h^2}{n_h} P_{k|f}^2\, Q_{f|h}\,(1 - Q_{f|h})$$

$$- \sum_h \sum_{\substack{f \\ f \neq f'}} \sum_{f'} \frac{N_h^2}{n_h} P_{k|f}\, P_{k|f'}\, Q_{f|h}\, Q_{f'|h}$$

$$+ \sum_h \sum_f \frac{N_h^2}{n_h} \frac{P_{k|f}(1 - P_{k|f}) Q_{f|h}}{\sum_h n_h Q_{f|h}}$$

$$\Bigg\{ (1 - Q_{f|h}) + n_h Q_{f|h}$$

$$+ \frac{[1 + (2n_h - 3) Q_{f|h} - 2(n_h - 1) Q_{f|h}^2]}{\sum_h n_h Q_{f|h}} \Bigg\} \tag{8}$$

and that $V_\xi(\hat{\Theta}_k'(1)) - V_\xi(\hat{\Theta}_k'(2)) \geq 0$.

Thus under the additional model assumptions $\hat{\theta}'_k(2)$ has smaller variance as would be expected. These expressions are model-based variances and no finite population corrections arise. A predictive approach to the unobserved elements in each poststratum would give rise to finite population correction factors.

The maximum likelihood estimator of the market penetration for product type k, R'_k, under assumption (5) is simply given by

$$\hat{\Omega}'_k(2) = \frac{\sum_f \hat{N}_f \dfrac{n_{kf}}{n_f}}{\sum_f \hat{N}_f \dfrac{n_{af}}{n_f}} = \frac{\sum_f \hat{N}_f \bar{y}'_{kf}}{\sum_f \hat{N}_f \bar{z}'_f}, \qquad (9)$$

where n_{af} is the sample count of consumers in fleet category f that buy an auto of any kind, and $\bar{z}'_f = n_{af}/n_f$ is the sample proportion of consumers in category f who buy an auto of any kind.

4.2 Efficiency of the Model-Based Estimator of Y'_k

To investigate the gain in efficiency of $\hat{\theta}'_k(2)$ over $\hat{\theta}'_k(1)$ we consider the efficiency of the model-based estimator, defined by

$$e[\hat{\theta}'_k(2)] = \frac{V_\xi(\hat{\theta}'_k(1)) - V_\xi(\hat{\theta}'_k(2))}{V_\xi(\hat{\theta}'_k(1))}, \qquad (10)$$

for various population structures in which assumption (5) holds.

We consider a population with strata $\{h\}$, stratum sizes $\{N_h\}$ and sample allocations $\{n_h\}$ as given in Table 1, and a conditioning factor with ten categories f ($f = 1, \ldots, 10$) of increasing fleet size. We compute the efficiency factor $e[\hat{\theta}'_k(2)]$ for various combinations of parameter values of $\{Q_{f|h}\}$ and $\{P_{k|f}\}$.

We consider five different structures for $\{Q_{f|h}\}$:

$$\text{(a) } Q_{f|h} = \begin{cases} 1 & f = h \\ 0 & f \neq h \end{cases} \quad \text{for } h = 1, \ldots, 10.$$

$$\text{(b) } Q_{f|h} = \begin{cases} 0.95 & f = h & \text{for} & h = 1, \ldots, 10 \\ 0.025 & f = h - 1 & \text{for} & h = 2, \ldots, 10 \\ 0.025 & f = h + 1 & \text{for} & h = 1, \ldots, 9 \\ 0.05 & h = 1, f = 2 & \text{and} & h = 10, f = 9 \\ 0 & \text{otherwise} \end{cases}$$

$$= \text{Band Matrix } (0.025, 0.95, 0.025).$$

(c) $Q_{f|h}$ = Band Matrix (0.05, 0.90, 0.05).

(d) $Q_{f|h}$ = Band Matrix (0.05, 0.10, 0.70, 0.10, 0.05).

(e) $Q_{f|h} = 0.1$ for $h = 1, \ldots, 10$
and $f = 1, \ldots, 10$.

We consider four different structures for $\{P_{k|f}\}$:

$$\text{(i) } P_{k|f} = \begin{cases} 0.1 & f = 1, 2 \\ 0 & \text{otherwise.} \end{cases}$$

(ii) $P_{k|f} = 0.1 - 0.01\,(f - 1)$ for $f = 1, \ldots, 10$.

(iii) $P_{k|f} = 0.1f$ for $f = 1, \ldots, 10$.

(iv) $P_{k|f} = 0.5$ for $f = 1, \ldots, 10$.

Structure (a) is one where the categorization f coincides with the stratification. In structures (b), (c) and (d), in any particular stratum h the majority of consumers fall into one fleet category ($f = h$) with a few consumers in neighbouring categories (*e.g.*, for (b) and (c) $f = h - 1$, $h + 1$). Finally, structure (e) implies that, in any stratum h, consumers will be equally likely to fall into any one of the fleet categories $f = 1, \ldots, 10$.

Structure (i) for $P_{k|f}$ implies a type of auto that is purchased with a small probability by consumers with small fleet sizes (*i.e.* that fall in categories $f = 1$ or 2), but not purchased by consumers with large (r) fleet sizes. Structure (ii) suggests a type of auto purchased with small probability which decreases as fleet size increases, whilst structure (iii) implies the reverse. In structure (iv) a popular model is bought with probability 0.5 *regardless* of the consumer's fleet size.

Table 3 gives the efficiency factor defined in (10) for each combination of structures for $Q_{f|h}$ and $P_{k|f}$ under the disproportionate allocation given in Table 1. Column (a) of the table is the special case where the stratification and the categorization f coincide, and the two estimators $\hat{\theta}'_k(1)$ and $\hat{\theta}'_k(2)$ are the same. The table shows that large gains in efficiency (*e.g.*, 70%) can be attained for certain parameter combinations: the weaker the association

Table 3
Efficiency Factors, $e[\hat{\theta}'_k(2)]$, for Various Combinations of $Q_{f|h}$ and $P_{k|f}$

		Structure for $Q_{f\|h}$				
		(a)	(b)	(c)	(d)	(e)
	(i)	0	0.108	0.196	0.355	0.648
Structure	(ii)	0	0.116	0.206	0.391	0.695
for $P_{k\|f}$	(iii)	0	0.103	0.181	0.387	0.695
	(iv)	0	0.115	0.203	0.391	0.706

between f and h the greater the efficiency gain. Even for structures (c) and (d) where the association between f and h is strong, substantial efficiency gains can be achieved. The structure $Q_{f|h}$ is much more important than $P_{k|f}$ in determining efficiency gain.

In the special case (e) where $Q_{f|h}$ is a constant for all f and h it can be shown that the efficiency factor can be expressed as

$$e[\hat{\Theta}_k'(2)] = \left(1 - \frac{\delta^2}{\bar{P}_{k|f}\,(1 - \bar{P}_{k|f})}\right) \frac{\sum\limits_h \tau_h N_h^2/n_h}{\sum\limits_h N_h^2/n_h}, \quad (11)$$

where

$$\bar{P}_{k|f} = \frac{1}{F} \sum_{f=1}^{F} P_{k|f} \quad \text{and} \quad \delta^2 = \frac{1}{F} \sum_{f=1}^{F} (P_{k|f} - \bar{P}_{k|f})^2$$

are the mean and variance of $\{P_{k|f}\}$ over the categories f, and $\tau_h = 1 - n_h/n + O(n^{-1})$. The term in parentheses in (11) lies between 0 and 1 and it's value depends on how the $\{P_{k|f}\}$ vary over the categories f. In case (iv) $P_{k|f}$ is constant and so this term is unity. The second term of (11) depends solely on the design, and its value for the sample allocation specified in Table 1 is 0.706.

4.3 Estimating Y_k: the Number of Autos Purchased of Product Type k

The previous approach in Section 4.1 may be extended to the number of purchases. We introduce a further conditioning factor which represents the total number of autos purchased, m, regardless of product type, and we extend the notation in the obvious manner to Y_{khfmi}, the random variable representing the number of autos of product type k purchased by consumer i in stratum h, fleet size f, and buying m autos of any kind. The idea is that the number of purchases of product k is likely to vary depending on the total number of autos purchased. Let

$$S_{m|hf} = \text{Prob}\{\text{consumer buys } m \text{ autos of any kind} \mid h,f\},$$
$$m = 0, 1, 2, \ldots,$$

$$T_{\ell|hfm} = \text{Prob}\{\text{consumer buys } \ell \text{ autos of type } k \mid h,f,m\},$$
$$\ell = 0, 1, \ldots, m.$$

The model-based target parameter, equivalent to the total purchases of product k, Y_k, is extended from (4) and may now be expressed as

$$\Theta_k = \sum_h \sum_f \sum_m \sum_\ell N_h Q_{f|h} S_{m|hf} T_{\ell|hfm}\, \ell. \quad (12)$$

We consider two sets of additional assumptions, the first of which is

$$T_{\ell|hfm} = T_{\ell|fm} \quad \text{for all} \quad h. \quad (13)$$

These assumptions imply that conditional on fleet size category, f, and the total number of new autos purchased, m, the distribution of the number of autos purchased of product type k is independent of stratum h.

The maximum likelihood estimator of Θ_k under assumptions (13) is

$$\hat{\Theta}_k(2) = \sum_f \sum_m \hat{N}_{fm}\, \bar{y}_{kfm}, \quad (14)$$

where $\hat{N}_{fm} = \sum_h N_h n_{hfm}/n_h$, and $\bar{y}_{kfm} = \sum_\ell \ell\, n_{fm\ell}/n_{fm}$ is the unweighted sample mean of the number of autos of product type k purchased by consumers of fleet size f that purchased a total of m autos of any kind.

The selection probabilities are used here to provide a weighted estimator of N_{fm}, the total number of consumers of fleet size f that buy m cars of any kind. The form of the estimator is analogous to that in equation (6). Under the model assumption (13) it may be shown that

$$V_\xi(\hat{\Theta}_k(2)) = \sum_h \sum_f \sum_m \frac{N_h^2}{n_h} \mu_{fm}^2 Q_{fm|h}\,(1 - Q_{fm|h})$$
$$- \sum_h \sum_f \sum_m \sum_{f'} \sum_{m'}{}_{(f,m) \neq (f',m')} \frac{N_h^2}{n_h} \mu_{fm}\,\mu_{f'm'}\, Q_{fm|h}\, Q_{f'm'|h}$$
$$+ \sum_h \sum_f \sum_m \frac{N_h^2}{n_h} \frac{\sigma_{fm}^2 Q_{fm|h}}{\sum\limits_h n_h\, Q_{fm|h}}$$
$$\left\{(1 - Q_{fm|h}) + n_h\, Q_{fm|h}\right.$$
$$\left. + \frac{[1 + (2n_h - 3)Q_{fm|h} - 2(n_h - 1)Q_{fm|h}^2]}{\sum\limits_h n_h\, Q_{fm|h}}\right\}, \quad (15)$$

where $Q_{fm|h} = Q_{f|h}\, S_{m|hf}$, $\mu_{fm} = E_\xi\{Y_{khfmi}\}$, and $\sigma_{fm}^2 = V_\xi\{Y_{khfmi}\}$.

In practice, $\bar{y}_{kfm}$ will be based on very few observations if few customers in fleet size category f purchase exactly m cars. For more stability m may be defined as an ordinal variable by grouping the total number of autos purchased into a small number of categories. In this case assumption (13) implies that the distribution of purchases for product type k is the same within fleet size category f and total purchase category m. Also, ℓ may be treated as a continuous random variable and distributional assumptions made about ℓ leading to ratio or regression estimators.

A second and even stronger set of parameter assumptions is

$$T_{\ell|hfm} = T_{\ell|fm} \quad \text{for all} \quad h,$$
$$S_{m|hf} = S_{m|f} \quad \text{for all} \quad h. \qquad (16)$$

These assumptions imply that conditional on fleet size category, f, the joint distribution of the number of autos purchased of type k and the total number of autos purchased of any kind, m, is independent of the stratum h. In this case the maximum likelihood estimator of Θ_k is given by

$$\hat{\Theta}_k(3) = \sum_f \hat{N}_f \bar{y}_{kf}, \qquad (17)$$

where $\bar{y}_{kf} = \sum_\ell \ell\, n_{f\ell}/n_f$ is the unweighted sample mean of the number of autos of product type k purchased by consumers in fleet size f regardless of how many autos the consumer bought in total, and $\hat{N}_f = \sum_h N_h\, n_{hf}/n_h$ is a weighted estimator of the number of consumers of fleet size f overall. It may be shown that under assumptions (16)

$$V_\xi(\hat{\Theta}_k(3)) = \sum_h \sum_f \frac{N_h^2}{n_h} \mu_f^2 Q_{f|h} (1 - Q_{f|h})$$
$$- \sum_h \sum_f \sum_{\substack{f' \\ f \neq f'}} \frac{N_h^2}{n_h} \mu_f \mu_{f'} Q_{f|h} Q_{f'|h}$$
$$+ \sum_h \sum_f \frac{N_h^2}{n_h} \frac{\sigma_f^2 Q_{f|h}}{\sum_h n_h Q_{f|h}}$$
$$\Bigg\{ (1 - Q_{f|h}) + n_h Q_{f|h}$$
$$+ \left[\frac{1 + (2n_h - 3) Q_{f|h} - 2(n_h - 1) Q_{f|h}^2}{\sum_h n_h Q_{f|h}} \right] \Bigg\}. \qquad (18)$$

If assumptions (16) were plausible then $\bar{y}_{kf}$ would be based on larger sample sizes than $\bar{y}_{kfm}$ in (14) and hence $\hat{\Theta}_k(3)$ would be more stable.

The maximum likelihood estimator of the market share for product type k, R_k, under assumption (16), is given by

$$\hat{\Omega}_k(3) = \frac{\sum_f \hat{N}_f \bar{y}_{kf}}{\sum_f \hat{N}_f \bar{z}_f}, \qquad (19)$$

where $\bar{z}_f$, defined analogously to $\bar{y}_{kf}$, is the unweighted sample mean number of autos of any kind purchased by consumers in fleet category f.

5. EMPIRICAL RESULTS

5.1 Estimating Consumers

In Section 4.2 the efficiency of $\hat{\Theta}_k'(2)$ was investigated for various population structures when assumption (5) held. Readers may find this measure unconvincing since (5) will not hold in practice. We now use the actual survey data to compute $\hat{\Theta}_k'(2)$ for a particular categorization of the conditioning factor that is defined by a combination of the fleet size *and* whether or not the consumer purchased any autos of any kind for fleet use (see Table 4). Empirical evaluations of synthetic estimators have been carried out by Schaible, Brock and Schnack (1977) and Drew, Singh and Choudhry (1982) in different contexts.

For each of the products A-G listed in Table 2 a χ^2 test was used to test the hypothesis that, conditional on the category of the conditioning factor (f), whether or not a consumer purchases that product is independent of stratum (h). Note that for our example the design is stratified random sampling and standard multinomial assumptions apply. For multistage designs, the standard χ^2 analysis would have to be adjusted by using Rao-Scott adjustments for example. In practice it is difficult to find a categorization f such that conditional independence assumptions (5) hold for every product type. However, for the categorization defined in Table 4 it was found that

Table 4
Definition of the Categories, f, of the Conditioning Factor

Categories f	Definition of f	
	Fleet Size	Fleet Purchases
1	Any	0
2	1-4	> 0
3	5-8	> 0
4	9-15	> 0
5	16-25	> 0
6	26-50	> 0
7	51-100	> 0
8	101-200	> 0
9	201-550	> 0
10	> 550	> 0

most of the variability in the probability of purchasing a particular product type was explained by the category f of the conditioning factor and very little of the residual variation was due to differences in strata.

The model-based estimates for consumers, $\hat{\theta}'_k(2)$ and $\hat{\Omega}'_k(2)$, obtained from (6) and (9) respectively, are given in Table 5. The model-based variances may give an optimistic view of the precision of the estimators since they depend on the conditional independence assumptions in the model which may be untrue in practice. Alternatively the usual survey estimate of the p-based variance of the model-based estimator may be derived (see Holt and Holmes 1993). This requires no distributional or conditional independence assumptions of any kind and might be considered a more objective measure. These estimates of standard errors are given in Table 5. Since the estimated standard errors are design-based, they include finite population corrections. [We note here that the model-based standard errors for $\hat{\theta}'_k(2)$ (not shown in Table 5) were consistently around 10% smaller than the p-based standard errors].

Table 5

Model-Based Estimates with p-Based Standard Errors for Selected Products

Product (k)	Estimating Consumers		Estimating Autos	
	Total $\hat{\theta}'_k(2)$	Penetration $\hat{\Omega}'_k(2)$	Total $\hat{\theta}_k(3)$	Share $\hat{\Omega}_k(3)$
A	63,433	.4070	263,511	.3722
	(2,230)	(.0105)	(13,007)	(.0048)
B	39,673	.2546	177,067	.2501
	(1,587)	(.0086)	(9,530)	(.0046)
C	21,930	.1407	65,357	.0923
	(1,142)	(.0066)	(3,836)	(.0027)
D	13,422	.0861	22,146	.0313
	(868)	(.0052)	(1,351)	(.0016)
E	7,366	.0473	15,798	.0223
	(675)	(.0041)	(1,223)	(.0014)
F	5,826	.0374	14,398	.0203
	(492)	(.0031)	(1,113)	(.0012)
G	7,686	.0493	11,207	.0158
	(633)	(.0039)	(813)	(.0011)

Row 1: estimate Row 2: p-based s.e.

Comparing these results with the usual survey results given in Table 2 we find that the standard errors for estimating totals are considerably smaller - around 30-40% smaller for all products except A and B (the major manufacturers) where the reduction is about 15-20%. This pattern is expected since the original survey design was optimal for the total sales of autos and therefore relatively efficient for products with a large market share. We expect the products with smaller market shares to benefit most from the model-based approach.

For estimating market penetration the reduction in standard error is again about 30-40% with slightly smaller reductions for products A and B.

5.2 Estimating Autos

Table 5 also contains model-based estimates for the total number of autos purchased of type k and the corresponding market share, $\hat{\theta}_k(3)$ and $\hat{\Omega}_k(3)$ as defined by (17) and (19) respectively, for the *same* categorization f of the conditioning factor as given in Table 4. P-based standard errors for these estimates are also presented in Table 5.

Comparing with the standard survey estimates given in Table 2 large reductions in standard errors for estimating totals are obtained (40-80%) apart from product type B. Similarly, for estimating the market shares the reduction in standard error is again substantial.

6. DISCUSSION

The model-based estimators are derived using conditional independence assumptions to partition the estimation problem into two components. The first, an estimate of N_f (the number of consumers of fleet size f), makes use of the unequal selection probabilities, whereas the second, an estimate of the proportion of consumers of fleet size f buying product type k (or the average number of autos of product type k purchased by consumers of fleet size f) does not. This can result in a substantial efficiency gain.

If the conditional independence assumptions are invalid then in ordinary design-based terms the estimators will have a residual bias but this may be an acceptable risk to achieve stability of the estimators over the whole product range. For the numerical results in previous sections, only the model-based estimates for product B are outside of the 95% confidence interval based on the direct survey estimator. The conditional independence assumptions will depend on the choice of the categories f, and can be tested using chi-square tests for contingency tables.

Whilst the results in Table 5 show that the design-based standard errors for the model-based estimates are generally smaller than for the direct estimates shown in Table 2, it may be argued that the model-based estimators may be biased and hence provide no gain in terms of mean-squared error (MSE). The bias will arise from the inappropriateness of the conditional independence assumptions (*e.g.*, equation (5)). This is not testable, but a comparison of Tables 2 and 5 can give some insight into the size of bias that would be required to cause the MSE to be the same for both the direct and the model-based estimators. Consider the estimate of total consumers for product E which is strongly affected by the procedure and hence perhaps most susceptible to bias. The variance (and hence MSE) of the direct estimator is $1{,}146^2 = 1{,}313{,}316$ whereas for the model-based estimator the variance is $675^2 = 455{,}625$. Hence, the model-based estimate of 7,366 would need a bias of 926 in order for the MSEs to be the same.

ACKNOWLEDGEMENTS

The authors would like to thank the referees for their helpful comments.

REFERENCES

DREW, J.D., SINGH, M.P., and CHOUDHRY, G.H. (1982). Evaluation of small area techniques for the Canadian Labour Force Survey. *Survey Methodology*, 8, 19-47.

FAY, R.E., and HERRIOT, R.A. (1979). Estimates of income for small places: An application of James-Stein procedures to census data. *Journal of the American Statistical Association*, 74, 269-277.

GHOSH, M., and RAO, J.N.K. (1994). Small area estimation: An appraisal. To appear in *Statistical Science*.

GONZALEZ, M.E. (1973). Use and evaluation of synthetic estimators. *Proceedings of the Social Statistics Section, American Statistical Association*, 33-36.

GONZALEZ, M.E., and HOZA, C. (1978). Small area estimation with application to Unemployment and Housing Estimates. *Journal of the American Statistical Association*, 73, 7-15.

HOLT, D., and HOLMES, D.J. (1993). Small domain estimation for unequal probability survey designs. Working Paper Series, No. 2, Department of Social Statistics, University of Southampton, UK.

HOLT, D., and SMITH, T.M.F. (1979). Poststratification. *Journal of the Royal Statistical Society*, Ser. A, 142, 33-46.

PLATEK, R., RAO, J.N.K., SÄRNDAL, C.-E., and SINGH, M.P. (1987). *Small Area Statistics*. New York: John Wiley and Sons.

SÄRNDAL, C.-E., and HIDIRIGLOU, M.A. (1989). Small domain estimation: a conditional analysis. *Journal of the American Statistical Association*, 84, 266-275.

SCHAIBLE, W.L., BROCK, D.B., and SCHNACK, G.A. (1977). An empirical comparison of the simple inflation, synthetic and composite estimators for small area statistics. *Proceedings of the Social Statistics Section, American Statistical Association*, 1017-1021.

Survey Methodology, June 1994
Vol. 20, No. 1, pp. 33–43
Statistics Canada

Time Series EBLUPs for Small Areas Using Survey Data

A.C. SINGH, H.J. MANTEL and B.W. THOMAS[1]

ABSTRACT

In estimation for small areas it is common to borrow strength from other small areas since the direct survey estimates often have large sampling variability. A class of methods called composite estimation addresses the problem by using a linear combination of direct and synthetic estimators. The synthetic component is based on a model which connects small area means cross-sectionally (over areas) and/or over time. A cross-sectional empirical best linear unbiased predictor (EBLUP) is a composite estimator based on a linear regression model with small area effects. In this paper we consider three models to generalize the cross-sectional EBLUP to use data from more than one time point. In the first model, regression parameters are random and serially dependent but the small area effects are assumed to be independent over time. In the second model, regression parameters are nonrandom and may take common values over time but the small area effects are serially dependent. The third model is more general in that regression parameters and small area effects are assumed to be serially dependent. The resulting estimators, as well as some cross-sectional estimators, are evaluated using bi-annual data from Statistics Canada's National Farm Survey and January Farm Survey.

KEY WORDS: Composite estimation; State space models; Kalman filter; Fay-Herriot estimator.

1. INTRODUCTION

There exists a considerable body of research on small area estimation using cross-sectional survey data in conjunction with supplementary data obtained from census and administrative sources. A good collection of papers on this topic can be found in Platek, Rao, Särndal and Singh (1987). Small area estimation techniques in use in U.S. federal statistical programs are reviewed by the Federal Committee on Statistical Methodology (1993). The basic idea underlying all small area methods is to borrow strength from other areas by assuming that different areas are linked via a model containing auxiliary variables from the supplementary data. It would also be important to borrow strength across time because many surveys are repeated over time. Recently time series methods have been employed to develop improved estimators for small areas; see Pfeffermann and Burck (1990) and Rao and Yu (1992). It is interesting to note that after the initiative of Scott and Smith (1974) on the application of time series methods to survey data, there has only lately been a resurgence of interest in developing suitable estimates of aggregates from complex surveys repeated at regular time intervals; see *e.g.*, Bell and Hillmer (1987), Binder and Dick (1989), Pfeffermann (1991), and Tiller (1992).

In this paper we consider some natural generalizations of the best linear unbiased predictor (BLUP) for small areas when a time series of direct small area estimates is available. An important example of the BLUP for small areas is the Fay-Herriot (FH) estimator, which entails smoothing of direct estimators by cross-sectional modelling of small area totals. The resulting estimators are composite estimators (*i.e.*, convex combinations of direct and synthetic estimators) and are called empirical BLUPs, or EBLUPs, whenever estimates of some variance components are substituted in the BLUPs. The work of Fay and Herriot (1979) represents an important milestone in the field of small area estimation because it is probably the first example of a large scale application of small area estimation by government agencies for policy analysis. With the use of structural models, we derive time series EBLUPs which combine both cross-sectional and time series data. The models underlying the time series EBLUPs were chosen on the basis of general heuristic considerations rather than formal model testing procedures. Formal testing of these types of models with survey data is very difficult and not very much is available. Instead, we begin with a regression model that is reasonable for the larger area, and then allow random small area effects to account for any local deviations from the global model. The regression parameters and random small area effects are allowed to evolve over time according to a state space model that was also formulated heuristically. We have not considered here the problem of mean squared error (MSE) estimation for our estimators. MSEs with respect to the motivating models could be defined and estimated for many of the estimators; however, the focus of this paper is on the performance of the estimators in a repeated sampling framework. MSE estimation is an important and difficult problem, and the availability of reliable MSE estimators could be an important consideration in the choice of estimators.

[1] A.C. Singh and H.J. Mantel, Social Survey Methods Division; B.W. Thomas, Business Survey Methods Division, Statistics Canada, Ottawa, Ontario, K1A 0T6.

The main purpose of this paper is to compare time series EBLUPs with cross-sectional estimators such as post-stratified domain, synthetic, FH and sample size dependent estimators. In the time series modelling of the direct small area estimates we assume that the survey errors are uncorrelated over time. When survey errors are correlated over time and can be modelled reasonably (*e.g.*, ARMA) the approach of Pfeffermann (1991) can be used to obtain time series EBLUPs via the Kalman filter. Rao and Yu (1992) obtain EBLUPs for a model, in which the Kalman filter cannot be applied, with survey errors having arbitrary correlation structure over time but being uncorrelated across areas. They also develop second order approximations to, and estimation of, the mean squared error under their model. When a model for the correlated survey errors is difficult to specify it may be possible, using a suitably modified Kalman filter, to get good sub-optimal estimators (Singh and Mantel 1991).

In this paper we report on an empirical study of the efficiency of time series EBLUPs. The study uses Monte Carlo simulations from real time series data obtained from Statistics Canada's biannual farm surveys. The main findings of the study are

(i) There can be reasonable gains in efficiency with time series EBLUPs over cross-sectional estimators.

(ii) Within the class of time series methods considered in this paper, introduction of serial dependence in the random small area effects is found to be beneficial.

(iii) Although any smoothed version of the direct small area estimator is expected to be biased, the time series EBLUPs exhibit less bias than cross-sectional smoothing methods.

Section 2 contains a description of various cross-sectional methods for small area estimation. Time series EBLUPs are described in Section 3 and the details and results of the Monte Carlo comparative study are given in Section 4. Finally, Section 5 contains concluding remarks.

2. METHODS BASED ON CROSS-SECTIONAL DATA

In this section we describe some well known small area estimation methods that use survey data from only the current time. Ghosh and Rao (1994) contains a good survey of various small area estimators.

Let $\underset{\sim}{\Theta}$ denote the vector of small area population totals Θ_k, $k = 1, \ldots, K$. In this section, which deals with methods based on cross-sectional data, we ignore the dependence of $\underset{\sim}{\Theta}$ on time t for simplicity.

2.1 Method 1 (Expansion Estimator for Domains)

This estimator is given by

$$g_{1k} = \sum_{j \epsilon s_k} d_j y_j,$$

where d_j is the survey weight for sample unit j. For stratified simple random sampling, which is used for our simulation study in Section 4, we have

$$g_{1k} = \sum_h (N_h/n_h) \sum_{j \epsilon s_{hk}} y_{hj}, \tag{2.1}$$

where y_{hj} is the j-th observation in the h-th stratum, s_{hk} denotes the set of n_{hk} sample units falling in the k-th small area in the h-th stratum and n_h, N_h denote respectively the sample and population sizes for the h-th stratum. This estimator is often unreliable because n_{hk}, the random sample size in the small area, may be small in expectation and could have high variability. Conditional on the realized sample size n_{hk}, g_{1k} is biased. However, unconditionally, it is unbiased for Θ_k.

2.2 Method 2 (Post-stratified Domain Estimator)

We will also refer to this estimator as the direct small area estimator. If the population size N_{lk} is known for some post-strata indexed by l, then the efficiency of the estimator g_{1k} could be improved by post-stratification. We define

$$g_{2k} = \sum_l N_{lk} \sum_{j \epsilon s_{lk}} d_j y_j \Big/ \sum_{j \epsilon s_{lk}} d_j = \sum_l N_{lk} \bar{y}_{lk}.$$

In our simulations our post-strata are the intersections of design strata with small areas which leads to

$$g_{2k} = \sum_h (N_{hk}/n_{hk}) \sum_{j \epsilon s_{hk}} y_{hj} = \sum_h N_{hk} \bar{y}_{hk}. \tag{2.2}$$

This estimator also may not be sufficiently reliable because of the possibility of n_{hk}'s being small in expectation. If $n_{hk} = 0$, the above estimator is not defined. It is conventional to replace $\bar{y}_{hk}$ by 0 when $n_{hk} = 0$. In the empirical study presented in this paper, we replaced $\bar{y}_{hk}$ by the synthetic estimate $(\bar{X}_{hk}/\bar{X}_h)\bar{y}_h$, where X is a suitable covariable, whenever $n_{hk} = 0$.

The estimator g_{2k} in (2.2) is conditionally (given $n_{hk} > 0$) unbiased and approximately unconditionally unbiased. Appendix A.1 gives details of estimation of the conditional mean squared error, v_k, of g_{2k}.

2.3 Method 3 (Synthetic Estimator)

It is possible to define a more efficient estimator by assuming a model which allows for "borrowing strength" from other small areas. This gives rise to synthetic estimators, see *e.g.*, Gonzalez (1973) and Ericksen (1974). Suppose different small area totals are connected via the auxiliary variable X_k by a linear model as

$$\Theta_k = \beta_1 + \beta_2 X_k, \; k = 1, \ldots, K, \tag{2.3a}$$

or in matrix notation

$$\underset{\sim}{\Theta} = F\underset{\sim}{\beta}, \qquad (2.3b)$$

where $F = (F_1, F_2, \ldots, F_K)'$, $F_k = (1, X_k)'$. Now consider a model for the direct small area estimators g_{2k}'s as

$$\underset{\sim}{g}_2 = F\underset{\sim}{\beta} + \underset{\sim}{\epsilon},$$

where $\underset{\sim}{g}_2 = (g_{21}, \ldots, g_{2K})'$, $\underset{\sim}{\epsilon} = (\epsilon_1, \ldots, \epsilon_K)'$, ϵ_ks are uncorrelated survey errors with mean 0 and variance v_k. Note that the g_{2k}s are uncorrelated over areas since they are conditionally (given n_{hk}) unbiased and the samples in different small areas are conditionally independent.

Denoting by $\hat{\underset{\sim}{\beta}}$ the weighted least squares (WLS) estimate of $\underset{\sim}{\beta}$, we obtain the regression-synthetic estimator of Θ_k under the assumed model as

$$\underset{\sim}{g}_3 = F\hat{\underset{\sim}{\beta}}.$$

The above estimator could be heavily biased unless the model (2.3) is satisfied reasonably well. The above model may not be realistic because no random fluctuation or random small area effect (a_k, say) is allowed.

2.4 Method 4 (Fay-Herriot Estimator or EBLUP)

Using the empirical Bayes approach of Fay and Herriot (1979) or the more general best linear unbiased predictor approach (see *e.g.*, Battese, Harter and Fuller 1988, and Pfeffermann and Barnard 1991), the bias of the synthetic estimator can be reduced considerably by using a composite estimator; for an early reference on composite estimation see Schaible (1978). The composite estimator is obtained as a convex combination of $\underset{\sim}{g}_2$ and a modified $\underset{\sim}{g}_3$. For this purpose, it is assumed that

$$\underset{\sim}{\Theta} = F\underset{\sim}{\beta} + \underset{\sim}{a}, \qquad (2.4)$$

where a_k's are uncorrelated random small area effects with mean 0 and variance w_k known up to a constant. In our empirical study later we take $w_k = w$. Thus we model $\underset{\sim}{g}_2$ as

$$\underset{\sim}{g}_2 = F\underset{\sim}{\beta} + \underset{\sim}{a} + \underset{\sim}{\epsilon}. \qquad (2.5)$$

Here $\underset{\sim}{a}$ is also assumed to be uncorrelated with $\underset{\sim}{\epsilon}$. The BLUP of $\underset{\sim}{\Theta}$ under the model defined by (2.4) and (2.5) is

$$\begin{aligned} \underset{\sim}{g}_4 &= \underset{\sim}{g}_3^* + \Lambda(\underset{\sim}{g}_2 - \underset{\sim}{g}_3^*) \\ &= \Lambda\underset{\sim}{g}_2 + (I - \Lambda)\underset{\sim}{g}_3^*, \end{aligned} \qquad (2.6)$$

where

$$\Lambda = (V^{-1} + W^{-1})^{-1}V^{-1} = WU^{-1},\ U \equiv V + W,$$

$$V = \mathrm{diag}(v_1, \ldots, v_K), \quad W = \mathrm{diag}(w_1, \ldots, w_K),$$

and $\underset{\sim}{g}_3^* = F\underset{\sim}{\beta}^*$, $\underset{\sim}{\beta}^*$ is the WLS estimate of $\underset{\sim}{\beta}$ under model (2.5). Here it is assumed that both the covariance matrices V and W are known in computing the BLUP.

The expression (2.6) follows from the general results on linear models with random effects, see *e.g.*, Rao (1973, p. 267) and Harville (1976). The BLUP or BLUE of $F\underset{\sim}{\beta}$ is $\underset{\sim}{g}_3^*$ and the BLUP of $\underset{\sim}{a}$ is $\Lambda(\underset{\sim}{g}_2 - \underset{\sim}{g}_3^*)$. It may be of interest to note that the structure of the BLUP does not change regardless of whether or not $\underset{\sim}{\beta}$ is known. However, its MSE does change as expected due to estimation of $\underset{\sim}{\beta}$.

When V and W are replaced by estimates, the estimator $\underset{\sim}{g}_4$ is termed EBLUP. Note that the model (2.4) is more realistic than (2.3), and therefore, the performance of $\underset{\sim}{g}_4$ is expected to be quite favourable. The estimator $\underset{\sim}{g}_4$ approaches $\underset{\sim}{g}_2$ when the v_ks get small, *i.e.*, when the n_{hk}s become large. However, it remains biased, in general, conditional on $\underset{\sim}{\Theta}$, with bias tending to 0 as the v_ks get small.

2.5 Method 5 (Sample Size Dependent Estimator)

An alternative composite estimator is given by the sample size dependent estimator of Drew, Singh and Choudhry (1982). It is defined as

$$\underset{\sim}{g}_5 = \Delta\underset{\sim}{g}_2 + (I - \Delta)\underset{\sim}{g}_3,$$

where $\Delta = \mathrm{diag}(\delta_1, \ldots, \delta_K)$,

$$\delta_k = \begin{cases} 1 & \text{if } \sum_{j \epsilon s_k} d_j \geq \lambda N_k, \\ \sum_{j \epsilon s_k} d_j/\lambda N_k & \text{otherwise} \end{cases} \qquad (2.7)$$

and the parameter λ is chosen subjectively as a way of controlling the contribution of the synthetic component. The above estimator takes account of the realized sample size n_{hk}'s and if these are deemed to be sufficiently large according to the condition in (2.7), then it does not rely on the synthetic estimator. This property is somewhat similar to that of $\underset{\sim}{g}_4$; however, unlike $\underset{\sim}{g}_4$, the above estimator does not take account of the relative sizes of the within area and between area variation. Rao and Choudhry (1993) have demonstrated empirically how EBLUPs can sometimes outperform sample size dependent estimators, especially when the between area variation is not large relative to the within area variation. Särndal and Hidiroglou (1989) also proposed estimators similar to the above sample size dependent estimator.

3. METHODS BASED ON POOLED CROSS-SECTIONAL AND TIME SERIES DATA

Suppose information is available for several time points, $t = 1, \ldots, T$, in the form of direct small area estimators g_{2t}, where g_{2t} is the vector of estimates g_{2k} in (2.2) based on data from time t, and also the small area population totals for the auxiliary variable. We will now introduce some estimators which generalize the Fay-Herriot estimator g_{4T} in different ways by taking account of the serial dependence of the direct estimates $\{g_{2t} : t = 1, \ldots, T\}$. Recall that for the Fay-Herriot estimator, the model for θ_T has two components, namely, the structural component $F_T \beta_T$ and the area component a_T. The estimator g_{4T} borrows strength over areas for the current time T and is given by the sum of two components, each being EBLUP (BLUE) for the corresponding random (fixed) effect, *i.e.*,

$$g_{4T} = F_T \beta_T^* + a_T^*. \tag{3.1}$$

Methods based on time series data could, however, borrow strength over time as well. Here we introduce three estimators which are motivated from specific structural models for serial dependence. All three of these estimators are optimal under different special cases of a structural time series model for the direct small area estimates $\{g_{2t} : t = 1, \ldots, T\}$ specified by the following state space model. Let α_t denote $(\beta_t', a_t')'$ and H_t denote (F_t, I). Then we have

$$\begin{aligned} g_{2t} &= \theta_t + \epsilon_t, \\ \theta_t &= F_t \beta_t + a_t \equiv H_t \alpha_t \end{aligned} \tag{3.2a}$$

and

$$\alpha_t = G_t \alpha_{t-1} + \zeta_t, \tag{3.2b}$$

where

$$G_t = \begin{pmatrix} G_t^{(1)} & 0 \\ 0 & G_t^{(2)} \end{pmatrix}, \quad \zeta_t = \begin{pmatrix} \xi_t \\ \eta_t \end{pmatrix}, \tag{3.2c}$$

along with the usual assumptions about random errors, *i.e.*, ϵ_t, ζ_t are uncorrelated, ζ_t is uncorrelated with α_s for $s < t$, and that $\epsilon_t \sim (0, V_t)$, $\zeta_t \sim (0, \Gamma_t)$ where $\Gamma_t = \text{block diag}\{B_t, Q_t\}$. The covariance matrices V_t, B_t, and Q_t are generally diagonal. If $G_t^{(1)} = I$ and $G_t^{(2)} = I$ then β_t and a_t evolve according to a random walk.

This model is in the general class defined by Pfeffermann and Burck (1991) using structural time series models. The main purpose of their study was to show how accounting for cross-sectional correlations between neighbouring small areas (in addition to serial correlations) and inclusion of certain robustness modifications (to protect against model breakdowns) could improve the performance of time series model based estimators. They also used the maximum likelihood method under normality to estimate model parameters. The focus of this paper, on the other hand, is on the Monte Carlo evaluation of a special class of time series estimators (related to Fay-Herriot) chosen on the basis of heuristic considerations and not on the basis of model fitting. The methods considered could, therefore, be viewed as model assisted methods whose performance will be evaluated in a design based (*i.e.*, repeated sampling) framework by Monte Carlo simulation. Moreover, it will be seen later that, for the types of serial dependence considered, the model parameters can be estimated relatively simply by the method of moments, without making any distributional assumptions such as normality.

To find the optimal estimator (BLUP) of θ_T in (3.2) based on all the direct estimates up to time T, we first found the BLUP $\tilde{\alpha}_T$ of α_T from which the BLUP of θ_T is obtained as $H_T \tilde{\alpha}_T$. It is possible, albeit cumbersome, to get $\tilde{\alpha}_T$ directly from the complete data using the theory of linear models with random effects. However, since the α_Ts are connected over time according to the transition equation (3.2b), it is more convenient to compute it recursively using the Kalman filter (KF). Traditionally KF is viewed as a Bayesian technique in which at each time t, the posterior distribution of α_t given data up to $t - 1$ is updated to get the posterior distribution of α_t given data up to time t. Although it is instructive to view KF in this manner, it is not necessary under mixed linear models. Suppose $\tilde{\alpha}_{T|s}$ denotes the BLUP of α_T based on data up to time s, $s < T$. It is known (see Duncan and Horn 1972) that, for the special structure of serial dependence considered here, the BLUP $\tilde{\alpha}_T$ of α_T based on data up to time T is the same as the BLUP of α_T based on $\tilde{\alpha}_{T|s}$ and the last $T - s$ observations. In other words, information in the previous data can be condensed into an appropriate BLUP before augmenting more current data points. A good description of the Kalman filter is given in chapter 3 of Harvey (1989).

3.1 Method 6 (Time Series EBLUP-I)

For the first estimator, we let β_t evolve over time (*e.g.*, according to a random walk), but assume that a_t is serially independent. The equations for the state space model for this case are similar to (3.2) except that the serial independence of the a_ts implies $G_t^{(2)} = 0$. This will give rise to a composite estimator

$$g_{6T} = F_T \tilde{\beta}_T + \tilde{a}_T. \tag{3.3}$$

Note that $\tilde{\beta}_T$ in (3.3) would now be based on all the small area estimates up to time T and therefore would be different from β_T^* of (3.1) which is based on only direct estimates at time T. The estimator $\tilde{a}_T$, as a result, would also be different from the corresponding component a_T^* of (3.1).

In the simulation study described later we take $G_t^{(1)} = I$, $B_t = \mathrm{diag}(\gamma_1^2, \gamma_2^2)$, corresponding to a random walk model, and $Q_t = \tau^2 I$. Appendix A.2 illustrates the method of moments estimation of the parameters γ_1^2, γ_2^2, and τ^2. The KF may then be run, with initial values for $\tilde{\utilde{\alpha}}_1$ and its MSE obtained from the *FH* estimator at $t = 1$, to obtain the EBLUP of $\tilde{\utilde{\alpha}}_T$. Then $H_T \tilde{\utilde{\alpha}}_T$ is the time series EBLUP-I estimator g_{6T} at time T.

As pointed out by a referee, when the number of small areas is quite large, or when the variation in $\utilde{\beta}_t$ over t is relatively large, there is little difference between $\utilde{g}_{6T}$ and $\utilde{g}_{4T}$. Indeed, there is little difference between the performances of these two estimators in our simulation study described in Section 4.

3.2 Method 7 (Time Series EBLUP-II)

For the second estimator, we let $\utilde{\beta}_t$ be fixed (it may or may not be common for different time points) and let the area effects $\utilde{a}_t$ be serially dependent according to, for example, a random walk. This time series generalization could be viewed as an analogue of the model proposed by Rao and Yu (1992). The resulting composite estimator will have the same form as (3.1), *i.e.*,

$$\utilde{g}_{7T} = F_T \tilde{\utilde{\beta}}_T + \bar{\utilde{a}}_T, \tag{3.4}$$

but the component estimates $\tilde{\utilde{\beta}}_T$ and $\bar{\utilde{a}}_T$ would be different. We have two cases.

3.2.1 Case 1: Suppose the $\utilde{\beta}_t$s are fixed and time-invariant but the $\utilde{a}_t$s are serially dependent. Then, in (3.2), $G_t^{(1)} = I$ and $B_t = 0$. If Q_t is taken as $\tau^2 I$, then the only unknown parameter τ^2 can be estimated by the method of moments; see Appendix A.2. We will denote by $\utilde{g}_{7T}$ the EBLUP obtained in this case when the parameter estimate is substituted.

3.2.2 Case 2: Here we assume that $\utilde{\beta}_t$s are fixed but different for different time points. The area effects $\utilde{\alpha}_t$ evolve over time as in Case 1. In (3.2) we have $G_t^{(1)} = 0$ and $B_t = mI$ where m is a large number. The expressions for $\tilde{\utilde{\alpha}}_T$ and its MSE obtained from the KF in this case give the correct formulas as $m \to \infty$ (see Sallas and Harville 1981). The KF updating equations for $\bar{\utilde{a}}_t$ in this case take the special form

$$\bar{\utilde{\beta}}_t = (F_t' A_t^{-1} F_t)^{-1} F_t' A_t^{-1} (\utilde{g}_{2t} - G_t^{(2)} \bar{\utilde{a}}_{t-1});$$

$$\bar{\utilde{a}}_t = G_t^{(2)} \bar{\utilde{a}}_{t-1} + P_{t|t-1} A_t^{-1} (\utilde{g}_{2t} - G_t^{(2)} \bar{\utilde{a}}_{t-1} - F_t \bar{\utilde{\beta}}_t);$$

$$P_t = P_{t|t-1} - P_{t|t-1} A_t^{-1} (A_t - F_t (F_t' A_t^{-1} F_t)^{-1} F_t') \\ A_t^{-1} P_{t|t-1},$$

where $A_t = P_{t|t-1} + V_t$, P_t is the MSE of $\bar{\utilde{a}}_t$ about $\utilde{a}_t$, and $P_{t|t-1} = G_t^{(2)} P_{t-1} \{G_t^{(2)}\}' + Q_t$ is the MSE of $G_t^{(2)}$ $\bar{\utilde{a}}_{t-1}$ as an estimator of $\utilde{a}_t$. The time series EBLUP in this case will be denoted by $\utilde{g}_{7T}^*$.

3.3 Method 8 (Time Series EBLUP-III)

For the third estimator, we let both $\utilde{\beta}_t$ and $\utilde{a}_t$ evolve over time. This will have more complex serial dependence than either (3.3) or (3.4). Its form will be similar to (3.1) and can be represented as

$$\utilde{g}_{8T} = F_T \tilde{\tilde{\bar{\utilde{\beta}}}}_T + \tilde{\bar{\utilde{a}}}_T. \tag{3.5}$$

As before, if $B_t = \mathrm{diag}\{\gamma_1^2, \gamma_2^2\}$ and $Q_t = \tau^2 I$, then the model parameters τ^2, γ_1^2, γ_2^2 can be estimated by the method of moments as in Appendix A.2. The resulting EBLUP of $\utilde{\theta}_T$ will be denoted by $\utilde{g}_{8T}$.

It may be of interest to note that many of the estimators considered so far are optimal under special cases of the model underlying $\utilde{g}_{8T}$. As has been shown, the time series EBLUPs of methods 6 and 7 result from making restrictions on the matrices G_t and Γ_t. The cross-sectional Fay-Herriot estimators of Section 2.4 result from restricting the data to a single time point. The synthetic estimators of section 2.3 are special cases of the Fay-Herriot estimators with zero variance for the random small area effects, and the direct (post-stratified) estimator is obtained in the limit as the variance of the small area effects goes to infinity.

A further generalization that could be useful is to allow correlations between neighbouring small area effects. This can be accomplished by allowing the matrix Q_t in (3.2) to be non-diagonal; however, it is not clear what would be an appropriate correlation structure in Q_t.

4. MONTE CARLO STUDY

The cross-sectional and time series methods were compared empirically by means of a Monte Carlo simulation from a real time series obtained from Statistics Canada's biannual farm surveys, namely, the National Farm Survey (in June) and the January Farm Survey. Due to the redesign after the census of Agriculture in 1986, the survey data for the six time points starting with the summer of 1988 were employed to create a pseudo-population for simulation purposes. To this, data from the census year 1986 was also added. Thus information at one more time point was available although this resulted in a 3-point gap in the series. The missing data points, however, can be easily handled by time series methods. It may be noted that although the data series is short, it is nevertheless believed to be adequate for illustrative purposes. The parameter of interest was taken as the total number of cattle and calves for each crop district (defined as the small area) at each time point. For simplicity, independent stratified random samples were drawn for each occasion from the pseudo-population, though the farm surveys use rotating panels over time. The dependence of direct small area estimates over time was modelled by assuming that the underlying small area population totals are connected according to some random process. The auxiliary variable used in the model was the ratio-adjusted census 1986 value of the total cattle and calves for each small area. This showed high correlations with the corresponding variable over time at the farm level. Specific details of the empirical study are described below.

4.1 Design of the Simulation Experiment

First we need to construct a pseudo-population from the survey data over six time points (June 1988, January 1989, ..., January 1991). The actual design involves two frames (list and area) with a one stage stratified sampling from the list frame and a two stage stratified sampling from the area frame, for details see Julien and Maranda (1990). We decided to use survey data from the list frame only because the list frame corresponds to farms existing at the time of Census 1986 and the chosen auxiliary variable for model building was based on Census 1986 information. Moreover, we chose to use the data from the province of Quebec because its area sample is only a minor component of the total sample and the estimated coefficient variation for the twelve crop-districts (*i.e.*, small areas of interest) of this province showed a wide range for the livestock variables. It was decided to avoid variability due to changes in the underlying population over time by retaining only those farms which responded to all the six occasions. Also, farm units who belonged to a multiholding arrangement in any one of the seven time points (including the census) were excluded because of the problems in finding individual farm's data from the multiholding summary record and changes in their reporting arrangement over time.

The various exclusions described above were motivated from considerations of yielding a sharper comparison between small area estimators. The total count of farm units after exclusions was found to be 1,160 out of a total of over 40,000 farms on the list frame. For the pseudo-population, we replicated the 1,160 farm units proportional to their sampling weight so that the total size N of the pseudo-population was 10,362, which was manageable for micro-computer simulation.

The pseudo-population was stratified into four takesome and one take-all strata using Census 1986 count data on cattle and calves as the stratification variable. Although we did not consider alternative stratifications or sample sizes in our simulation study, there is no reason to think that our conclusions would alter significantly if we were to do so. The sigma-gap rule (Julien and Maranda 1990) was used for defining the take-all stratum. To apply the sigma-gap rule we look at the smallest population value greater than the population median where the distance to the next population value, in order of size, is at least one population standard deviation; all units above this point are placed into the take-all stratum. The algorithm of Sethi (1963) was used for determining optimal stratification boundaries for take-some strata. Neyman's optimum allocation was used for sample sizes for strata in order to optimize the precision of the provincial estimate of total count. This resulted in, from a total sample size of 207 (2% sampling rate), allocations of 51, 62, 48 and 35 from takesome strata with 5,001, 3,188, 1,850 and 312 farms, respectively, and the size of the take all stratum was 11. The expected number of sample farms in each small area varied from 4.6 in area 9 up to 27.5 in area 6, with an average of 17.3. The expected number of sample farms with some cattle and calves varied from 3.6 in area 9 to 18.8 in area 3, and the average over the small areas was 11.7. A total of 30,000 simulations were performed. For each simulation, samples were drawn independently for each time point using stratified simple random sampling without replacement. The 30,000 simulations were conducted in 15,000 sets of 2 simulations where each set corresponds to a different vector of realized sample sizes in the twelve small areas within each stratum. This was required to compute certain conditional evaluation measures as described in the next subsection, see also Särndal and Hidiroglou (1989).

4.2 Evaluation Measures

Suppose m simulations are performed in which m_1 sets of different vectors of realized sample sizes in domains (h,k) are replicated m_2 times. The following measures can be used for comparing performance of different estimators at time T. Let i vary from 1 to m_1 and j from 1 to m_2.

(i) Absolute Relative Bias for area k:

$$\mathrm{ARB}_k = | m^{-1} \sum_i \sum_j (\mathrm{est}_{ijk} - \mathrm{true}_k)/\mathrm{true}_k |. \quad (4.1)$$

The average of ARB_k over areas k will be denoted by AARB. We take the absolute relative bias since our primary interest in this study is in an overall measure like AARB; however, in other contexts the actual biases for individual small areas may also be of considerable interest.

The following measure is motivated by a desire to evaluate the conditional performance of estimators, conditional on the vectors of realized sample sizes in domains. It is conventional to measure performance conditional on fixed domain sample sizes; here we consider the standard deviation of the conditional bias, B_{ik}, as a simple summary measure. If this standard deviation is small then the method is robust to variations in the realized sample sizes. Note that the expected value of B_{ik} is just the unconditional bias which is estimated by ARB_k. Let B_k^2 denote the unconditional expected value of B_{ik}^2. We define the following Monte Carlo measure:

(ii) Standard Deviation of Conditional Relative Bias for area k:

$$\mathrm{SDCRB}_k = \left\{ m_1^{-1} \sum_i \left(\hat{B}_{ik}^2 - \hat{C}_{ik}\right) \Big/ \mathrm{true}_k - \mathrm{ARB}_k^2 \right\}^{½};$$

$$\hat{B}_{ik} = m_2^{-1} \sum_j \mathrm{est}_{ijk} - \mathrm{true}_k, \qquad (4.2)$$

$$\hat{C}_{ik} = m_2^{-1}(m_2 - 1)^{-1}\left(\sum_j \mathrm{est}_{ijk}^2 - \left(\sum_j \mathrm{est}_{ijk}\right)^2 \Big/ m_2\right).$$

The correction term $\hat{C}_{ik}$ adjusts for bias in $\hat{B}_{ik}^2$, as an estimate of B_{ik}^2, due to m_2 being finite. $\hat{B}_{ik}^2 - \hat{C}_{ik}$ is conditionally unbiased for B_{ik}^2; it is also unconditionally unbiased for B_k^2. The Monte Carlo average $m_1^{-1}\sum_i(\hat{B}_{ik}^2 - \hat{C}_{ik})$ converges to B_k^2 with probability 1 as $m_1 \to \infty$. $\hat{B}_{ik}^2 - \hat{C}_{ik}$ may be negative for some i, due to finite m_2. For large m_1 the average over i is usually very close to B_k^2; whenever the average is less than ARB_k^2 we set SDCRB_k to 0. ASDCRB will denote the average of SDCRB_k over areas k.

(iii) Mean Absolute Relative Error for area k:

$$\mathrm{MARE}_k = m^{-1} \sum_i \sum_j \left| \mathrm{est}_{ijk} - \mathrm{true}_k \right| / \mathrm{true}_k \qquad (4.3)$$

and AMARE denotes the average of MARE_k over areas.

(iv) Mean Squared Error for area k:

$$\mathrm{MSE}_k = m^{-1} \sum_i \sum_j (\mathrm{est}_{ijk} - \mathrm{true}_k)^2 \qquad (4.4)$$

and AMSE as before denotes the average over areas.

(v) Relative Root Mean Squared Error for area k:

$$\mathrm{RRMSE}_k = \{\mathrm{MSE}_k\}^{½}/\mathrm{true}_k. \qquad (4.5)$$

Again, ARRMSE denotes the average over areas.

The precision (*i.e.*, the Monte Carlo standard error) of each measure depends on m_1, m_2. For all measures except (ii), the optimal choice of m_1, m_2 under the restriction that $m_2 > 1$ is $m_1 = m/2$, $m_2 = 2$, since this minimizes the Monte Carlo standard error. To see this, let A be the average of an evaluation measure from m_2 samples all with the same sample configuration (set of random sample sizes in domains) which we call C. Then the expected value of A conditional on C is a function of C, say $E(C)$, and the conditional variance of A is proportional to m_2^{-1}, say $V(C)/m_2$. The unconditional variance of A is then $V\{E(C)\} + E\{V(C)\}/m_2$, and the overall Monte Carlo variance of an evaluation measure based on m_1 sample configurations replicated m_2 times is $V\{E(C)\}/m_1 + E\{V(C)\}/m_1m_2$ which is minimized, since $m = m_1m_2$ is fixed, by taking m_1 as large as possible. For the second measure, the appropriate choice of m_1, m_2 is less straightforward. In the simulation study, m was chosen as 30,000 and the corresponding values of m_1, m_2 were set at 15,000 and 2.

4.3 Estimators Used in the Comparative Study

There were nine estimators included in the study, namely, g_1 to g_8 and g_7^*, all calculated for time $T = 10$. We used a simple linear regression model for the synthetic component with the auxiliary variable defined as

$$X_{kt} = (\hat{\Theta}_t/\Theta_1)\Theta_{k1}, \qquad (4.6)$$

where Θ_{k1}, Θ_1 respectively denote the population totals for small area k and the province at $t = 1$, *i.e.*, at Census 1986. The estimator $\hat{\Theta}_t$ denotes the post-stratified estimator of Θ_t from the farm survey at time t at the province level. Thus X_{kt} is simply a ratio-adjusted synthetic variable. The variances of error components in the regression model were assumed to be constant over areas. For time series models, it was assumed that the serial dependence was generated by a random walk. The above type of model assumptions have been successfully used in many applications and the main reason for our choice was simplicity. It was hoped, however, that the chosen models might be adequate for our purpose and might illustrate the differential gains with different types of model assisted small area estimators, *i.e.*, both cross-sectional and time series smoothing methods.

Since the Census 1986 data was included in the time series, the direct estimate $\underset{\sim}{g}_{21}$ corresponds to Census 1986 and therefore the survey error $\underset{\sim}{\epsilon}_1$ would be identically $\underset{\sim}{0}$. Moreover, from the definition of X_{kt}, it follows that a reasonable choice of (β_{11}, β_{21}) would be (0,1) which implies that $\underset{\sim}{a}_1$ must be $\underset{\sim}{0}$. Thus the covariance matrices B_t and W_t at $t = 1$ are null and, therefore, the distribution of $\underset{\sim}{\alpha}_t$ at $t = 1$ would not require estimation. The above modification in the initial distribution of $\underset{\sim}{\alpha}_t$ is natural in view of the extra information available from the census. Moreover, since the direct estimates g_{2t} were not available for $t = 2, 3, 4$, equations for estimating model variance components in Appendix A.2 were modified accordingly.

For method 7 (case 1), $\underset{\sim}{\beta}_t$ was assumed to have a common fixed value only for $t \geq 2$ because at $t = 1$, $\underset{\sim}{\beta}_t = (0,1)'$. For the sample size dependent estimator g_5 the parameter λ was taken to be 1.

4.4 Empirical Results

The main findings were listed in Section 1. Here we give some detailed comparisons and some possible explanations. We do not show separate results for g_7^* which performs slightly worse than, though overall similarly to, g_7. The estimators are summarized in Table 1. Figures 1 to 3 and Tables 2 to 4 present some of the empirical results. We have not shown the Monte Carlo standard errors but they were all found to be quite negligible.

Table 1

Summary of Estimators

g_1 - Expansion

g_2 - Post-stratified

g_3 - Synthetic

g_4 - Fay-Herriot

g_5 - Sample Size Dependent

g_6 - Time Series EBLUP-I, βs evolve over time, αs independent over time

g_7 - Time Series EBLUP-II, αs evolve over time, fixed common β

g_8 - Time Series EBLUP-III, βs and αs evolve over time

Table 2 gives the five evaluation measures averaged over small areas, Figure 1 shows plots of the averaged evaluation measures relative to the Fay-Herriot (g_4) value. There is a clear pattern in the behaviour of various measures across different estimators. The direct estimator g_2 does very well with respect to the bias measure (AARB) but does somewhat poorly with respect to the other measures. The cross-sectional smoothing method g_3 (synthetic) does quite poorly with respect to the bias measures. The Fay-Herriot method g_4 performs somewhat better than post-stratified on average with respect to the MSE measure but is much worse in terms of bias. The sample size dependent method g_5 is quite similar to g_2, slightly worse with respect to the bias measures and slightly better with respect to the other measures. The time series methods g_7 and g_8 perform quite well overall, though they are somewhat worse than g_2 with regard to bias. The performance of the time series estimator g_6 is generally between that of Fay-Herriot and the time series estimators g_7 and g_8. For all of the estimators (including the synthetic g_3) the standard deviation of the conditional relative bias (ASDCRB) is appreciable; however, it is smallest for the time series methods. As expected, the expansion estimator g_1 does well with respect to the unconditional bias measure, AARB, but its conditional performance (ASDCRB) is quite poor.

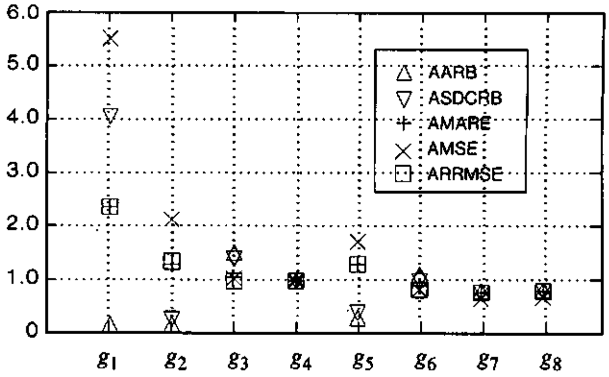


Figure 1. Evaluation Measures Relative to Fay-Herriot
Note: Relative ASDCRB for g_1 (= 18.98) not shown.

Table 2

Average Evaluation Measures

	g_1	g_2	g_3	g_4	g_5	g_6	g_7	g_8
AARB	.001	.007	.097	.065	.018	.070	.053	.053
ASDCRB	.282	.016	.016	.015	.023	.010	.010	.010
AMARE	.269	.147	.115	.108	.136	.097	.087	.088
ARRMSE	.339	.192	.137	.137	.176	.120	.109	.111
AMSE (1,000's)	72,979	27,596	13,382	12,898	22,760	10,603	8,610	8,829

Figure 2 plots averages of RRMSE$_k$ for three size groups, namely small, medium and large small areas, based on the ranking of their true population totals at time T. They are divided up into these three groups because the relative errors of estimation would be expected to be larger for the smaller totals, and the plots do not contradict this expectation. Again, the time series methods g_7 and g_8 perform best. Note that the time series method g_6, which assumes the small area effects to be independent over time, does not do as well. The unaveraged values of RRMSE$_k$ are given in Table 3. RRMSE$_9$ is relatively large because the total number of cattle and calves for area 9 is less than half that of any other small area. Areas 6 and 8 stand out within the medium size small areas as being most difficult to estimate by the smoothing methods. The reason for this is that, while there was an overall decline of about 16% in the total number of cattle and calves in the pseudo-population from June 1986 to January 1991, the decreases for areas 6 and 8 were the furthest from the average at 33% and 1%, respectively, so the ratio adjusted covariate would be least appropriate for those areas. Nevertheless, the time series methods g_7 and g_8 performed significantly better than the post-stratified estimator for areas 6 and 8. This is because the random walk model for the small area effects is able to track small areas which, like areas 6 and 8, progressively deviate from the model.

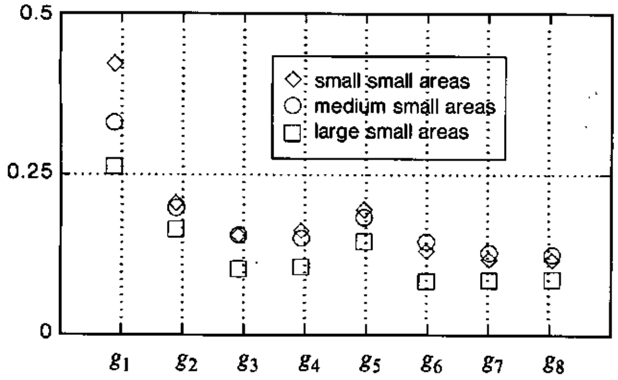


Figure 2. Relative Root Mean Squared Errors: Averaged within Size Groups

Table 3

Relative Root Mean Squared Errors and True Total Cattle and Calves for Small Areas

	Area	True Values	g_1	g_2	g_3	g_4	g_5	g_6	g_7	g_8
Small Size	9	8,502	.580	.277	.342	.275	.277	.199	.160	.174
	10	18,990	.360	.196	.078	.113	.175	.097	.103	.104
	11	18,776	.339	.122	.122	.103	.112	.096	.086	.087
	12	19,819	.409	.237	.076	.152	.212	.123	.117	.117
	Average	16,522	.422	.208	.154	.161	.194	.129	.116	.120
Medium Size	1	27,595	.312	.206	.117	.130	.185	.120	.100	.102
	6	29,012	.306	.241	.256	.216	.224	.224	.168	.172
	7	23,600	.341	.121	.107	.094	.110	.088	.092	.092
	8	23,627	.383	.250	.155	.165	.219	.155	.146	.144
	Average	25,959	.336	.205	.159	.151	.185	.147	.126	.127
Large Size	2	35,592	.268	.171	.113	.110	.156	.096	.089	.088
	3	40,582	.241	.151	.087	.090	.137	.070	.072	.073
	4	42,396	.256	.160	.099	.103	.144	.080	.088	.089
	5	35,996	.270	.176	.091	.097	.160	.088	.085	.088
	Average	38,642	.259	.164	.098	.100	.149	.083	.083	.084

Figure 3 and Table 4 are identical to Figure 2 and Table 3 in format, but show relative biases instead of relative root mean squared errors. The biases for both the expansion estimator g_1 and the post-stratified g_2 are negliglible. For the smoothing methods the average absolute relative biases for medium size small areas are relatively large, mainly because of areas 6 and 8 for which the covariate is least appropriate. Among smoothing methods, the sample size dependent g_5 has the least bias because it is usually very close to the direct g_2; however, it also gains very little over g_2 with respect to mean squared error. Of the remaining smoothing methods the time series estimators g_7 and g_8, which had the smallest mean squared error, also have the smallest bias. Nevertheless, the relative bias of these methods can be quite large, as in areas 6 and 8. In practice it would not be possible to estimate these biases; however, the possible size of the bias could be assessed using simulated sampling from a variety of plausible populations.

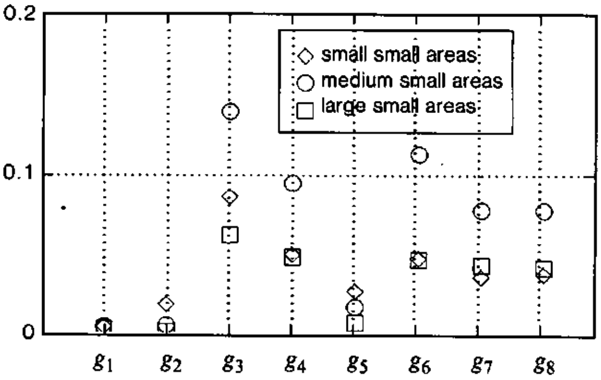


Figure 3. Absolute Relative Biases: Averaged within Size Groups

Table 4

Absolute Relative Biases and True Total Cattle and Calves for Small Areas

	Area	True Values	g_1	g_2	g_3	g_4	g_5	g_6	g_7	g_8
Small Size	9	8,502	.002	.047	.232	.139	.085	.099	.061	.069
	10	18,990	.002	.002	.006	.007	.003	.015	.026	.025
	11	18,776	.002	.009	.090	.052	.021	.062	.039	.037
	12	19,819	.000	.007	.019	.011	.007	.023	.024	.023
	Average	16,522	.001	.016	.087	.052	.029	.050	.037	.039
Medium Size	1	27,595	.001	.003	.093	.063	.007	.078	.044	.045
	6	29,012	.000	.001	.239	.157	.023	.195	.120	.123
	7	23,600	.000	.005	.088	.053	.014	.058	.062	.061
	8	23,627	.002	.008	.143	.106	.024	.124	.093	.091
	Average	25,959	.001	.004	.141	.095	.017	.114	.080	.080
Large Size	2	35,592	.000	.000	.095	.071	.009	.068	.049	.047
	3	40,582	.000	.001	.047	.041	.005	.029	.026	.025
	4	42,396	.001	.002	.066	.056	.008	.044	.057	.056
	5	35,996	.000	.000	.045	.029	.005	.048	.035	.039
	Average	38,642	.000	.001	.063	.049	.006	.047	.042	.042

5. CONCLUDING REMARKS

It was seen by means of a simulation study that small area estimation methods obtained by combining both cross-sectional and time series data can perform better than those based only on cross-sectional data, with respect to both bias and mean squared error. However, the cost in terms of bias could still be substantial. A question of obvious importance is whether it is possible in practical situations to judge if the gains from any type of smoothing would outweigh the costs, and how to make this judgement.

The models for the simulation study were chosen on general considerations. However, in practice, suitable diagnostics similar to those employed in Pfeffermann and Barnard (1991) should be developed for survey data before any model-assisted method can be recommended. It should also be noted that the small area estimators could be modified to make them robust to mis-specification of the underlying model as suggested by Pfeffermann and Burck (1990), see also Mantel, Singh and Bureau (1993). Finally, modification and further extension of the methods presented in this paper to the more realistic case of correlated sampling errors should be investigated in the future.

ACKNOWLEDGEMENT

We would like to thank Jon Rao, Danny Pfeffermann and M.P. Singh for useful discussions and comments on earlier versions of this paper. The comments and suggestions of an anonymous referee and an Associate Editor are also very much appreciated. The first author's research was supported in part by a grant from the Natural Sciences and Engineering Research Council of Canada held at Carleton University.

APPENDIX

A.1 Variance Estimation for g_{2kt}

Let v_{kt} denote the conditional (given n_{hkt}) variance of g_{2kt} in (2.2). Then v_{kt} is given by (whenever $n_{hkt} > 0$ for all h at time t),

$$v_{kt} = \sum_h N_{hkt}^2 \left(n_{hkt}^{-1} - N_{hkt}^{-1} \right) \sigma_{hkt}^2, \qquad \text{(A.1)}$$

where σ_{hkt}^2 is the population variance for the intersection of the h-th stratum with the k-th small area at time t. The variance σ_{hkt}^2 can be estimated by the usual estimator s_{hkt}^2 for $n_{hkt} \geq 2$. Note that the estimate of the conditional variance v_{kt} also provides an estimate of the unconditional variance of g_{2kt}.

If $n_{hkt} = 1$, then we can use a synthetic value as an estimate of σ_{hkt}^2 which can be defined as $\sum (n_{hkt} - 1) s_{hkt}^2 / \sum (n_{hkt} - 1)$, the summation being over all k for which $n_{hkt} \geq 2$ within each (h,t). If $n_{hkt} = 0$, v_{ht} of (A.1) is of course not defined. With the synthetic value of $\bar{y}_{hkt}$ used in this case, we need a synthetic value of its mean squared error. For each (h,t), it can be defined as

$$(\bar{X}_{hkt} / \bar{X}_{ht})^2 (n_{ht}^{-1} - N_{ht}^{-1}) s_{ht}^2 + (\widehat{\text{bias}})^2,$$

where $(\widehat{\text{bias}})^2$ will be taken as

$$\sum_{n_{hlt} > 0} ((\bar{X}_{hlt} / \bar{X}_{ht}) \bar{y}_{ht} - \bar{y}_{hlt})^2 / m_{ht},$$

where m_{ht} is the number of small areas with sample in stratum h at time t.

A.2 Estimation of Variance Components

Using the notation of (3.2), we here illustrate the method of moments for estimating variance components for the model of Section 3.1 in the special case when there is only one auxiliary variable X_{ht}, $Q_t = \tau^2 I$ and $\underset{\sim}{\beta}_t$ follows a random walk, *i.e.*, $G_t^{(1)} = I$. Let $F_t = (F_{1t}, \ldots, F_{Kt})'$, $F_{kt} = (1, X_{kt})'$, $\underset{\sim}{\beta}_t = (\beta_{1t}, \beta_{2t})'$, and $B_t = \text{diag}(\gamma_1^2, \gamma_2^2)$. The parameter τ^2 is estimated by the solution of

$$\sum_{t=1}^{T} \sum_{k=1}^{K} (g_{2kt} - F_{kt}' \hat{\underset{\sim}{\beta}}_t)^2 / (v_{kt} + \tau^2) = T(K - 2).$$

If there is no positive solution, we set $\hat{\tau}^2 = 0$. Here $\hat{\underset{\sim}{\beta}}_t$ denotes the WLS estimate of $\underset{\sim}{\beta}_t$ based on only the cross-sectional data at t. This is analogous to the method used in Fay and Herriot (1979) for cross-sectional data. An estimate of γ_i^2 can be obtained by solving (for $i = 1,2$)

$$\sum_{t=2}^{T} (\hat{\beta}_{i,t} - \hat{\beta}_{i,t-1})^2 / (\gamma_i^2 + d_{ii}^{(t)}) = T - 1,$$

where $d_{ii}^{(t)}$ is the (i,i)-th element of $(F_{t-1}' U_{t-1}^{-1} F_{t-1})^{-1} + (F_t' U_t^{-1} F_t)^{-1}$.

REFERENCES

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

BELL, W.R., and HILLMER, S.C. (1987). Time series methods for survey estimation. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 83-92.

BINDER, D.A., and DICK, J.P. (1989). Modelling and estimation for repeated surveys. *Survey Methodology*, 15, 29-45.

DUNCAN, D.B., and HORN, S.D. (1972). Linear dynamic recursive estimation from the viewpoint of regression analysis. *Journal of the American Statistical Association*, 67, 815-821.

ERICKSEN, E.P. (1974). A regression method for estimating populations of local areas. *Journal of the American Statistical Association*, 69, 867-875.

FAY, R.E., and HERRIOT, R.A. (1979). Estimates of income for small places: an application of James-Stein procedures to census data. *Journal of the American Statistical Association*, 74, 269-277.

FEDERAL COMMITTEE ON STATISTICAL METHODOLOGY (1993). Indirect Estimators in Federal Programs. Statistical Policy Working Paper 21, U.S. Office of Management and Budget.

GHOSH, M., and RAO, J.N.K. (1994). Small Area Estimation: an Appraisal. *Statistical Science*, 9, to appear.

GONZALEZ, M.E. (1973). Use and evaluation of synthetic estimates. *Proceedings of the Social Statistics Section, American Statistical Association*, 33-36.

HARVEY, A.C. (1989). *Forecasting, Structural Time Series Models and the Kalman Filter.* Cambridge: University Press.

JULIEN, C., and MARANDA, F. (1990). Sample design of the 1988 national farm survey. *Survey Methodology*, 16, 117-129.

MANTEL, H.J., SINGH, A.C., and BUREAU, M. (1993). Benchmarking of small area estimators. *Proceedings of the International Conference on Establishment Surveys, Buffalo, June 1993*, 920-925.

PFEFFERMANN, D. (1991). Estimation and seasonal adjustment of population means using data from repeated surveys. *Journal of Business and Economics Statistics*, 9, 163-175.

PFEFFERMANN, D., and BARNARD, C.H. (1991). Some new estimators for small area means with application to the assessment of farmland values. *Journal of Business and Economics Statistics*, 9, 73-84.

PFEFFERMANN, D., and BURCK, L. (1990). Robust small area estimation combining time series and cross-sectional data. *Survey Methodology*, 16, 217-237.

PLATEK, R., RAO, J.N.K., SÄRNDAL, C.-E., and SINGH, M.P. Eds. (1987). *Small Area Statistics: An International Symposium*; New York; John Wiley and Sons.

RAO, J.N.K., and CHOUDHRY, G.H. (1993). Small area estimation: Overview and empirical study. Presented at the International Conference on Establishment Surveys, Buffalo, June 1993, to appear.

RAO, J.N.K., and YU, M. (1992). Small Area Estimation by Combining Time Series and Cross-sectional Data. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 1-9.

SALLAS, W.M., and HARVILLE, D.A. (1981). Best linear recursive estimation for mixed linear models. *Journal of the American Statistical Association*, 76, 860-869.

SÄRNDAL, C.-E., and HIDIROGLOU, M.A. (1989). Small domain estimation: a conditional analysis. *Journal of the American Statistical Association*, 84, 266-275.

SCHAIBLE, W.L. (1978). Choosing weights for composite estimation for small area statistics. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 741-746.

SCOTT, A.J., and SMITH, T.M.F. (1974). Analysis of repeated surveys using time series methods. *Journal of the American Statistical Association*, 69, 674-678.

SETHI, V.K. (1963). A note on optimum stratification of populations for estimating the population means. *Australian Journal of Statistics*, 5, 20-33.

SINGH, A.C., and MANTEL, H.J. (1991) State space composite estimation for small areas. *Proceedings: Symposium 91: Spatial Issues in Statistics*, Statistics Canada, Ottawa, November 1991, 17-25.

TILLER, R. (1992). Time series modelling of sample survey data from the U.S. Current Population Survey. *Journal of Official Statistics*, 8, 149-166.

Survey Methodology, June 1994
Vol. 20, No. 1, pp. 45–52
Statistics Canada

Jackknife Variance Estimation of Imputed Survey Data

JOHN G. KOVAR and EDWARD J. CHEN[1]

ABSTRACT

Imputation is a common technique employed by survey-taking organizations in order to address the problem of item nonresponse. While in most of the cases the resulting completed data sets provide good estimates of means and totals, the corresponding variances are often grossly underestimated. A number of methods to remedy this problem exists, but most of them depend on the sampling design and the imputation method. Recently, Rao (1992), and Rao and Shao (1992) have proposed a unified jackknife approach to variance estimation of imputed data sets. The present paper explores this technique empirically, using a real population of businesses, under a simple random sampling design and a uniform nonresponse mechanism. Extensions to stratified multistage sample designs are considered, and the performance of the proposed variance estimator under non-uniform response mechanisms is briefly investigated.

KEY WORDS: Item nonresponse; Hot deck imputation; Nearest neighbour imputation; Nonrandom nonresponse; Complex survey design.

1. INTRODUCTION

All sample surveys suffer from varying degrees of nonresponse. While total or unit nonresponse is often redressed by appropriate survey weight adjustment, most survey taking organizations resort to imputation in the case of item nonresponse. In this way, plausible values are inserted in place of missing or inconsistent entries, thus simplifying estimation of means and totals at all levels of aggregation. As early as the 1950's however, Hansen, Hurwitz and Madow (1953) recognized that treating the imputed values as observed values can lead to underestimation of variances of these estimators if standard formulae are used; underestimation which becomes more appreciable as the proportion of imputed items increases.

A number of remedies to overcome this problem have been advanced. In particular, Rubin (1987) proposed multiple imputation to estimate the variance due to imputation by replicating the process a number of times and estimating the between replicate variation. More recently, Särndal (1990) outlined a number of model assisted estimators of variance, while Rao and Shao (1992) proposed a technique that adjusts the imputed values to correct the usual or naive jackknife variance estimator. The Särndal, and Rao and Shao methods, are appealing in that only the imputed file (with the imputed fields flagged) is required for variance estimation. No auxiliary files are needed. Särndal's model assisted approach yields unbiased variance estimators, provided the model holds (Lee, Rancourt and Särndal 1991). The Rao and Shao adjusted jackknife method is design consistent as well as model unbiased (Rao 1992). But while the model assisted approach requires different variance estimators for each imputation method, the adjusted jackknife method provides a unified approach that requires the implementation of only one estimator, the jackknife estimator, provided the imputed values are adjusted appropriately during the variance estimation stage.

In this paper we describe a simulation study that evaluates the adjusted jackknife variance estimator of Rao and Shao (1992). In Section 2 we motivate the present empirical study by demonstrating the characteristics of the naive variance estimator under four imputation methods in the case of simple random sampling. In Section 3 we briefly outline the Rao and Shao adjustment procedure and present the empirical results. Extensions to more complex designs and experiments with nonrandom nonresponse mechanisms are elaborated in Section 4. Finally, in Section 5 we offer some concluding remarks and recommendations, including areas for future study.

2. BACKGROUND

Following the notation of Rao (1992), we suppose that in a sample s, of size n, m units respond to item y, while $n - m$ units do not. Denote by y_i^* the imputed value for unit i, $i \epsilon s\text{-}s_r$, where s_r is the set of responding units. The usual estimator of the mean $\bar{Y}$ under simple random sampling, based on the imputed file is given by

$$\bar{y}_I = \frac{1}{n}\left(\sum_{i \epsilon s_r} y_i + \sum_{i \epsilon s\text{-}s_r} y_i^* \right). \qquad (1)$$

[1] John G. Kovar, Business Survey Methods Division; Edward J. Chen, Social Survey Methods Division, Statistics Canada, R.H. Coats Building, Tunney's Pasture, Ottawa, Ontario, Canada K1A 0T6.

2.1 Imputation Methods

In the present simulation study we consider four simple methods of imputation, namely the mean of respondents, ratio, nearest neighbour and hot deck imputation methods. The reader is referred to Kalton and Kasprzyk (1986) for a thorough review of the topic of imputation. The simplest and most intuitive method of imputation, when the interest lies in estimating the mean of the item y, is to impute all missing items with the mean of the observed responding units. The imputed value y_i^*, for unit i, under the mean imputation method, is thus given by

$$y_i^* = \bar{y}_m = \sum_{j \in s_r} y_j / m. \qquad (2)$$

In this case, the estimator of the population mean $\bar{Y}$ in (1) reduces to the estimator $\bar{y}_I = \bar{y}_m$. Due to the fact that this method has the undesirable property of distorting the distributions, it is used in practice usually only as a last resort. It is included here for illustrative purposes.

Secondly, we consider a ratio imputation method based on the assumption that a correlated auxiliary variable x, is available, and that the ratio $\bar{y}/\bar{x}$ is the same in the s_r and $s - s_r$ sets, as would be the case if the nonresponse occurred at random, for example. Under the ratio imputation method, we impute the predicted value in place of the missing y_i as follows:

$$y_i^* = \frac{\bar{y}_m}{\bar{x}_m} x_i, \qquad (3)$$

where $\bar{x}_m$ is the mean of the x values of the respondent set s_r. The estimator of the population mean $\bar{Y}$ in (1) reduces to the double sampling estimator $\bar{y}_I = (\bar{y}_m/\bar{x}_m)\bar{x}$, by considering the respondents as the second phase sample.

The third imputation technique we consider is the nearest neighbour (NN) method. Under this method, the missing value is filled in by an observed value of another unit from the set s_r, whose distance to the nonresponding unit is minimum. In practice the distance functions used are usually the ζ_1, ζ_2, or ζ_∞ Minkowski's norms based on the auxiliary x-variables, assumed observed for all units in s. Thus

$$y_i^* = y_j, j \in s_r, \text{ such that } \| x_i - x_j \| \text{ is minimized}, \qquad (4)$$

where $\| \cdot \|$ is one of the above mentioned norms.

The above three methods are often labelled deterministic, since, given the sample of respondents, the imputed values are determined uniquely. The fourth imputation method considered in this study, the hot deck method (HD), is non-deterministic, since the imputed values are chosen at random from the respondent set. While in practice imputation classes are often created and some sort of sequential procedure is usually implemented, we consider here the pure hot deck, whereby the donor unit (j) is chosen at random, with replacement, from the entire set s_r, that is,

$$y_i^* = y_j, j \in s_r. \qquad (5)$$

2.2 Variance Due to Imputation

Treating the imputed values as observed values, leads to the incorrect variance estimator

$$v_{naive} = (1 - f) s_I^2 / n, \qquad (6)$$

where s_I^2 is the sample variance of the complete sample of responding and imputed values, and $(1 - f)$ is the finite population correction factor ($f = n/N$). It can be easily shown that the true variance of the estimator $\bar{y}_I$ in (1), $V(\bar{y}_I)$, can be written as (Särndal 1990)

$$V(\bar{y}_I) = V_{sam} + V_{imp} + V_{mix}, \qquad (7)$$

where V_{sam} is the sampling variance component, V_{imp} is the variance introduced by the imputation method in question and V_{mix} is a covariance term between V_{sam} and V_{imp} which in most cases is negligible or zero. An estimator of V_{sam} could be obtained by adding to v_{naive} a term to correct for the fact that the standard formula understates the sampling variance component when there are imputed values in the data set. To estimate $V(\bar{y}_I)$, however, an additional component of variance due to the imputation mechanism, V_{imp}, must be estimated. This may be done explicitly, as in Rubin's (1987) multiple imputation, or by modifying common variance formulae as in Särndal (1990) and Rao and Shao (1992). Note that the interest lies in estimating the variance of the estimator at hand, that is, $V(\bar{y}_I)$, not the variance of an estimator that would have been obtained had there been no nonresponse.

2.3 Variance Underestimation

To illustrate the seriousness of the underestimation of $V(\bar{y}_I)$ by v_{naive}, and the dependence of the degree of underestimation on the imputation method, we first describe the simulation study used for this purpose. We consider a data set of 5,620 units with two variables: An auxiliary variable x, the Gross Business Income, available for all units, that can be used as a measure of size, and a related purchase variable y. The correlation between x and y in this particular data set is of the order of 0.92. Simple random samples of size 200 were selected without replacement. A fixed proportion of units were identified at random as nonrespondents, having their y-values deleted and imputed according to one of the four methods described above. Various rates of nonresponse were generated, though, for the most part, we confine our reporting to results based on 5 and 30% nonresponse rates.

To evaluate the performance of the proposed variance estimators, we calculate the percent relative bias of the variance estimator $v.$, given by

$$\text{Rel.Bias}(v.) = \sum_{k=1}^{K} \frac{(v_k - V(\bar{y}_I))/K}{V(\bar{y}_I)} \times 100, \qquad (8)$$

where $V(\bar{y}_I)$ is obtained through simulation, and v_k is the k-th realization of the K simulated variance estimates in question. Similarly, the percent relative stability of the variance estimators is given by

$$\text{Rel.Stab.}(v.) = \sum_{k=1}^{K} \frac{\sqrt{(v_k - V(\bar{y}_I))^2/K}}{V(\bar{y}_I)} \times 100. \quad (9)$$

All simulations were performed on an IBM PC, using Microsoft's Fortran 77, Version 5.0. In the case of simple random sampling, results are based on averages of 100,000 replications (K = 100,000). With this number of replicates, the reported relative bias values were observed not to vary by more than one percentage point. The results are summarized below in Table 1 for the case of 5 and 30% nonresponse rates.

Table 1
Underestimation of Variance of $\bar{y}_I$ by the Naive Estimator Under Four Imputation Methods, and 5 and 30% Nonresponse Rates

Non-response Rate	Variance Estimator	Imputation Method			
		Mean	HD	Ratio	NN
5%	$V(\bar{y}_I)$	9.9	10.3	9.5	9.5
	v_{naive}	8.9	9.4	9.2	9.3
	Rel.Bias(v_{naive})	−10.7%	−9.4%	−2.5%	−2.2%
30%	$V(\bar{y}_I)$	13.5	16.5	10.1	10.3
	v_{naive}	6.5	9.4	8.5	9.0
	Rel.Bias (v_{naive})	−51.4%	−43.4%	−15.3%	−12.8%

First, we note in Table 1, that the naive estimator underestimates the true variance of $\bar{y}_I$ by 10.7% in the case of mean imputation at a 5% level of nonresponse. About half of this underestimation is due to the fact that v_{naive} underestimates V_{sam} and the other half is due to the fact that v_{naive} ignores the component V_{imp}. Särndal (1990) obtains very similar results with respect to the partitioning of the underestimation in the case of mean imputation. Secondly, in the first row of Table 1, the true variance of $\bar{y}_I$ is larger in the case of the hot deck imputation as compared to the mean imputation, due to the procedure's inherent variability (*i.e.*, the V_{imp} component is larger). By contrast, $V(\bar{y}_I)$ is slightly lower in the case of the ratio and nearest neighbour imputation methods, since V_{imp} decreases as the imputation procedure is better able to predict the true unobserved values (Särndal 1990), as is the case in the present study due to the relatively high correlation between the x and y variables. Thirdly, as can be seen in Table 1, $V(\bar{y}_I)$ increases while v_{naive} decreases as the nonresponse rate becomes more elevated. As such, the underestimation of $V(\bar{y}_I)$, when the imputed values are treated as observed values, becomes more serious as the proportion of missing items increases. The problem is more pronounced in the case of the mean and hot deck imputation methods, which do not use auxiliary information. Note that underestimation of variance in the order of 50%, as was observed in this case, can lead to confidence intervals that are about 30% too short and to declaration of significance when none exists. Also of note is the similar behaviour of the ratio and nearest neighbour methods which will be exploited later.

3. JACKKNIFE VARIANCE ESTIMATOR

Let $\bar{y}_I(j)$ be the imputed estimator of $\bar{Y}$ obtained when the j-th unit is deleted from the sample. Then, in the case of simple random sampling, a naive jackknife variance estimator of $\bar{y}_I$ is given by

$$\tilde{v}_J = \frac{n-1}{n} \sum_{j=1}^{n} [\bar{y}_I(j) - \bar{y}_I]^2, \qquad (10)$$

which can be shown to reduce to v_{naive} (Rao 1992).

3.1 Imputed Value Adjustment

In order to produce the "correct" (Rao 1990) jackknife variance estimator, Rao (1992) proposed to adjust the imputed values as described below. Intuitively, the adjustment is necessary whenever a responding unit is deleted from a jackknife replicate, since in the case of most imputation methods, all the imputed values depend directly or indirectly on the observed value that was deleted. This is clear in the case of mean imputation and ratio imputation, where all respondents contribute directly to the mean $\bar{y}_m$, but is less evident in nearest neighbour and hot deck imputation methods where the deleted unit contributes to the imputation process only in the sense that it is not available to be selected as a donor. Thus, whenever a responding unit is deleted, *all* imputed values in the sample must be adjusted before the "delete-one" imputed estimator of the mean is computed. The adjustment must clearly be a function of the imputation method used. In the case of the mean and the hot deck imputation methods, it can be shown that the following adjustment is appropriate (Rao 1992; Rao and Shao 1992). Let $z_i^*(j)$ be the adjusted value of the i-th imputed unit y_i^*, when the j-th unit has been deleted. Then $z_i^*(j)$ is given by

$$z_i^*(j) = \begin{cases} y_i^* + [\bar{y}_m(j) - \bar{y}_m] & \text{if } j \in s_r \\ y_i^* & \text{if } j \in s\text{-}s_r. \end{cases} \quad (11)$$

In other words, no adjustment is necessary if the deleted unit (j), has itself been imputed; that is, unit j is a nonrespondent. In the case of the mean imputation, for example, when $j \in s_r$, the adjusted value reduces to $\bar{y}_m(j)$, the mean of the remaining $m - 1$ respondents, as desired.

The jackknife variance estimator is evaluated by first computing the adjusted imputed estimator $\bar{y}_I^a(j)$, as

$$\bar{y}_I^a(j) = \sum_{\substack{i \in s \\ i \neq j}} z_i^*(j)/(n - 1), \quad (12)$$

and then letting

$$v_J(\bar{y}_I) = \frac{n-1}{n} \sum_{j=1}^{n} [\bar{y}_I^a(j) - \bar{y}_I]^2. \quad (13)$$

It can be shown that the adjusted jackknife variance estimator reduces to the correct variance estimator in the case of the mean imputation (Rao 1990), and provides a consistent estimator in the case of the hot deck imputation (Rao and Shao 1992).

In the case of the ratio imputation, the adjusted values are given by

$$z_i^*(j) = \begin{cases} y_i^* + \left[\dfrac{\bar{y}_m(j)}{\bar{x}_m(j)} x_i - \dfrac{\bar{y}_m}{\bar{x}_m} x_i\right] & \text{if } j \in s_r \\ y_i^* & \text{if } j \in s\text{-}s_r, \end{cases} \quad (14)$$

where $\bar{x}_m(j)$ is the mean of the $m - 1$ sample values of x of the responding units when unit j is deleted. The jackknife variance estimator $v_J(\bar{y}_I)$ is then computed as in (13) above, yielding the correct variance estimator. Furthermore, Rao (1992) shows that not only is the adjusted jackknife variance estimator design consistent (p-consistent) under uniform nonresponse irrespective of the model, but is also design-model unbiased (pm-unbiased) under the model (15) and any nonresponse mechanism that does not depend on the y-values.

$$E_m(y_i) = \beta x_i, \quad V_m(y_i) = \sigma^2 x_i,$$
$$\text{cov}_m(y_i, y_j) = 0 \quad i \neq j \in s. \quad (15)$$

Since the naive variance estimator under the nearest neighbour imputation was observed to behave much like the naive variance estimator under the ratio imputation, the adjustment for the ratio imputation given in (14) was used in the case of the nearest neighbour imputation. As well, an alternate adjustment was considered, whereby unit i was re-imputed using the nearest neighbour method, whenever the deleted unit (j) was used to impute unit i. That is, adjustment takes place only if the deleted unit is a respondent (as above), but only those nonrespondents in the j-th jackknife replicate that were actually imputed using unit j are re-imputed by one of the $m - 1$ remaining donors. (This corresponds to imputing the second nearest neighbour for these units.) We note that no theoretical justification exists for either of these adjustments. Since the latter adjustment performed worse than the ratio adjustment in our examples, and since its eventual implementation in production would be cumbersome, we omitted it from further consideration, even though it was always observed to be conservative.

We would like to stress here that for all imputation methods the adjustments are only performed for the purpose of variance estimation and can be made temporarily while the variance estimation program executes. No permanent adjustments are required on the imputed file used for the estimation of means and totals, though the imputed fields must be flagged appropriately.

3.2 Empirical Results

The jackknife variance estimator with adjustments corresponding to the four imputation methods described above, was computed in addition to v_{naive} in the simulation study outlined in Section 2. Nonresponse rates of 5 and 30% were considered and the relative biases were calculated. They are summarized in Table 2 below.

Table 2

Relative Biases of the Naive Variance Estimator and the Adjusted Jackknife Variance Estimator Under 5 and 30% Nonresponse Rates

Non-response Rate	Variance Estimator	Imputation Method			
		Mean	HD	Ratio	NN
		in percent			
5%	v_{naive}	−10.7	−9.4	−2.5	−2.2
	v_J	2.7	3.6	3.4	3.7
30%	v_{naive}	−51.4	−43.4	−15.3	−12.8
	v_J	3.3	1.9	3.0	5.3

Since the adjusted jackknife variance estimator is design consistent (p-consistent) (Rao 1992), it performs well in the case of the mean, hot deck and ratio imputation under uniform response mechanism, as expected. (Equally good performance was observed with other data sets which do not follow the model (15) as well, but more work is needed on this front.) Of note is the relatively good performance under the nearest neighbour imputation. The proposed estimator tends to be somewhat conservative, due, in small part, to the fact that it does not incorporate the finite population correction.

4. EXTENSIONS

While the adjusted jackknife variance estimator has been shown to perform well in the case of simple random sampling under uniform nonresponse mechanism in one imputation class, we consider here extensions to more complex design, to more than one imputation class, and to nonrandom response mechanisms.

4.1 Complex Designs

In this section we describe a simulation study that evaluates the Rao and Shao (1992) adjusted jackknife variance estimator in comparison to the naive variance estimator, in the case of stratified multistage sampling and hot deck imputation. In particular, data from the Canadian Survey of Consumer Finances (SCF) that follows the design of the Canadian Labour Force Survey will be used. The variable of interest, y, is the total household income. The SCF follows a complex stratified multistage design with the primary sampling units (psu's) in the strata used in this study selected with probability proportional to the number of dwellings. Generally speaking, the psu's are collections of dwellings, corresponding to city blocks in urban areas and to groups of Census Enumeration Areas (EA's) in rural regions. We used as a population a sample of 3,870 households in 30 strata and sampled two psu's in each stratum. As in the case of the simple random sampling study, 5 and 30% uniform nonresponse rates were generated at the household level. The missing values were then imputed using the hot deck imputation method described in Rao and Shao (1992). Briefly, the imputation method consists of selecting the donors from the respondent set with replacement, with probability proportional to the survey weight of the donors.

We first consider the case of a single imputation class. Let y_{hik} be the observed value for the k-th unit in the i-th psu and the h-th stratum ($k = 1, \ldots, n_{hi}, i = 1, \ldots, n_h, h = 1, \ldots, L, n = \sum \sum n_{hi}$), and let y^*_{hik} be the corresponding imputed value whenever the (hik) unit is a nonrespondent, that is, whenever $(hik) \epsilon s\text{-}s_r$. The imputed estimator of Y is then given by

$$\hat{Y}_I = \sum_{(hik)\epsilon s_r} w_{hik}\, y_{hik} + \sum_{(hik)\epsilon s\text{-}s_r} w_{hik}\, y^*_{hik}, \qquad (16)$$

where w_{hik} is the survey weight corresponding to unit (hik). Under the above hot deck imputation scheme, $\hat{Y}_I$ is asymptotically unbiased (Rao and Shao 1992).

The expectation of $\hat{Y}_I$ under the hot deck imputation procedure can be written as (Rao and Shao 1992):

$$E_*(\hat{Y}_I) = \left[\sum_{(hik)\epsilon s_r} w_{hik}\, y_{hik} \Big/ \sum_{(hik)\epsilon s_r} w_{hik}\right] \times \sum_{(hik)\epsilon s} w_{hik}$$

$$= [\hat{S}/\hat{T}] \times \hat{U}, \qquad (17)$$

thus defining the terms $\hat{S}$, $\hat{T}$ and $\hat{U}$. The jackknife "delete-one" values are then given by

$$\hat{S}(gj) = \sum_{\substack{(hik)\epsilon s_r \\ h \neq g}} w_{hik} y_{hik} + \frac{n_g}{n_g - 1} \sum_{\substack{(gik)\epsilon s_r \\ i \neq j}} w_{gik} y_{gik},$$

$$\hat{T}(gj) = \sum_{\substack{(hik)\epsilon s_r \\ h \neq g}} w_{hik} + \frac{n_g}{n_g - 1} \sum_{\substack{(gik)\epsilon s_r \\ i \neq j}} w_{gik}, \qquad (18)$$

whenever the j-th psu in the g-th stratum is deleted. The adjustment of the imputed values is performed whenever the (gj)-th psu is deleted, $(hi) \neq (gj)$, and $(hik) \epsilon s\text{-}s_r$, by letting

$$z^{(gj)}_{hik} = y^*_{hik} + \left[\frac{\hat{S}(gj)}{\hat{T}(gj)} - \frac{\hat{S}}{\hat{T}}\right]. \qquad (19)$$

Then, analogous to (12) and (13), the jackknife variance estimator is evaluated by first computing the adjusted imputed estimator $\hat{Y}^a_I$ when the (gj)-th psu is deleted as

$$\hat{Y}^a_I(gj) = \hat{S}(gj) + \sum_{(hik)\epsilon s\text{-}s_r} w_{hik}\, z^{(gj)}_{hik}$$

$$+ \frac{n_g}{n_g - 1} \sum_{\substack{(hik)\epsilon s\text{-}s_r \\ i \neq j}} w_{gik}\, z^{(gj)}_{gik}, \qquad (20)$$

and then setting

$$v_J(\hat{Y}_I) = \sum_{g=1}^{L} \frac{n_g - 1}{n_g} \sum_{j=1}^{n_g} (\hat{Y}^a_I(gj) - \hat{Y}_I)^2. \qquad (21)$$

It can be shown that v_J as defined in (21), is a consistent estimator of the variance of $\hat{Y}_I$ (Rao and Shao 1992).

We generated 10,000 samples of 60 psu's selected with probability proportional to size, and subjected the selected households to 5 and 30% uniform nonresponse. We then computed the naive variance estimator, and the adjusted jackknife variance estimator, v_J, in (21). The relative bias (8) and the relative stability (9) were computed for both of the variance estimators, and are summarized in Table 3 below.

Table 3

Relative Bias and Relative Stability (in Parentheses) of the Naive Variance Estimator and the Adjusted Jackknife Variance Estimator Under 5 and 30% Nonresponse, in the Case of Stratified Multistage Sampling

Variance Estimator	Nonresponse Rate	
	5%	30%
	in percent	
v_{naive}	−10.3 (88)	−43.7 (84)
v_J	−0.9 (97)	1.2 (124)

As can be seen in Table 3, the naive variance estimator underestimates the true variance of Y at rates comparable to the simple random sampling case (Table 2), with the underestimation becoming more serious as the nonresponse rate increases. The adjusted jackknife variance estimator, on the other hand, performs well at both levels of nonresponse, at a relatively modest cost of a slight decrease in the stability of the variance estimator, as compared to v_{naive}.

4.2 Imputation Classes

Under the same sample design as in Section 4.1, we also considered the case of more than one imputation class as is the case in practice. The household size, known for all households in the sample, was used to form two imputation classes, namely one member households and more than one member households. This was done under the assumption that the propensity to respond is different between these two classes, while uniform response probability was assumed within the imputation classes. Two nonresponse schemes were evaluated. The first assumes a 5% uniform nonresponse in the single member household class and 10% uniform nonresponse in the multiple member household class, while the second scheme assumes rates of 25 and 30% in each of the classes respectively. The hot deck imputation, the imputed value adjustments, and the adjusted total calculations in (20), $\hat{Y}^{a}_{Iv}(gj)$, were performed independently within each imputation class denoted by v. The terms $\hat{Y}^{a}_{Iv}(gj)$ were then summed over the two imputation classes, yielding $\hat{Y}^{a}_{I}(gj)$, which was used in (21) to provide the estimate v_J. The results are summarized in Table 4.

Table 4
Relative Bias and Relative Stability (in Parentheses) of the Naive Variance Estimator and the Adjusted Jackknife Variance Estimator Under Two Nonresponse Schemes, in the Case of Stratified Multistage Sampling and Two Imputation Classes

Variance Estimator	Nonresponse Rate	
	5% and 10%	25% and 30%
	in percent	
v_{naive}	− 16.7 (87)	− 40.2 (84)
v_J	− 1.0 (103)	1.1 (127)

As can be seen in Table 4, the adjusted jackknife variance estimator v_J, performs well under both nonresponse schemes. The results, along with those in Table 3, demonstrate the consistency and the reasonably good stability of the adjusted jackknife variance estimator, even in cases of elevated nonresponse rates.

4.3 Nonrandom Nonresponse

As demonstrated above, the adjusted jackknife variance estimator performs well when the nonresponse is random within imputation classes. To study its robustness against the uniform response mechanism assumption, we use the data set described in Section 2, and generated nonresponse as outlined in Lee, Rancourt and Särndal (1991). In particular, the probability of nonresponse is assumed to be related to the x-variable in two distinct ways:

$$P_L = 1 - \exp(-c_L x), \tag{22}$$

$$P_s = \exp(-c_s x), \tag{23}$$

where the constants c_L and c_S are chosen such that an expected 30% nonresponse rate is achieved. In the model P_L given in (22) the nonresponse is positively correlated with the x-variable, implying that large (L) units are more likely not to respond. The opposite is true in the model P_S given in (23), under which smaller (S) units are more likely not to respond. Imputation methods which ignore the x-variable (mean and hot deck) are expected to yield estimators of $\bar{Y}$ that underestimate the true mean under nonresponse model (22) and over estimate the true mean under the model (23). However, imputation methods that incorporate the auxiliary variable into the procedure (ratio and nearest neighbour), can be expected to produce better estimates of the mean. This has been confirmed by simulation as shown in Table 5 below. As before, 100,000 replicates were used.

Table 5
Estimates of the Mean $\hat{Y}$ as Percent of the True Mean when the Nonresponse is not Random, and the Nonresponse Rate is an Expected 30%

Nonresponse Model	Imputation Method			
	Mean	HD	Ratio	NN
	in percent			
P_L	60.4	60.4	94.7	93.5
P_S	132.7	132.7	102.0	101.4

Clearly, variance estimation is of no interest when the point estimators themselves are highly biased as is the case for the mean and hot deck methods. However, in the case of the ratio and nearest neighbour methods, under which the point estimators perform better, we investigated the performance of the adjusted jackknife variance estimator, as well as an estimator proposed by Särndal (1990), which can be written as (Rao 1992):

$$v_s(\bar{y}_I) = \left(\frac{\bar{x}}{\bar{x}_m}\right)^2 \frac{1}{m(m-1)} \sum_{i \in s_r} \left(y_i - \frac{\bar{y}_m}{\bar{x}_m} x_i\right)^2 + \left(\frac{\bar{y}_m}{\bar{x}_m}\right) \frac{2m}{n^2(m-1)} \sum_{i \in s_r} \left(y_i - \frac{\bar{y}_m}{\bar{x}_m}\right) x_i + \left(\frac{\bar{y}_m}{\bar{x}_m}\right)^2 \frac{1}{n(n-1)} \sum_{i \in s} (x_i - \bar{x})^2, \tag{24}$$

provided that the finite population correction factor is ignored, and that $(n-1)/n \cong 1$ and $(m-1)/m \cong 1$. The results are summarized in Table 6.

Table 6

Relative Bias of the Naive Variance Estimator, the Adjusted Jackknife Variance Estimator and Särndal's Variance Estimator Under 30% Nonrandom Nonresponse

Nonresponse Model	Variance Estimator	Imputation Method	
		Ratio	NN
		in percent	
P_L	v_{naive}	−22.7	−54.6
	v_J	3.9	−37.5
	v_S	−2.6	−36.8
P_S	v_{naive}	−4.0	−0.7
	v_J	3.7	7.2
	v_S	2.8	4.5

In the case of the ratio imputation, the naive variance estimator performs quite differently under the two nonresponse models (−22.7 versus −4.0%). This is due to the fact that while the reduction in effective sample size tends to decrease the variance in both cases, under the P_L model disproportionately more large units are missing which tends to accentuate this effect, whereas under the P_S model, where disproportionately more small units are missing, this effect tends to be partly compensated for. Secondly, the adjusted jackknife variance estimator performs well in the case of ratio imputation, but relatively poorly in the case of nearest neighbour imputation. This is due to the fact that the present data set follows the usual linear model (15) fairly well and the adjusted jackknife variance estimator has been shown to be model unbiased (Rao 1992) in the case of the ratio imputation. On the other hand, the ratio adjustment does not work well in the case of nearest neighbour imputation when the nonresponse is not uniform. The alternate adjustment for the nearest neighbour imputation described in Section 3, performs equally poorly in absolute terms (not shown here), though the estimates are always conservative. Thirdly, the performance of Särndal's estimator, v_S, is roughly equivalent to that of the adjusted jackknife estimator under either the ratio or the nearest neighbour imputation methods, and nonrandom nonresponse that depends only on x.

In cases where the response mechanism is not random, and when the propensity to respond is related to the variable subject to nonresponse (y), the point estimators are themselves severely biased under all four imputation methods. As such, variance estimation is of little interest, as the real interest lies in estimating the mean squared error. That is, more attention needs to be concentrated on improving the point estimates and their bias. Some preliminary results on this front have been put forth by Rancourt, Lee and Särndal (1992).

5. CONCLUDING REMARKS

It is well known that the usual variance estimator understates the variance of the estimate of $\bar{Y}$ in the presence of imputed values if these values are treated as having been observed. In this study we again demonstrated the high degree of underestimation of the naive variance estimator in the presence of imputed data. Several imputation methods were considered in order to illuminate the dependence of the degree of underestimation on the method of imputation. We evaluated a unified jackknife variance estimator proposed by Rao and Shao (1992), an estimator that incorporates the variance due to imputation component. The study demonstrated some desirable properties of the proposed estimator in the case of both simple random sampling as well as complex survey designs. Our findings can be summarized as follows.

(1) The extent of variance underestimation is highly dependent on both the imputation method's ability to predict the true values, and its ability to preserve the natural variation in the data.

(2) The proposed adjusted jackknife variance estimator offers a unified approach to variance estimation of imputed data, that is easy to implement under a number of imputation methods and under designs of varying complexity.

(3) Operationally, no modifications to the original imputed file are necessary and the estimation of means and totals is thus unaffected by the need to estimate variances.

(4) The proposed method is easily extended to more complex designs, more than one imputation class and, with care, to the case of nonrandom nonresponse that depends only on available auxiliary variables.

(5) The adjusted jackknife variance estimator performs well whenever the nonresponse is uniform or the usual linear model holds, demonstrating the fact that the estimator is both design consistent as well as design-model unbiased.

(6) In the case of the P_L model, under which units with large y-values are more likely to not respond, all three variance estimators perform extremely poorly.

(7) In the case of y-dependent nonresponse, better imputation techniques are needed and the bias of the point estimators needs to be studied further. Here the issue is primarily that of estimating the mean square error rather than the variance.

Given the relatively high degree of imputation in today's surveys, at least within some imputation classes, it is clear that the effect of imputation on variance estimation cannot be ignored. An overestimation of precision can lead to confidence intervals that are too short and to spurious declaration of significance. If implementation of the above suggested methods is deemed too onerous in any particular circumstance, at the very least studies should be conducted to evaluate the impact of imputation in some representative cases. An *ad hoc* variance inflation factor could then be implemented. With the emergence of generalized estimation software, however, there seems to remain little reason for not implementing variance estimators which correctly account for the effect of imputation.

There clearly remain many unsolved, and perhaps unsolvable problems. To begin with, much more theoretical work is needed with respect to nearest neighbour imputation. The jackknife adjustments considered for this imputation method fail to perform as well as those applied to the other methods. Perhaps smoother alternatives to the nearest neighbour method need to be developed. Secondly, the robustness of the proposed estimator must be investigated. It is clear that satisfactory performance can be obtained if the model (15) holds, and when nonresponse is random. Limited failure of either one of these conditions did not seem to detract from the good performance of the jackknife estimator in our limited experience, but further research along these lines is warranted. Departures from both of the conditions simultaneously are yet to be investigated. Cases of nonrandom nonresponse when the propensity of nonresponse is related to the y-variable are even less well understood, though the emphasis in this case must be placed on the estimation of the mean square error rather than the variance. Thirdly, comparisons to multiple imputation results should be considered. It must be recognized, however, that proper imputation methods (Rubin 1987) must first be established. We note that none of the imputation methods studied within are proper with respect to multiple imputation.

Extensions to other imputation methods and other parameters of interest should be undertaken. This study was limited to four simple imputation methods. In practice, much more complicated methods are used, often in conjunction with each other. The impact of more than one imputation method on the estimation of variance has been studied by Rancourt, Lee and Särndal (1993); more work is needed. With respect to other, more complicated methods of imputation, the effect of adding theoretical residuals to imputed data can, for example, be considered. However, this technique only addresses the underestimation of V_{sam} by v_{naive} and ignores the effect of V_{imp}. Finally, other parameters, such as the median for example, and the effect of imputation on their variance are yet to be evaluated. Multivariate extensions can likewise be considered: estimation of correlations, ratios and regression parameters in the presence of imputation would likely be of interest.

ACKNOWLEDGEMENTS

The authors are grateful to Prof. J.N.K. Rao for his continuous encouragement and support, and the Associate Editor for his constructive comments.

REFERENCES

HANSEN, M., HURWITZ, W., and MADOW, W. (1953). *Sample Survey Methods and Theory.* (Volume 2), New York: J. Wiley, 139-141.

KALTON, G., and KASPRZYK, D. (1986). The treatment of missing survey data. *Survey Methodology*, 12, 1-16.

LEE, H., RANCOURT, E., and SÄRNDAL, C.-E. (1991). Experiments with variance estimation from survey data with imputed values. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 690-695.

RAO, J.N.K. (1990). Variance estimation under imputation for missing data. Unpublished report, Statistics Canada.

RAO, J.N.K. (1992). Jackknife variance estimation under imputation for missing survey data. Unpublished report, Statistics Canada.

RAO, J.N.K., and SHAO, J. (1992). Jackknife variance estimation with survey data under hot deck imputation. *Biometrika*, 79, 811-822.

RANCOURT, E., LEE, H., and SÄRNDAL, C.-E. (1992). Bias corrections for survey estimates from data with imputed values for nonignorable nonresponse. *Proceedings 1992 Annual Research Conference*, Bureau of the Census, 523-539.

RANCOURT, E., LEE, H., and SÄRNDAL, C.-E. (1993). Variance estimation under more than one imputation method. *Proceedings of the International Conference on Establishment Surveys, American Statistical Association*, 374-379.

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

SÄRNDAL, C.-E. (1990). Methods for estimating the precision of survey estimates when imputation has been used. Special invited lecture. *Proceedings: Symposium 90, Measurement and Improvement of Data Quality*, Statistics Canada, 337-350.

Survey Methodology, June 1994
Vol. 20, No. 1, pp. 53–57
Statistics Canada

Estimation in Overlapping Clusters with Unknown Population Size

D.S. TRACY and S.S. OSAHAN[1]

ABSTRACT

Two sampling strategies for estimation of population mean in overlapping clusters with known population size have been proposed by Singh (1988). In this paper, ratio estimators under these two strategies are studied assuming the actual population size to be unknown, which is the more realistic situation in sample surveys. The sampling efficiencies of the two strategies are compared and a numerical illustration is provided.

KEY WORDS: Overlapping clusters; Clustering before sampling; Mean square error; Relative efficiency.

1. INTRODUCTION

In cluster sampling, clusters are formed either before selecting the sample (CBS) or after selecting the sample (CAS). In both cases, clusters may be overlapping or non-overlapping. For non-overlapping clusters, much work by several researchers is available in the literature. However, there are many practical sampling situations where one gets overlapping clusters. For example, overlapping clusters may exist in some regional epidemiological survey for a contagious disease like mycobacterim tuberculosis (T.B.), becoming very prevalent with the spread of AIDS (Gifford-Jones 1993). Clusters here may be formed around infected individuals or closely associated individuals who are more vulnerable to the same type of infection. A similar situation may exist in an ecological survey where clusters are formed around the factories burning coal and emitting polyaromatic hydrocarbons (PAH's) which are potent cancer causing compounds. Clusters are formed on the basis of the intensity of such gases, and surveys may be required in order to control air pollution which causes lung diseases like bronchitis. For overlapping clusters, one can refer to the limited work done by Goel and Singh (1977), Agarwal and Singh (1982) and Amdekar (1985). But the methodologies developed by them suffer from one limitation or the other.

Recently, Singh (1988) has developed a very simple estimator for a population mean using two sampling strategies in the CBS system assuming known population size. In the first strategy, clusters are selected with equal probabilities, whereas in the second case selection probabilities are taken proportional to cluster size. The elements within the clusters are selected with equal probability in both the cases. But it is unrealistic to assume that the actual population size is known. If it is the case, then all the duplicates in the population are known *a priori*, and one could easily remove them to increase the efficiency of the sampling design. Hence, the estimators of the population mean studied by Singh (1988) need an improvement in order to be practicable, as they depend on the actual population size. This limitation in the methodology has motivated the present work.

We propose two sampling strategies in the CBS system with simple ratio estimators for the population mean, which do not depend on the actual population size. As in Singh (1988), an equal probability with replacement sampling scheme is used for selecting the clusters in the first strategy, whereas in the second, an unequal probability sampling scheme is used. The elements within the clusters are selected with an equal probability without replacement sampling scheme in both strategies.

The population of N units under consideration is expressible in the form of K overlapping clusters with N_i units in the i-th cluster and $\sum_{i=1}^{K} N_i = M \geq N$, the unknown actual population size, (equality holds only for non-overlapping clusters). A population unit may be included in more than one cluster. Let y be the characteristic of interest and let the population mean be $\bar{Y}$.

Define

$$Z_{ij} = Y_{ij}/F_{ij}, \quad W_{ij} = 1/F_{ij}; \quad i = 1, 2, \ldots, K, \quad \text{and} \quad j = 1, 2, \ldots, N_i$$

where Y_{ij} is the value of y for the j-th unit in the i-th cluster and F_{ij} its frequency of occurring in K clusters.

When clusterwise data on units are available on the computer, the values of these frequencies for overlapping clusters may be easily available. As for the example considered earlier in epidemiology, suppose we have data available for households or individuals along with their identification labels like house numbers or social insurance

[1] D.S. Tracy and S.S. Osahan, Department of Mathematics and Statistics, University of Windsor, Windsor, Ontario N9B 3P4.

numbers/health card numbers on the computer. Then, by giving a simple command to the computer, a researcher can easily extract information about the repetition of a certain unit from its label in different clusters. Also, in case we have a map of the overlapping clusters and the criterion for forming clusters does not allow the elimination of duplicacy of units in the different clusters, the values of such frequencies may be known.

The two strategies are discussed in section 2 and their efficiencies are compared in section 3.

2. THE TWO STRATEGIES

The two proposed strategies are discussed in Sections 2.1 and 2.2. Their comparison is undertaken in Section 3.

2.1 Strategy A

This strategy consists of the following steps:

(a) Select k clusters out of K by simple random sampling with replacement (SRSWR).

(b) From the i-th selected cluster of size $N_i (i = 1, \ldots, K)$, select n_i elementary units by simple random sampling without replacement (SRSWOR).

Theorem 1. The ratio estimator under SRS

$$\bar{z}_{RS} = \hat{Y}_{RS}/\hat{N}_{RS} = \frac{K}{k}\sum_{i=1}^{k} N_i \bar{z}_i \Big/ \frac{K}{k}\sum_{i=1}^{k} N_i \bar{w}_i \tag{1}$$

has relative bias, to the first order of approximation,

$$RB(\bar{z}_{RS}) \doteq \frac{K}{k}\left[\left(\frac{\sigma_{bw}^2}{N^2} - \frac{\sigma_{bzw}}{NY}\right)K + \sum_{i=1}^{K} N_i^2\left(\frac{1}{n_i} - \frac{1}{N_i}\right)\left(\frac{S_{iw}^2}{N^2} - \frac{S_{izw}}{NY}\right)\right] \tag{2}$$

where

$$\sigma_{bzw} = \sum_{i=1}^{K} (N_i\bar{Z}_i - Y/K)(N_i\bar{W}_i - N/K)/K$$

$$S_{izw} = \sum_{j=1}^{N_i} (Z_{ij} - \bar{Z}_i)(W_{ij} - \bar{W}_i)/(N_i - 1),$$

$$\bar{Z}_i = \sum_{j=1}^{N_i} Z_{ij}/N_i \quad \text{and} \quad \bar{z}_i = \sum_{j=1}^{n_i} z_{ij}/n_i,$$

and σ_{bw}^2, S_{iw}^2, $\bar{W}_i$ and $\bar{w}_i$ are the expressions of σ_{bzw}, S_{izw}, $\bar{Z}_i$ and $\bar{z}_i$ respectively, with z replaced by w and Y replaced by N.

Proof. Following a standard result, the relative bias of the estimator $\bar{z}_{RS}$, to the first order of approximation, is

$$RB(\bar{z}_{RS}) \doteq [V(\hat{N}_{RS})/N^2] - \mathrm{Cov}(\hat{Y}_{RS},\hat{N}_{RS})/YN. \tag{3}$$

Let E_2 and V_2 denote the conditional expectation and variance for a given sample of clusters and E_1 and V_1 the expectation and variance over all such samples. Then, we have

$$\begin{aligned} V(\hat{N}_{RS}) &= V_1E_2(\hat{N}_{RS}) + E_1V_2(\hat{N}_{RS}) \\ &= V_1\left[\frac{K}{k}\sum_{i=1}^{k} N_iE_2(\bar{w}_i)\right] + E_1\left[\frac{K^2}{k^2}\sum_{i=1}^{k} N_i^2V_2(\bar{w}_i)\right] \\ &= V_1\left(\frac{K}{k}\sum_{i=1}^{k} N_i\bar{W}_i\right) + E_1\left[\frac{K^2}{k^2}\sum_{i=1}^{k} N_i^2\left(\frac{1}{n_i} - \frac{1}{N_i}\right)S_{iw}^2\right] \\ &= \frac{K^2}{k}\sigma_{bw}^2 + \frac{K}{k}\sum_{i=1}^{K} N_i^2\left(\frac{1}{n_i} - \frac{1}{N_i}\right)S_{iw}^2. \end{aligned} \tag{4}$$

Similarly, we have

$$\mathrm{Cov}(\hat{Y}_{RS},\hat{N}_{RS}) = \frac{K^2}{k}\sigma_{bzw} + \frac{K}{k}\sum_{i=1}^{K} N_i^2\left(\frac{1}{n_i} - \frac{1}{N_i}\right)S_{izw}. \tag{5}$$

By substituting (4) and (5) in (3), we obtain (2), which completes the proof of the theorem.

Theorem 2. The mean square error (MSE) of the estimator $\bar{z}_{RS}$, to the first order of approximation, is

$$\mathrm{MSE}(\bar{z}_{RS}) \doteq \frac{K}{kN^2}\sum_{i=1}^{K} N_i^2\left[(\bar{Z}_i - \bar{Y}\bar{W}_i)^2 + \left(\frac{1}{n_i} - \frac{1}{N_i}\right)D_i^2\right] \tag{6}$$

where $D_i^2 = S_{iz}^2 - 2\bar{Y}S_{izw} + \bar{Y}^2S_{iw}^2$, and $S_{iz}^2 = \sum_{j=i}^{N_i} (Z_{ij} - \bar{Z}_i)^2/(N_i - 1)$.

Proof. To the first order of approximation, we have

$$\text{MSE}(\bar{z}_{RS}) \doteq [V(\hat{Y}_{RS}) - 2\bar{Y}\,\text{Cov}(\hat{Y}_{RS},\hat{N}_{RS}) + \bar{Y}^2 V(\hat{N}_{RS})]/N^2. \quad (7)$$

The expression for $V(\hat{Y}_{RS})$ may be written, following (4), as

$$V(\hat{Y}_{RS}) = \frac{K^2}{k}\sigma_{bz}^2 + \frac{K}{k}\sum_{i=1}^{K} N_i^2\left(\frac{1}{n_i} - \frac{1}{N_i}\right)S_{iz}^2 \quad (8)$$

where $\sigma_{bz}^2 = \sum_{i=1}^{K} (N_i\bar{Z}_i - Y/K)^2/K$.

By substituting (4), (5) and (8) in (7), we obtain upon simplification

$$\text{MSE}(\bar{z}_{RS}) \doteq \frac{K^2}{kN^2}(\sigma_{bz}^2 - 2\bar{Y}\sigma_{bzw} + \bar{Y}^2\sigma_{bw}^2)$$

$$+ \frac{K}{kN^2}\sum_{i=1}^{K} N_i^2\left(\frac{1}{n_i} - \frac{1}{N_i}\right)(S_{iz}^2 - 2\bar{Y}S_{izw} + \bar{Y}^2 S_{iw}^2). \quad (9)$$

Substitution of the expressions for σ_{bz}^2, σ_{bzw} and σ_{bw}^2 into (9) and simplification yields (6). Now, we provide an estimator of MSE$(\bar{z}_{RS})$ below.

Theorem 3. A consistent estimator of MSE$(\bar{z}_{RS})$, to the first order of approximation, is given by

$$\widehat{\text{MSE}}(\bar{z}_{RS}) = \frac{K^2}{k\hat{N}_{RS}^2}\cdot\frac{1}{k-1}\sum_{i=1}^{k} N_i^2(\bar{z}_i - \bar{z}_{RS}\bar{w}_i)^2. \quad (10)$$

Proof. We note that the first-stage sampling is done with SRSWR sampling scheme and the random variables $N_i\bar{z}_i$ and $N_i w_i$ in the ratio estimator are independently and identically distributed. Hence, the mean square error of $\bar{z}_{RS}$ can be estimated using the well-known result that a variance estimator for a multi-stage design can consider the first stage only (see Särndal, Swensson and Wretman, 1992, Results 2.9.1 and 4.5.1).

From (9), an unbiased estimator of

$$\sigma_{bz}^2 + \frac{1}{K}\sum_{i=1}^{K} N_i^2\left(\frac{1}{n_i} - \frac{1}{N_i}\right)S_{iz}^2$$

can be written as

$$s_{bz}^2 = \frac{1}{k-1}\sum_{i=1}^{k}\left(N_i\bar{z}_i - \sum_{i=1}^{k} N_i\bar{z}_i/k\right)^2, \quad (11)$$

and an unbiased estimator of

$$\sigma_{bzw} + \frac{1}{K}\sum_{i=1}^{K} N_i^2\left(\frac{1}{n_i} - \frac{1}{N_i}\right)S_{izw}$$

is

$$s_{bzw} = \frac{1}{k-1}\sum_{i=1}^{k}\left(N_i\bar{z}_i - \sum_{i=1}^{k} N_i\bar{z}_i/k\right) \times \left(N_i\bar{w}_i - \sum_{i=1}^{k} N_i\bar{w}_i/k\right). \quad (12)$$

Similarly, an independent estimator of

$$\sigma_{bw}^2 + \frac{1}{K}\sum_{i=1}^{K} N_i^2\left(\frac{1}{n_i} - \frac{1}{N_i}\right)S_{iw}^2$$

is s_{bw}^2, defined parallel to (11).

Using these results, one can easily show that a consistent estimator of MSE$(\bar{z}_{RS})$ given in (6) is provided by

$$\widehat{\text{MSE}}(\bar{z}_{RS}) = \frac{K^2}{k\hat{N}_{RS}^2}(s_{bz}^2 - 2\bar{z}_{RS}s_{bzw} + \bar{z}_{RS}^2 s_{bw}^2),$$

which can be written as (10).

2.2 Strategy B

This strategy consists of the following steps:

(a) Select k clusters out of K by probability proportional to size with replacement (PPSWR) sampling with selection probabilities $P_i = N_i/M$, $i = 1, \ldots, K$.

(b) Same as for strategy A.

Theorem 4. The ratio estimator under PPS sampling

$$\bar{z}_{RP} = \hat{Y}_{RP}/\hat{N}_{RP} = \frac{M}{k}\sum_{i=1}^{k}\bar{z}_i \Big/ \frac{M}{k}\sum_{i=1}^{k}\bar{w}_i \quad (13)$$

has relative bias, to the first order of approximation,

$$RB(\bar{z}_{RP}) \doteq \frac{M^2}{k}\left[\left(\frac{\sigma_{bw'}^2}{N^2} - \frac{\sigma_{bzw'}}{YN}\right) + \sum_{i=1}^{K}\frac{N_i}{M}\left(\frac{1}{n_i} - \frac{1}{N_i}\right)\left(\frac{S_{iw}^2}{N^2} - \frac{S_{izw}}{YN}\right)\right] \quad (14)$$

where

$$\sigma_{bzw'} = \sum_{i=1}^{K}(\bar{Z}_i - Y/M)(\bar{W}_i - N/M)(N_i/M)$$

and $\sigma_{bw'}^2$ is the expression of $\sigma_{bzw'}$ with z replaced by w and Y replaced by N.

Proof. Using a standard result, the approximate relative bias, to the first order of approximation, is

$$RB(\bar{z}_{RP}) \doteq [V(\hat{N}_{RP})/N^2] - \mathrm{Cov}(\hat{Y}_{RP},\hat{N}_{RP})/YN. \quad (15)$$

We have

$$V(\hat{N}_{RP}) = V_1E_2(\hat{N}_{RP}) + E_1V_2(\hat{N}_{RP})$$

$$= M^2\left[V_1\frac{1}{k}\sum_{i=1}^{k} E_2(\bar{w}_i) + E_1\frac{1}{k^2}\sum_{i=1}^{k} V_2(\bar{w}_i)\right]$$

$$= M^2\left[V_1\frac{1}{k}\sum_{i=1}^{k} \bar{W}_i + E_1\frac{1}{k^2}\sum_{i=1}^{k}\left(\frac{1}{n_i}-\frac{1}{N_i}\right)S_{iw}^2\right]$$

$$= \frac{M^2}{k}\left[\sigma_{bw'}^2 + \sum_{i=1}^{K}\frac{N_i}{M}\left(\frac{1}{n_i}-\frac{1}{N_i}\right)S_{iw}^2\right]. \quad (16)$$

Similarly, one can write

$$\mathrm{Cov}(\hat{Y}_{RP},\hat{N}_{RP}) = \frac{M^2}{k}\left[\sigma_{bzw'} + \sum_{i=1}^{K}\frac{N_i}{M}\left(\frac{1}{n_i}-\frac{1}{N_i}\right)S_{izw}\right]. \quad (17)$$

Substituting (16) and (17) in (15), we obtain (14).

Theorem 5. The MSE of the estimator $\bar{z}_{RP}$, to the first order of approximation, is

$$\mathrm{MSE}(\bar{z}_{RP}) \doteq \frac{M}{kN^2}\sum_{i=1}^{K} N_i \times \left[(\bar{Z}_i - \bar{Y}\bar{W}_i)^2 + \left(\frac{1}{n_i}-\frac{1}{N_i}\right)D_i^2\right]. \quad (18)$$

Proof. We write, to the first order of approximation,

$$\mathrm{MSE}(\bar{z}_{RP}) \doteq [V(\hat{Y}_{RP}) - 2\bar{Y}\,\mathrm{Cov}(\hat{Y}_{RP},\hat{N}_{RP}) + \bar{Y}^2V(\hat{N}_{RP})]/N^2. \quad (19)$$

Also, from Theorem 2.5 of Singh (1988), we have by analogy

$$V(\hat{Y}_{RP}) = \frac{M^2}{k}\sigma_{bz''}^2 + \frac{M^2}{k}\sum_{i=1}^{K}\frac{N_i}{M}\left(\frac{1}{n_i}-\frac{1}{N_i}\right)S_{iz}^2, \quad (20)$$

where $\sigma_{bz'}^2 = \sum_{i=1}^{K}(N_i/M)\,(\bar{Z}_i - Y/M)^2$. On substituting (16), (17) and (20) in (19) and simplifying, we obtain (18).

Theorem 6. A consistent estimator of $\mathrm{MSE}(\bar{z}_{RP})$, to the first order of approximation, is

$$\widehat{\mathrm{MSE}}(\bar{z}_{RP}) = \frac{M^2}{\hat{N}_{RP}^2}\cdot\frac{1}{k(k-1)}\sum_{i=1}^{k}(\bar{z}_i - \bar{z}_{RP}\bar{w}_i)^2. \quad (21)$$

Proof. As the first-stage units are selected with PPSWR, the justification given in the proof of theorem 3 applies here, as well.

From (20), using Results 2.9.1 and 4.5.1 of Särndal, Swensson and Wretman (1992), an unbiased estimator of

$$\sigma_{bz'}^2 + \sum_{i=1}^{K}\frac{N_i}{M}\left(\frac{1}{n_i}-\frac{1}{N_i}\right)S_{iz}^2$$

can be written as

$$s_{bz'}^2 = \frac{1}{k-1}\sum_{i=1}^{k}\left(\bar{z}_i - \sum_{i=1}^{k}\bar{z}_i/k\right)^2. \quad (22)$$

Similarly, defining $s_{bzw'}$ and $s_{bw'}^2$, one can show that

$$\widehat{\mathrm{MSE}}(\bar{z}_{RP}) = \frac{M^2}{\hat{N}_{RP}^2k}(s_{bz'}^2 - 2\bar{z}_{RP}s_{bzw'} + \bar{z}_{RP}^2s_{bw'}^2),$$

which can be written as (21).

3. EFFICIENCY COMPARISON

The efficiencies of the estimators are compared below under the two strategies.

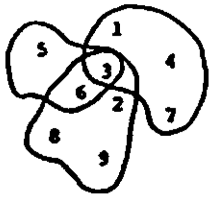
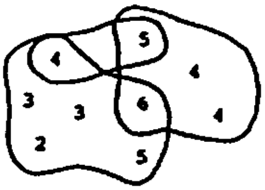
Remark. The estimator $\bar{z}_{RP}$ under strategy B is expected to be more efficient than the estimator $\bar{z}_{RS}$ under strategy A.

We provide a justification. From (6) and (18), we obtain

$$\mathrm{MSE}(\bar{z}_{RS}) - \mathrm{MSE}(\bar{z}_{RP}) \doteq \frac{M}{kN^2}\sum_{i=1}^{K}N_i \times \left[(\bar{Z}_i - \bar{Y}\bar{W}_i)^2 + \left(\frac{1}{n_i}-\frac{1}{N_i}\right)D_i^2\right]\left(\frac{KN_i}{M}-1\right).$$

As the cluster size N_i increases, the factor $(KN_i/M - 1)$ will also increase. The other factor of the term under summation is $N_i[(\bar{Z}_i - \bar{Y}\bar{W}_i)^2 + (1/n_i - 1/N_i)D_i^2]$, which represents the contribution due to variability in z and w present in the i-th cluster (without the constant M/kN^2) towards $\mathrm{MSE}(\bar{z}_{RP})$ in (18). As cluster size N_i increases, the contribution of the i-th cluster towards $\mathrm{MSE}(\bar{z}_{RP})$ is also expected to increase. This makes the covariance between these two factors positive. Hence, the estimator $\bar{z}_{RP}$ is expected to have a smaller MSE than $\bar{z}_{RS}$.

Table 1
Comparison of the Two Strategies for Two Small Populations

	Population No. 1			Population No. 2		
						
N_i	3	4	5	2	4	6
n_i	1	2	2	1	2	2
Y_{ij}	3,5,6	1,3,4,7	2,3,6,8,9	4,5	4,4,5,6	2,3,3,4,5,6
F_{ij}	3,1,2	1,3,1,1	1,3,2,1,1	2,2	1,1,2,2	1,1,1,2,1,2
Z_{ij}	1,5,3	1,1,4,7	2,1,3,8,9	2,2.5	4,4,2.5,3	2,3,3,2,5,3
W_{ij}	⅓,1,½	1,⅓,1,1	1,⅓,½,1,1	½,½	1,1,½,½	1,1,1,½,1,½
F	1.38	10.16	18.12	.24	.77	2.94
MSE$(\bar{z}_{RS})$	2.09			0.45		
MSE$(\bar{z}_{RP})$	1.83			0.33		
R.E.	114.21			136.36		
R.B.$(\bar{z}_{RS})$	$-.0105$			.0348		
R.B.$(\bar{z}_{RP})$	$-.0047$			$-.0037$		

Numerical Illustration. Here the two proposed sampling strategies are applied to two small populations to shed light on the computations of F_{ij}, Z_{ij} and W_{ij}, and on their comparison. For both the populations $K = 3$, $k = 2$, $M = 12$ and $N = 9$. A unit repeated in two or more clusters represents overlapping. The populations are described in Table 1.

The analysis of the results in Table 1 supports the theoretical developments of the present paper. For both the populations, the factor $F = N_i\ [(\bar{Z}_i - \bar{Y}\bar{W}_i)^2 + (1/n_i - 1/N_i)D_i^2]$ increases with N_i, resulting in MSE$(\bar{y}_{RP}) <$ MSE$(\bar{y}_{RS})$, as remarked above.

CONCLUSION

This paper removes the realistic limitation of known population size in the earlier work of Singh (1988) while considering overlapping clusters. Also comparison of the two strategies here is more direct, whereas in Singh (1988) the support of evidence given by Hansen and Hurwitz (1943) was needed.

ACKNOWLEGDEMENTS

This research was partially supported by NSERC Grant A-3111. The comments of the referees and the editor on an earlier version were most helpful in improving the paper. These are gratefully acknowledged.

REFERENCES

AGARWAL, D.K., and SINGH, P. (1982). On cluster sampling strategies using ancillary information. *Sankhyā*, B, 44, 184-192.

AMDEKAR, S.J.(1985). An unbiased estimator in overlapping clusters. *Bulletin of the Calcutta Statistical Association*, 15, 231-232.

GIFFARD-JONES, W. (1993). The doctor game. *The Windsor Star*, April 15, 1993.

GOEL, B.B.P.S., and SINGH, D. (1977). On the formation of clusters. *Journal of the Indian Society of Agricultural Statistics*, 29, 53-68.

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

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

SINGH, S. (1988). Estimation in overlapping clusters. *Communications in Statistics, Theory and Methods*, 17, 613-621.

Survey Methodology, June 1994
Vol. 20, No. 1, pp. 59–64
Statistics Canada

PPS Sampling over Two Occasions

N.G.N. PRASAD and J.E. GRAHAM[1]

ABSTRACT

The Random Group Method for sampling with probability proportional to size(PPS) is extended to sampling over two occasions. Information on a study variate observed on the first occasion is used to select the matched portion of the sample on the second occasion. Two real data sets are considered for numerical illustration and for comparsion with other existing methods.

KEY WORDS: Composite estimator; Efficiency comparisons; Random group method; Probability proportional to size.

1. INTRODUCTION

The practice of using a partial replacement sampling scheme in repeated surveys is quite common due, in part, to an anticipated increase in the efficiency of estimation as well as a reduction in the burden of response. Essentially, after each sampling occasion a fraction of the units observed on that occasion is rotated out of the sample and replaced by a fresh sub-sample from the population.This set of unmatched units is then observed on the next sampling occasion along with the remaining set of matched units. The literature abounds with discussions of sampling and estimation procedures for sampling with equal selection probabilities on two occasions. A particularly important case is the situation where the units are chosen on a given occasion with unequal selection probabilities. In the literature to date,information collected on the previous occasion is used to improve upon the customary estimator of the total or mean for the current occasion by using a difference method of estimation. In this article we present a sampling and estimation procedure for sampling on two occasions which incorporates information collected on the first (previous) occasion in selecting the sub-sample for observation on the second (current) occasion. For the sake of completeness and parsimony, we review only unequal probability selection procedures for two occasions in this section.

Consider a finite population of N units, labelled $1, 2, \ldots, N$, and two sampling occasions: 1 (previous occasion) and 2 (current occasion). Let y_{1i} and y_{2i} denote the values of a characteristic y for the i-th unit observed on the first and second occasions respectively. Let Y_1 and Y_2 denote the respective population totals. Suppose a size measure x is known for each of the population units.

1.1 The Des Raj Scheme

Raj (1965) considered the following PPS (probability proportional to size) sampling scheme: On the first occasion a sample s of size n is selected with probabilities p_i proportional to the x_i values, $i = 1, 2, \ldots, N$, and with replacement (wr). On the second occasion a simple random sample s_1 of m units is selected from s without replacement (wor) and an independent PPS sample s_2 of $u = n - m$ units is selected wr from the entire population. Then Y_1 and Y_2 are respectively unbiasedly estimated by:

$$\hat{Y}_1 = \sum_{i \in s} y_{1i}/(np_i) \tag{1.1}$$

and

$$\hat{Y}_2 = Q\hat{Y}_{2u} + (1 - Q)\hat{Y}_{2m}, \tag{1.2}$$

where

$$\hat{Y}_{2u} = \sum_{i \in s_2} y_{2i}/(up_i), \tag{1.3}$$

$$\hat{Y}_{2m} = \sum_{i \in s} y_{1i}/(np_i) + \sum_{i \in s_1} (y_{2i} - y_{1i})/(mp_i), \tag{1.4}$$

and Q is a weight, $0 \le Q \le 1$. Assuming that

$$V_1 = \sum_{i=1}^{N} (y_{1i}/p_i - Y_1)^2 p_i = V_2$$

$$= \sum_{i=1}^{N} (y_{2i}/p_i - Y_2)^2 p_i = V, \tag{1.5}$$

[1] N.G.N. Prasad, Associate Professor, Department of Statistics and Applied Probability, University of Alberta, Edmonton, Alberta, Canada T6G 2G1; J.E. Graham, Professor, Department of Mathematics and Statistics, Carleton University, Ottawa, Ontario, Canada K1S 5B6.

the minimum variance of $\hat{Y}_2$ was found to be

$$V_{min}(\hat{Y}_2) = V[1 + \sqrt{2(1-\delta)}/(2n)], \qquad (1.6)$$

where δ is given by

$$V\delta = \sum_{i=1}^{N} (y_{1i}/p_i - Y_1)(y_{2i}/p_i - Y_2)p_i. \qquad (1.7)$$

1.2 The Ghangurde-Rao (G-R) Scheme

Under the PPSWOR framework, Ghangurde and Rao (1969) extended the Rao-Hartley-Cochran (RHC) Method, also known as the Random Group Method (See: Rao, Hartley and Cochran 1962) to sampling on two occasions. Under the RHC Method, the population of N units is split at random into n groups of sizes $N_1, N_2, \ldots, N_n$ such that $\sum_{h=1}^{n} N_h = N$, and a sample of one unit is drawn independently from each of the n groups with probabilities proportional to the initial selection probabilities, p_i. Under the G-R Method, the population is first divided at random into n groups, each of size N/n (assumed to be an integer). On the first occasion, one unit is drawn from each random group (as described above), giving a sample s of n units. On the second occasion, a simple random wor sample s_1 of $m = \lambda n (o < \lambda < 1)$ matched units is selected from s and an independent sample s_2 of $u = n - m$ units is drawn from the whole population of N units by the same method that was used in obtaining s. Then, a composite estimator of Y_2 is given by

$$\hat{Y}_2' = Q'\hat{Y}_{2u}' + (1 - Q')\hat{Y}_{2m}', \qquad (1.8)$$

where $0 \le Q' \le 1$,

$$\hat{Y}_{2u}' = \sum_{i \epsilon s_2} \frac{y_{2i}P_i^*}{p_i}, \qquad (1.9)$$

and

$$\hat{Y}_{2m}' = \sum_{i \epsilon s} \frac{y_{1i}P_i}{p_i} + nm^{-1} \sum_{i \epsilon s_1} \frac{(y_{2i} - y_{1i})P_i}{p_i}, \qquad (1.10)$$

with P_i and P_i^* denoting the totals of the p_i values for the groups containing the i-th unit $(i = 1, 2, \ldots, N)$ in the selection of s and s_2 respectively. Under assumption (1.5), the variance of $\hat{Y}_2'$ (with optimum values of Q' and λ) is given by

$$V_{min}(\hat{Y}_2') = \frac{NV}{2n(N-1)}$$

$$\times [1 - n/N + \sqrt{2(1-\delta)}(1+\gamma)n/N], \qquad (1.11)$$

where

$$\gamma = \frac{(1-\rho')V'}{(1-\delta)V} - 1,$$

$$V' = N^{-1} \sum_{i=1}^{N} (y_{1i} - \bar{Y}_1)^2 = N^{-1} \sum_{i=1}^{N} (y_{2i} - \bar{Y}_2)^2$$

and

$$\rho' = N^{-1} \sum_{i=1}^{N} (y_{1i} - \bar{Y}_1)(y_{2i} - \bar{Y}_2)/V'.$$

1.3 The Chotai Scheme

Chotai (1974), under the additional assumption that n/m is an integer, modified the G-R sampling design on the second occasion. A sample s is selected as in the G-R scheme on the first occasion. On the second occasion, the n units in the sample s are split at random into $m (= \lambda n)$ groups each of size n/m. One unit is selected from each of the m groups independently with probabilities proportional to the P_i's as defined in the G-R scheme. This selection yields the sample s_1. The selection of s_2 is the same as in the G-R scheme. Then a composite estimator of Y_2 is given by

$$\hat{Y}_2^C = Q^C\hat{Y}_{2u}^C + (1 - Q^C)\hat{Y}_{2m}^C, \qquad (1.12)$$

where $0 \le Q^C \le 1$,

$$\hat{Y}_{2u}^C = \sum_{i \epsilon s_2} \frac{y_{2i}P_i^*}{p_i}, \qquad (1.13)$$

and

$$\hat{Y}_{2m}^C = \sum_{i \epsilon s_1} \frac{(y_{2i} - y_{1i})P_i^+}{p_i} + \sum_{i \epsilon s} \frac{y_{1i}P_i}{p_i}. \qquad (1.14)$$

Here, P_i and P_i^* are as defined in the G-R scheme, and P_i^+ denotes the total of the P_i-values for those random groups of s containing the i-th unit $(i = 1, 2, \ldots, N)$ in the selection of s_1. The minimum variance of $\hat{Y}_2^C$ under assumption (1.5), obtained by using the optimum values of Q^C and λ, is given by

$$V_{min}(\hat{Y}_2^C) = \frac{NV}{2n(N-1)}[1 - n/N + \sqrt{2(1-\delta)}]. \qquad (1.15)$$

Under this scheme, but without assumption (1.5), Chotai also considered an estimator of Y_2 (similar to Kulldorff's estimator for simple random sampling: See Kulldorff 1963), given by

$$\hat{Y}_2^{CM} = Q^{CM}\hat{Y}_{2u}^{C} + (1 - Q^{CM})\hat{Y}_{2m}^{CM}, \qquad (1.16)$$

where $\hat{Y}_{2u}^{C}$ is as defined in (1.13), Q^{CM} $(0 \le Q^{CM} \le 1)$ is an assigned weight to be determined and

$$\hat{Y}_{2m}^{CM} = \sum_{i \in s_1} \frac{(y_{2i} - \beta y_{1i})P_i^{+}}{p_i} + \beta \sum_{i \in s} \frac{y_{1i}P_i}{p_i}, \qquad (1.17)$$

with

$$\beta = \delta \left[\frac{\sum_{i=1}^{N} p_i (y_{2i}/p_i - Y_2)^2}{\sum_{i=1}^{N} p_i (y_{1i}/p_i - Y_1)^2} \right] = \delta \frac{V_2}{V_1}, \qquad (1.18)$$

and δ as defined in (1.7). The minimum variance of $\hat{Y}_2^{CM}$, using optimum values of Q^{CM} and λ, is given by

$$V_{min}(\hat{Y}_2^{CM}) = \frac{N}{2n(N-1)}(1 + \sqrt{1 - \delta^2} - n/N)V_2. \qquad (1.19)$$

To actually use $\hat{Y}_2^{CM}$ it is evidently necessary to first assess the value of β, which is usually not possible in practice. An estimate of β based on the available sample can be used but this will induce a bias in the estimator $\hat{Y}_2^{CM}$.

2. ALTERNATIVE SCHEMES FOR SAMPLING PPS OVER TWO OCCASIONS

We now present an alternative sampling and estimation procedure which does not require a known value of β as defined in (1.18). In this scheme information collected on the first occasion is used in selecting the sample on the second occasion. The approach is based upon a procedure developed by Prasad and Srivenkataramana (1980) and was used there in the context of double sampling where a second phase sub-sample is selected using information obtained from an initial sample. For simplicity, we first consider its implementation in Raj's (1965) scheme (described earlier).

2.1 A Modification of Des Raj's Scheme

On the first occasion a sample s of size n is selected with probabilities p_i proportional to the x_i values and with replacement. On the second occasion, instead of choosing a SRSWOR sub-sample, a sub-sample s_1 of m units is selected from s using a PPSWR scheme with size measure $z_i = y_{1i}/x_i$, where y_{1i} is the observed value for the y characteristic for unit i on the first occasion. A sample s_2 of size $u = n - m$ is drawn, independent of s, as in Raj (1965). A composite estimator of Y_2 is given by

$$\tilde{Y}_2 = Q\,\hat{Y}_{2u} + (1 - Q)\tilde{Y}_{2m},$$

where $\hat{Y}_{2u}$ is as defined in (1.3) and

$$\tilde{Y}_{2m} = \frac{1}{nm}\sum_{i \in s_1} \frac{(y_{2i}/p_i)}{(y_{1i}/p_i)} \sum_{i \in s} (y_{1i}/p_i),$$

with Q being a weight, $0 \le Q \le 1$. The minimum variance of $\tilde{Y}_2$, obtained by minimizing the variance of $\tilde{Y}_2$ with respect to Q, is given by

$$V_{min}(\tilde{Y}_2) = V_1 C_1 (n + C_1 m)^{-1},$$

where $C_1 = \sum_{i=1}^{N} (y_{2i}/p_{1i} - Y_2)^2 p_{1i} V_1^{-1}$, with $p_{1i} = y_{1i}/Y_1$ and V_1 as defined in (1.5).

2.2 A Modification to Chotai's Scheme

As in Chotai (1974), assume that N, n, and $m(< n)$ are all positive integers such that N/n, N/u and n/m are also all integers. Then:

1. For the first occasion select a sample s of size n in the same manner as that adopted in the G-R procedure. For this set of units, observations y_{1i}, $i = 1, \ldots, n$, are made on a characteristic y.
2. For the second occasion, (a) split the n units in s at random into m groups, each of size n/m and draw one unit with PPS, $p_i^* = (y_{1i}P_i)/p_i$, independently from each of the m groups, yielding a sub-sample s_1, where P_i is as defined in the G-R scheme; (b) select s_2, a fresh sample of $u = n - m$ units from the entire population, and observe the second occasion y values, y_{2i}, for these u units in the same manner as in the G-R scheme.

Note that the difference between the proposed procedure and that of Chotai (1974) lies in the selection of s_1: in the former, information collected on the first occasion is used in selecting s_1 on the second occasion.

We now consider an estimator of the second occasion total Y_2 that exploits the proposed procedure. Let

$$y_{2i}^{*} = \frac{y_{2i}P_i}{p_i}.$$

A composite estimator of Y_2 is given by

$$\hat{Y}_2^{*} = Q^{**}\hat{Y}_{2u}^{C} + (1 - Q^{**})\hat{Y}_{2m}^{*}, \qquad (2.1)$$

where $\hat{Y}^C_{2u}$ is defined as in (1.13), $0 \le Q^{**} \le 1$ and

$$\hat{Y}^*_{2m} = \sum_{i \in s_1} \frac{y^*_{2i}\tilde{P}_i}{p^*_i}.$$

Here $\tilde{P}_i$ denotes the total of the p^*_i values associated with those units that belong to the random group from which the i-th unit was selected in s_1. Let E_1 and E_2 denote expectation and V_1 and V_2 denote variance over all s and for a given s, respectively. The unbiasedness of $\hat{Y}^*_{2m}$ and hence of $\hat{Y}^*_2$ for Y_2 follows by noting that the expected value of $\hat{Y}^*_{2m}$ is

$$E(\hat{Y}^*_{2m}) = E_1E_2(\hat{Y}^*_{2m}) = E_1\left(\sum_{i \in s} \frac{y_{2i}P_i}{p_i}\right) = Y_2. \quad (2.2)$$

To obtain the variance of $\hat{Y}^*_{2m}$, consider

$$V_2(\hat{Y}^*_{2m}) = \frac{n-m}{m(n-1)} \sum_{i \in s} \left(\frac{y^*_{2i}}{p^*_i} - \sum_{i \in s} y^*_{2i}\right)^2 p^*_i$$

$$= \frac{n-m}{m(n-1)}\left[\sum_{i \in s} \frac{(y^2_{2i}/y_{1i})}{p_i} P_i \sum_{i \in s} \frac{y_{1i}P_i}{p_i} - \left(\sum_{i \in s} \frac{y_{2i}P_i}{p_i}\right)^2\right],$$

which leads, after considerable algebraic simplication, to

$$E_1V_2(\hat{Y}^*_{2m}) = \frac{N(n-m)}{mn(N-1)}\sigma^2_3,$$

where

$$\sigma^2_3 = \sum_{i=1}^{N} \left(\frac{y_{2i}}{y_{1i}}Y_1 - Y_2\right)^2 \frac{y_{1i}}{Y_1}.$$

Noting that

$$V_1E_2(\hat{Y}^*_{2m}) = \frac{N-n}{n(N-1)}\sigma^2_2,$$

it follows that

$$V(\hat{Y}^*_{2m}) = \frac{N}{n(N-1)}\left[(1 - n/N) + \frac{1-\lambda}{\lambda}h\right]\sigma^2_2, \quad (2.3)$$

where

$$h = \frac{\sigma^2_3}{\sigma^2_2}, \; \sigma^2_2 = V_2 = \sum_{i=1}^{N} (y_{2i}/p_i - Y_2)^2 p_i \quad \text{and} \quad \lambda = \frac{m}{n}.$$

Because $\hat{Y}^C_{2u}$ and $\hat{Y}^*_{2m}$ are independent, the variance of $\hat{Y}^*_2$ is given by

$$V(\hat{Y}^*_2) = Q^{**2}V(\hat{Y}^C_{2u}) + (1 - Q^{**})^2 V(\hat{Y}^*_{2m}),$$

where

$$V(\hat{Y}^C_{2u}) = \frac{N-u}{u(N-1)}\sigma^2_2,$$

and $V(\hat{Y}^*_{2m})$ is given by (2.3).

The minimum variance of $V(\hat{Y}^*_2)$ is obtained by using optimum values of Q^{**} and λ, respectively given by

$$Q^{**} = \frac{(1 - n/N) + \frac{(1-\lambda)}{\lambda}h}{(1 - n/N) + \frac{(1-\lambda)}{\lambda}h + \frac{(1 - (1-\lambda)n/N)}{(1-\lambda)}},$$

and

$$\lambda = \frac{\sqrt{h}}{1 + \sqrt{h}}.$$

Hence, the minimum variance of $V(\hat{Y}^*_2)$ is given by

$$V_{min}(\hat{Y}^*_2) = \frac{N\sigma^2_2}{n(N-1)}[1 - n/N + \sqrt{h}]. \quad (2.4)$$

Note that the quantity h reflects the efficiency of the estimator using the p_i's as initial selection probabilities over the estimator with initial selection probabilities y_{1i}/Y_1. A "small" value of h leads to an increase in the efficiency of the proposed method over Chotai's.

3. NUMERICAL EFFICIENCY COMPARISONS

The composite estimators $\hat{Y}^C_2$ defined in (1.12), $\hat{Y}^{CM}_2$ defined in (1.16) and $\hat{Y}^*_2$ defined in (2.1) are now compared at their respective optimum Q and λ values. The efficiency of the scheme proposed in 2.2 relative to Chotai's (1974) procedure is examined through a comparison of the following two relative efficiencies:

$$\text{RE1} = \frac{V_{min}(\hat{Y}^C_2)}{V_{min}(\hat{Y}^*_2)} = \frac{(1 - n/N) + \sqrt{2(1-\delta)}}{(1 - n/N) + \sqrt{h}}$$

and

$$\text{RE2} = \frac{V_{min}(\hat{Y}^{CM}_2)}{V_{min}(\hat{Y}^*_2)} = \frac{(1 - n/N) + \sqrt{1 - \delta^2}}{(1 - n/N) + \sqrt{h}},$$

evaluated respectively obtained using (1.15) and (2.4), and (1.19) and (2.4). It follows that the proposed scheme is superior to that of Chotai using Kulldorff's estimator (which depends on the unknown constant β) for those populations having $h < (1 - \delta^2)$. In order to permit meaningful numerical comparisons, two data sets that have appeared elsewhere in the literature are used here.

Data Set A: This data set relates to the area under wheat in 1964 (y_2), in 1963 (y_1) and cultivated area in 1961 (x) for 34 villages in India (See Murthy 1967). The parameter values for this data set are $\delta = 0.6404$ and $h = 0.1868$.

Data Set B: This data set relates to the area under wheat in 1937 (y_2) and in 1936 (y_1) and cultivated area in 1930 (x) for a sample of 34 villages in India (see: Sukhatme, P.V. and Sukhatme, B.V. 1970). The corresponding parameter values for this data set are $\delta = 0.7635$ and $h = 0.3811$.

Using these values for δ and h the two relative efficiencies values RE1 and RE2 (expressed as percentages) were computed for selected values of n/N and are given in Tables 1 and 2.

Table 1

RE1% - Values for Data Sets A and B

n/N	Data Set A	Data Set B
0.05	130.09	124.30
0.10	131.22	125.21
0.15	132.43	126.19
0.20	133.75	127.25
0.25	135.18	128.41
0.30	136.73	129.66

Table 2

RE2% - Values for Data Sets A and B

n/N	Data Set A	Data Set B
0.05	104.49	101.82
0.10	104.64	101.88
0.15	104.80	101.94
0.20	104.97	102.01
0.25	105.15	102.08
0.30	105.34	102.16

An examination of Table 1 leads to the conclusion that the proposed scheme out performs that of Chotai (1974). The gain in the efficiency ranges from 30% to 37% for Data Set A and from 24% to 30% for Data Set B as the sampling fraction varies from 0.05 to 0.30. Note that the increase in efficiency is greater for Data Set A than for Data Set B because of the difference in the value of the parameters h (0.1868 *vs.* 0.3811) and of δ (0.6404 *vs.* 0.7635). Recall that h measures the efficiency of p_i as a size measure for unit i compared to the use of y_{1i} as a size measure in estimating the total Y_2 for the current occasion and δ is the correlation between y_{1i}/p_i and y_{2i}/p_i as defined in (1.7). When h is relatively small, greater gains in efficiency are realized with the proposed scheme than when h is not small. In both cases, however, the efficiency gains using the proposed procedure are worthwhile.

The efficiency gains using the proposed method compared to the use of Chotai's scheme with Kulldorff's estimator (as reported in Table 2) are minimal, varying from 4.5% to 5.3% for Data Set A and from 1.8% to 2.2% from Data Set B. But in order to use Kulldorff's estimator, the value of β must be available. In practice this is not the case. It follows that the proposed strategy performs well from the point of view of actual implementation and of efficiency gain.

There are situations where the auxiliary information needed to compute the initial selection probabilities is not available. A simple random sampling scheme may then be used in place of the RHC procedure in selecting the sample for the first occasion enumeration; the RHC procedure can then be adopted in selecting s_1 by using the SRS information on the study variable collected on the first occasion. The theory for such a procedure follows directly as a special case of that presented by taking $p_i = 1/N$, $i = 1, \ldots, N$. One would anticipate that substantial gains in efficiency would then result in this situation.

ACKNOWLEDGEMENTS

The authors thank Professor J.N.K. Rao for suggesting this problem, and the referee for constructive suggestions. This research was supported by grants from the Natural Sciences and Engineering Research Council of Canada.

REFERENCES

CHOTAI, J. (1974). A note on the Rao-Hartley-Cochran method for PPS sampling over two occasions. *Sankhyā*, Series C, 36, 173-180.

GHANGURDE, P., and RAO, J.N.K. (1969). Some results on sampling over two occasions. *Sankhyā*, Series A, 31, 463-472.

KULLDORFF, G. (1963). Some problems of optimum allocation for sampling on two occasions. *Review of the International Statistical Institute*, 31, 24-57.

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

PRASAD, N.G.N., and SRIVENKATARAMANA, T. (1980). Double sampling with PPS selection. *Vignana Bharathi*, 6, 52-58.

RAJ, D. (1965). On sampling over two occasions with probabilities proportional to size. *Annals of Mathematical Statistics*, 36, 327-330.

RAO, J.N.K., HARTLEY, H.O., and COCHRAN, W.G. (1962). On a simple procedure of unequal probability sampling without replacement. *Journal of the Royal Statistical Society*, 24, 482-491.

SUKHATME, P.V., and SUKHATME, B.V. (1970). *Sampling Theory of Surveys With Applications*. Ames, Iowa: Iowa State University Press.

Survey Methodology, June 1994
Vol. 20, No. 1, pp. 65–73
Statistics Canada

Multi-way Stratification by Linear Programming

R.R. SITTER and C.J. SKINNER[1]

ABSTRACT

Rao and Nigam (1990, 1992) showed how a class of controlled sampling designs can be implemented using linear programming. In this article their approach is applied to multi-way stratification. A comparison is made with existing methods both by illustrating the sampling schemes generated for specific examples and by evaluating mean squared errors. The proposed approach is relatively simple to use and appears to have reasonable mean squared error properties. The computations required can, however, increase rapidly as the number of cells in the multi-way classification increase. Variance estimation is also considered.

KEY WORDS: Controlled selection; Linear programming; Multistage sampling; Stratified sampling.

1. INTRODUCTION

There are often several stratifying variables available to the sample designer and it is natural in such cases for the designer to consider defining strata as the cells formed by cross-classifying categories of these variables. A problem with this approach, particularly common when selecting primary sampling units (psu's) in household surveys, is that the desired sample size may be less than the total number of cells and hence conventional methods of stratification may be inapplicable.

An illustration, based on a hypothetical example of Bryant *et al.* (1960), is given in Table 1. Communities (psu's) are classified by two stratifying factors: type of community with three categories and region with five categories. The desired sample size of $n = 10$ is less than the total number of cells, 15. This example also illustrates a related problem. The entries in Table 1 are the expected counts under proportionate stratification, that is the population proportions multiplied by the sample size. Even if the sample size was doubled to exceed the number of cells, the expected sample counts would still not be integers. Whilst the effect of rounding such values to integers may not be practically significant for large expected counts, the choice of how to round with very small expected counts may be of greater concern.

One reaction to the problem of many cells is simply to drop one or more of the stratifying variables or to group some of the categories. Alternatively, a number of procedures have been proposed which attempt to retain some 'control' for all the categories of all the stratifying variables by permitting different forms of random selection of cells.

Goodman and Kish (1950) proposed one procedure under the title 'controlled selection'. Jessen (1970) suggests that 'this method is somewhat complicated and its use in applied sampling appears limited' (p. 778). Waterton (1983) illustrates this complexity. Bryant *et al.* (1960) propose a much simpler method for two-way stratification. Their method has the property that the expected sample counts display independence between the rows and columns of the two-way table. If the rows and columns are also independent in the population then there is no problem but if, as will often be the case, there is an appreciable lack of independence then some reweighting will usually be necessary and this can be unattractive in practice and can inflate the variance as is shown in Section 5. Jessen (1970) points out that a further limitation of the method of Bryant *et al.* (1960) is that it is not possible to constrain specified cell sizes to be zero. He proposes two approaches for both two-way and three-way stratification but both approaches remain fairly complicated to implement and, as noted by Causey *et al.* (1985), do not always lead to a solution.

All the above methods may be carried out by hand with varying degrees of laboriousness, but none take advantage of the power of modern computing. In this paper we shall show how computational procedures of linear programming can be applied to the multi-way stratification problem following Rao and Nigam (1990, 1992). Our proposed approach may be viewed as complementing the linear

Table 1

Expected Sample Cell Counts Under Proportionate Stratification with $n = 10$

Regions	Type of Community			
	Urban	Rural	Metropolitan	Total
1	1.0	0.5	0.5	2.0
2	0.2	0.3	0.5	1.0
3	0.2	0.6	1.2	2.0
4	0.6	1.8	0.6	3.0
5	1.0	0.8	0.2	2.0
Total	3.0	4.0	3.0	10.0

[1] R.R. Sitter, Department of Mathematics and Statistics, Carleton University, Ottawa, Ontario K1S 5B6; C.J. Skinner, Department of Social Statistics, University of Southampton S09 5NH, U.K.

programming approach proposed by Causey *et al.* (1985). Which of the two approaches is preferable will depend on the nature of the stratification problem and on the software available. The potential disadvantage of our approach is that it can be much more computationally intensive, since the number of unknowns in our linear programming problem may be as large as $\binom{k}{n}$, when k is the number of cells in the table and n is the sample size, whereas the number of unknowns in the approach of Causey *et al.* (1985) is only k. A number of suggestions will be made, however, to reduce the computational demands of our approach. There are several potential advantages of our approach. First, the stratification problem corresponds directly to the linear programming problem and so the computer programming is straightforward, whereas the approach of Causey *et al.* is less direct, involving mimicking the behaviour of nonlinear functions by linear functions (p. 904) and nesting repeated linear programming problems within a further recursive algorithm. Second, our procedure always has a solution, whereas the procedure of Causey *et al.* need not, for example in cases of three-way stratification. Third, the objective function in our linear programming problem can be naturally modified to reflect the different objectives of the stratification problem, for example in a three-way problem where it is more important to 'balance' the sample with respect to the first two stratifying variables than the third. Fourth, our procedure can be naturally modified to constrain the joint inclusion probabilities of cells to be positive in order to permit unbiased variance estimation.

2. THE PROPOSED APPROACH

2.1 Basic Ideas

We begin with the simplest kind of two-way stratification. Let a population of N units be classified into the RC cells of a two-way table formed by cross-classifying a row stratification factor with R categories and a column factor with C categories. Let N_{ij} be the number of units in cell ij, that is the set of units in both row i and column j, and let $P_{ij} = N_{ij}/N$ be the corresponding proportion. The parameter of interest is taken to be the population mean, $\bar{Y}$, of a variable Y.

Consider the following two-stage sampling procedure. First, sample sizes n_{ij} are determined for each cell according to a specified randomized procedure. Letting s denote the $R \times C$ array (n_{ij}, $i = 1, \ldots, R, j = 1, \ldots, C$), this procedure assigns a probability $p(s)$ to each s in a set S of possible arrays. To emphasize the dependence of n_{ij} on s we write $n_{ij}(s)$. Second, a simple random sample of $n_{ij}(s)$ units is selected from cell ij and the values of Y are recorded for the sample units.

We restrict attention to designs of fixed sample size $n > 0$, that is we restrict S to be the set S_n of all arrays such that

$$\sum_{i=1}^{R} \sum_{j=1}^{C} n_{ij}(s) = n.$$

We also restrict attention to proportionate stratification so that

$$\sum_{s\in S_n} n_{ij}(s)p(s) = nP_{ij} \quad \text{for} \quad i = 1, \ldots, R, \quad j = 1, \ldots, C. \tag{2.1}$$

It follows from (2.1) that the simple unweighted sample mean $\bar{y}(s)$ is an unbiased estimator of $\bar{Y}$. We propose to choose a (or the) sampling design $p(s)$ which minimizes the expected lack of 'desirability' of the sample s by solving the problem:

$$\underset{p\in P}{\text{minimize}} \sum_{s\in S_n} w(s)p(s), \tag{2.2}$$

subject to the constraint (2.1), where $w(s)$ is a loss function for the sample s to be specified and P is the class of possible sample designs on S_n obeying

$$0 \le p(s) \le 1 \quad \text{for all} \quad s \in S_n. \tag{2.3}$$

Note that (2.1) implies $\sum_{s\in S_n} p(s) = 1$. The key observation of Rao and Nigam (1990, 1992) is that the objective function in (2.2) and the equality and inequality constraints in (2.1) and (2.3) are all linear in $p(s)$ and hence this problem may be solved directly by linear programming with the $p(s)$, $s \in S_n$, as unknowns. The main obstacle to this approach is that the number of elements in S_n is often very large and even with modern computing power it becomes difficult to carry out linear programming if the number of unknowns is large.

It is therefore desirable to restrict attention to a subset of S_n. One natural restriction is to consider only arrays s for which $n_{ij}(s)$ is either equal to $I_{ij} = [nP_{ij}]$, the greatest integer less than nP_{ij}, or $I_{ij} + 1$. Letting $\tilde{n}_{ij}(s) = n_{ij}(s) - I_{ij}$ and $r_{ij} = nP_{ij} - I_{ij}$ the problem becomes

$$\underset{p\in P}{\text{minimize}} \sum_{s\in \tilde{S}_{\tilde{n}}} w(s)p(s), \tag{2.4}$$

subject to

$$\sum_{s\in \tilde{S}_{\tilde{n}}} \tilde{n}_{ij}(s)p(s) = r_{ij}, \tag{2.5}$$

$$\sum_{s\in \tilde{S}_{\tilde{n}}} p(s) = 1, \; 0 \le p(s) \le 1 \quad \text{for all} \quad s \in \tilde{S}_{\tilde{n}}, \tag{2.6}$$

where $\tilde{S}_{\tilde{n}}$ is the set of $R \times C$ arrays, where all elements are 0 or 1 and the sum of elements is $\tilde{n} = n - \sum_{ij} I_{ij}$. Note, of course, that if all the I_{ij} are zero, then this is just the same problem as before. The number of elements in $\tilde{S}_{\tilde{n}}$, which determines the magnitude of the computational task for linear programming, is now $\binom{RC}{\tilde{n}}$. This number can still be very large, however, and some further reduction can be achieved by sensible choice of the loss function $w(s)$ as discussed in the next section.

For Table 1, this would amount to considering the situation represented by Table 2, while only allowing a 0 or 1 cell sample size, and then adding back 1 to cells (1,1), (3,3), (4,2) and (5,1) in the final solution. Thus $n = 10$, but $\tilde{n} = 6$.

Table 2

Table of r_{ij} Values from Table 1 with $\tilde{n} = 6$

Regions	Type of Community			
	Urban	Rural	Metropolitan	Total
1	0.0	0.5	0.5	1.0
2	0.2	0.3	0.5	1.0
3	0.2	0.6	0.2	1.0
4	0.6	0.8	0.6	2.0
5	0.0	0.8	0.2	1.0
Total	1.0	3.0	2.0	6.0

2.2 Choice of Loss Function *w(s)*

The major flexibility of the proposed approach derives from the user's freedom to choose the function $w(s)$ which enters the objective function in (2.2). The conventional approach to two-way stratification (*e.g.*, Jessen 1970; Causey *et al.* 1985) is to require that the selected sample s obey the marginal constraints:

$$| n_{i\cdot}(s) - nP_{i\cdot} | < 1 \quad i = 1, \ldots, R, \tag{2.7}$$

$$| n_{\cdot j}(s) - nP_{\cdot j} | < 1 \quad j = 1, \ldots, C, \tag{2.8}$$

where

$$n_{i\cdot}(s) = \sum_j n_{ij}(s), \quad n_{\cdot j}(s) = \sum_i n_{ij}(s)$$

$$P_{i\cdot} = \sum_j P_{ij}, \quad P_{\cdot j} = \sum_i P_{ij}.$$

This requirement can be accommodated in our approach by setting $w(s)$ as (effectively) infinite for samples s not satisfying (2.7) or (2.8) or more simply by excluding such samples from the set S_n. The problem with this conventional approach is that no solution to the constrained optimization-problem may exist.

In our approach, however, if we use a loss function such as

$$w(s) = \sum_{i=1}^{R} (n_{i\cdot}(s) - nP_{i\cdot})^2 + \sum_{j=1}^{C} (n_{\cdot j}(s) - nP_{\cdot j})^2, \tag{2.9}$$

then an optimal solution will always exist within a large enough set S_n. In practice, it may be advantageous computationally to restrict the set S_n initially to only those samples obeying (2.7) and (2.8), or even a subset of these, and then to expand the set if necessary, say by changing 1 to 2 in (2.7) and (2.8), until a solution is found.

Let us now consider the more fundamental question of why constraints such as (2.7) and (2.8) are sensible anyway. From a non-statistical point of view, the balancing of a sample with respect to factors with a known population distribution may reassure users about the 'representativeness' of the sample. From a statistical point of view, given our unbiasedness constraint (2.1), it is natural to consider how the loss function might be chosen to improve efficiency. This question may be examined by taking $w(s)$ as the mean squared error $E_m(\bar{y}(s) - \bar{Y})^2$ under a model m. Then the solution to the optimization problem (2.2) minimizes the design-expected model-mean squared error or equivalently, since we require design-unbiasedness, the model-expected design variance.

Consider, for example, the main-effects analysis of variance model

$$y_{ijk} = \mu + \alpha_i + \beta_j + \epsilon_{ijk},$$

where y_{ijk} is the kth value of Y in cell ij, μ is a fixed mean and α_i, β_j and ϵ_{ijk} are independent zero-mean random effects with variances σ_α^2, σ_β^2 and σ_ϵ^2, respectively. Then, ignoring finite population correction terms,

$$E_m(\bar{y}(s) - \bar{Y})^2 = \sigma_\alpha^2 \sum_i (n_{i\cdot}(s)/n - P_{i\cdot})^2 + \sigma_\beta^2 \sum_j (n_{\cdot j}(s)/n - P_{\cdot j})^2 + \sigma_\epsilon^2/n. \tag{2.10}$$

Hence, if $\sigma_\alpha^2 = \sigma_\beta^2$ the expected design variance of $\bar{y}(s)$ under this model is minimized by taking the loss function in (2.9). Alternatively, if one had some prior information about the likely ratio of the between row variance relative to the between column variance then it may be sensible, on efficiency grounds, to modify the loss function in (2.9) by multiplying the first term on the right hand side of (2.9) by this estimated ratio.

On the other hand if it is thought *a priori* that there is likely to be a strong interaction between the row and column factors in their effect on Y then simply attempting to balance on the margins may be inappropriate. For example, if one stratification factor is urban/rural and the other is an economic indicator X and it is known that Y is positively related to X in urban areas and negatively related in rural areas then it is likely to be more efficient to stratify partially by *X separately* within rural and urban areas than to balance fully on both margins. See Bryant *et al.* (1960, section 9) for related comments on efficiency for two-way stratification.

2.3 Higher-way Stratification

The proposed approach extends naturally to 3 or more stratifying factors by letting s denote the corresponding r-way array. The loss function will typically include further terms, for example for three-way stratification we might take

$$w(s) = \lambda_1 \sum_{i=1}^{R_1} (n_{i\cdot\cdot}(s) - nP_{i\cdot\cdot})^2 + \lambda_2 \sum_{j=1}^{R_2} (n_{\cdot j\cdot}(s) - nP_{\cdot j\cdot})^2 + \lambda_3 \sum_{k=1}^{R_3} (n_{\cdot\cdot k}(s) - nP_{\cdot\cdot k})^2$$

in obvious notation, where λ_1, λ_2 and λ_3 are included to represent the relative importance of balancing on the three factors and might consist of prior estimates of the variances of the Y means between categories of the three stratifying factors, as in (2.10).

2.4 Multistage Sampling

One important practical application of multi-way stratification is to the selection of primary sampling units (psu's) in multistage sampling, where it is common for information of several stratifying factors to be available.

In the approach of Section 2.1, the inclusion probabilities of each population unit are $E(n_{ij}(s)/N_{ij}) = n/N$. If it is desired to select psu's with equal probability then this approach extends immediately with the psu's constituting the units and with the observed values of Y replaced by unbiased estimators of the psu totals. Suppose instead that it is desired to select psu's with unequal probabilities, say nz_{ijk} for psu k in cell ij, where usually z_{ijk} will equal $M_{ijk}/\sum_{ijk} M_{ijk}$, with M_{ijk} being some measure of size of psu k in cell ij. Then the procedure may be simply modified by setting P_{ij} equal to the sum of z_{ijk} over psu's k in cell ij. Then, if $n_{ij}(s) > 0$, a sample of $n_{ij}(s)$ psu's in cell ij is selected by some probability proportional to z_{ijk} method.

3. EXAMPLES

Example 1: Bryant, Hartley and Jessen (1960)

We will first demonstrate the method on the hypothetical example of Bryant *et al.* (1960) given in Table 1. We first reduce the problem to the form of (2.4), (2.5) and (2.6), where the r_{ij}'s are given in Table 2. The weight function in (2.9) in this reduced linear programming problem becomes

$$w(s) = \sum_{i=1}^{5} (\tilde{n}_{i\cdot}(\tilde{s}) - r_{i\cdot})^2 + \sum_{j=1}^{3} (\tilde{n}_{\cdot j}(\tilde{s}) - r_{\cdot j})^2.$$

Applying a standard linear programming package in the NAG FORTRAN library, we obtain the solution given in Table 3. The I_{ij} values have been added to the solution so that $n_{ij} = I_{ij} + \tilde{n}_{ij}(\tilde{s})$. It turns out for this solution that each s, for which $p(s) > 0$, has margins $n_{i\cdot}(s)$ and $n_{\cdot j}(s)$ which match the desired margins exactly, that is the solution makes (2.4) zero.

Table 3

Solution to Example 1

s	$p(s)$	s	$p(s)$
1 1 0 1 0 0 0 1 1 0 2 1 1 0 1	0.2	1 1 0 0 0 1 0 1 1 1 1 1 1 1 0	0.1
1 1 0 0 0 1 0 0 2 1 2 0 1 1 0	0.2	1 0 1 0 1 0 1 0 1 0 2 1 1 1 0	0.2
1 0 1 0 1 0 0 1 1 1 1 1 1 1 0	0.1	1 0 1 0 0 1 0 1 1 1 2 0 1 1 0	0.2

Example 2: Jessen (1970)

Jessen (1970) proposed two methods for two-way and three-way stratification. Both of these are quite complicated and involve determining the set of samples which exactly match the margins. Neither method is guaranteed to yield a solution. Jessen (1970) applies both methods to a simple hypothetical example for which both yield a solution. This example is reproduced in Table 4. In this example, since all of the $nP_{ij} < 1$, the linear programming problems defined by (2.1), (2.2) and (2.3) and by (2.4), (2.5) and (2.6), respectively, are identical. We applied our method to this problem, again using the $w(s)$ as defined in (2.9). By trying a number of different seeds in the optimization routine, we were able to obtain three different solutions, all of which make (2.2) zero and satisfy the constraints. These are given in Table 5. The first two solutions are the same two as obtained by Jessen's method 2 and method 3, respectively.

Table 4

Example 2: Jessen (1970)
Expected Sample Cell Counts Under Proportionate Stratification with $n = 6$

Rows	Columns			$nP_{i\cdot}$
	1	2	3	
1	0.8	0.5	0.7	2.0
2	0.7	0.8	0.5	2.0
3	0.5	0.7	0.8	2.0
$nP_{\cdot j}$	2.0	2.0	2.0	6.0

Table 5

Solution to Example 2

s	$p_1(s)$	$p_2(s)$	$p_3(s)$
1 0 1 1 1 0 0 1 1	0.5	0.4	0.3
1 1 0 0 1 1 1 0 1	0.3	0.2	0.1
0 1 1 1 0 1 1 1 0	0.2	0.1	0.0
1 1 0 1 0 1 0 1 1	0.0	0.1	0.2
1 0 1 0 1 1 1 1 0	0.0	0.1	0.2
0 1 1 1 1 0 1 0 1	0.0	0.1	0.2

Example 3: Causey, Cox and Ernst (1985)

Causey *et al.* (1985) give an example of three-way stratification for which their method fails to yield a solution. They consider a population subject to a 2 × 2 × 2 stratification from which a sample of size $n = 2$ is to be drawn, with the expected sample size in the ijk-th cell, n_{ijk}, as follows:

$$n_{111} = n_{221} = n_{122} = n_{212} = .5$$

$$n_{121} = n_{211} = n_{112} = n_{222} = 0.$$

If we apply our method in a similar manner to Examples 1 and 2 we obtain the solution given in Table 6. In this case, the objective function did not attain zero so that the margins are not exactly matched in each sample.

Table 6

Solution to Example 3

s		$p(s)$
$i = 1$	$i = 2$	
1 0 0 0	0 1 0 0	0.5
0 0 0 1	0 0 1 0	0.5

4. COMPARISON OF MSE

In this section the mean squared error (MSE) of the proposed design with estimator $\bar{y}$ will be compared with the MSE of the design of Bryant *et al.* (1960) with either of the two estimators they propose, namely $\bar{y}_U$ and $\bar{y}_B$, where the U and B subscripts indicate that the first estimator is unbiased and the second is not. Let the cells be denoted c (ij in the two-way case), let k (and where necessary l) denote a unit within a cell, and suppress the s in $n_c(s)$ for simplicity of notation. The inclusion probability of any unit k in cell c is

$$\pi_{ck} = E[n_c]/N_c = E[n_c]/(NP_c) \tag{4.1}$$

and the joint inclusion probability of unit k in cell c and unit k' in cell c' is

$$\pi_{ckc'k'} = \begin{cases} \dfrac{E[n_c(n_c-1)]}{N_c(N_c-1)} & \text{if} \quad c = c' \\ \dfrac{E[n_c n_{c'}]}{N_c N_{c'}} & \text{if} \quad c \neq c'. \end{cases} \tag{4.2}$$

For large N this is approximately

$$\pi_{ckc'k'} \doteq \frac{E(n_c n_{c'})}{N^2 P_c P_{c'}} - \frac{E(n_c)}{N^2 P_c^2} I_{[c=c']}, \tag{4.3}$$

where

$$I_{[c=c']} = \begin{cases} 1 & \text{if} \quad c = c' \\ 0 & \text{if} \quad c \neq c'. \end{cases}$$

The expectations will differ for our design compared to the Bryant *et al.* design and thus the π_{ck} and $\pi_{ckc'k'}$ will differ. Keeping this in mind we can obtain the variance of $\bar{y}$, $\bar{y}_U$ and $\bar{y}_B$ in a generalized form in terms of the π_{ck}

and $\pi_{ckc'k'}$ values and thus have some basis for comparison. To do this, let us consider an estimator of the form $\bar{z} = \sum_c \sum_k w_c y_{ck}/n$, where the w_c values are fixed known constants independent of k. If we restrict to the case where $n_{i.} = nP_{i.}$ and $n_{.j} = nP_{.j}$, that is, integer marginal requirements, then both of the estimators given in Bryant *et al.* as well as our estimator are of this form. We will assume this to be the case in the sequel. Replacing the subscript c with ij for two-way stratification, $\bar{y}_U$ and $\bar{y}_B$ are of the same form as $\bar{z}$ with $w_c = w_{ij} = G_{ij} = P_{ij}/(P_{i.}P_{.j})$ and $w_c = w_{ij} = 1$, respectively. The estimator $\bar{y}$ is also of the form $\bar{z}$ with $w_c = w_{ij} = 1$.

We can now obtain a general form for the variance of $\bar{z}$ keeping in mind that the π_{ck} and $\pi_{ckc'k'}$ values will differ for the Bryant *et al.* design and our design:

$$V(\bar{z}) = \frac{1}{2n^2} \sum_c \sum_{c'} \sum_k \sum_{k'} (\pi_{ck}\pi_{c'k'} - \pi_{ckc'k'}) (w_c y_{ck} - w_{c'} y_{c'k'})^2. \quad (4.4)$$

Using (4.1) and (4.3) this becomes

$$V(\bar{z}) = \frac{1}{2n^2} \sum_c \frac{w_c^2 E(n_c)}{N^2 P_c^2} \sum_k \sum_{k'} (y_{ck} - y_{ck'})^2 - \frac{1}{2n^2} \sum_c \sum_{c'} \frac{\mathrm{Cov}(n_c, n_{c'})}{N^2 P_c P_{c'}} \sum_k \sum_{k'} (w_c y_{ck} - w_{c'} y_{c'k'})^2. \quad (4.5)$$

Noting that

$$\sum_k \sum_l (y_{ck} - y_{cl})^2 = 2N^2 P_c^2 S_c^2$$

and

$$\sum_k \sum_{k'} (w_c y_{ck} - w_{c'} y_{c'k'})^2 = N^2 P_c P_{c'} [w_c^2 S_c^2 + w_{c'}^2 S_{c'}^2 + (w_c \bar{Y}_c - w_{c'} \bar{Y}_{c'})^2],$$

where S_c^2 refers to the population variance of cell c, (4.5) reduces to

$$V(\bar{z}) = \frac{1}{n^2} \sum_c w_c^2 E(n_c) S_c^2 - \frac{1}{2n^2} \sum_c \sum_{c'} \mathrm{Cov}(n_c, n_{c'}) [w_c^2 S_c^2 + w_{c'}^2 S_{c'}^2 + (w_c \bar{Y}_c - w_{c'} \bar{Y}_{c'})^2]$$
$$= v_1 + v_2, \quad \text{say.} \quad (4.6)$$

The first term v_1 may be interpreted as the usual stratified variance for fixed sample sizes $E(n_c)$ within the two-way 'strata' (of course in our case the $E(n_c)$ will generally not be integers). The second term v_2 may be interpreted as the increase in variance arising from the variability of the n_c and the correlation between them. We discuss this further at the end of this section. We now revert to the notation $c = ij$ and compare the variances for two-way stratification.

First let us consider v_1 in (4.6). For the Bryant *et al.* method $E(n_{ij}) = nP_{i.}P_{.j}$, $\bar{y}_U = \sum_i \sum_j \sum_k G_{ij} y_{ijk}/n$, $G_{ij} = P_{ij}/(P_{i.}P_{.j})$ and $\bar{y}_B = \sum_i \sum_j \sum_k y_{ijk}/n$.

Thus

$$v_1(\bar{y}_U) = \sum_i \sum_j P_{ij} G_{ij} S_{ij}^2 / n,$$

(this is the same as the first term of equation (12) in Bryant *et al.*) and

$$v_1(\bar{y}_B) = \sum_i \sum_j P_{i.} P_{.j} S_{ij}^2 / n.$$

In the case of our approach $E(n_{ij}) = nP_{ij}$ and $\bar{y} = \sum_i \sum_j \sum_k y_{ijk}/n$ so that

$$v_1(\bar{y}) = \sum_i \sum_j P_{ij} S_{ij}^2 / n.$$

Next let us consider v_2. It is not difficult to show that for both the Bryant *et al.* method and our approach (see Appendix)

$$\sum_i \mathrm{Cov}(n_{ij}, n_{i'j'}) = \sum_j \mathrm{Cov}(n_{ij}, n_{i'j'}) = 0. \quad (4.7)$$

Using this and replacing c and c' by ij and $i'j'$, respectively, in v_2 given in (4.6), it follows that v_2 reduces to

$$v_2 = \frac{1}{n^2} \sum_i \sum_j \sum_{i'} \sum_{j'} \mathrm{Cov}(n_{ij}, n_{i'j'}) w_{ij} w_{i'j'} \bar{Y}_{ij} \bar{Y}_{i'j'}.$$

Replacing w_{ij} with G_{ij} we get $v_2(\bar{y}_U)$, and using simple algebra one can show that this is the same as term 2 of equation (12) in Bryant *et al.* Replacing w_{ij} with 1 gives the form of $V(\bar{y}_B)$ and of $V(\bar{y})$, noting that the $\mathrm{Cov}(n_{ij}, n_{i'j'})$ will not be the same for both. So we see that v_2 depends only on the cell means while v_1 depends only on the within cell variances.

Finally, we should note that

$$\mathrm{bias}(\bar{y}_B) = -\sum_i \sum_j (P_{ij} - P_{i.}P_{.j}) \bar{Y}_{ij}, \quad (4.8)$$

since to compare the three estimators the mean square error (MSE) will be the relevant measure, and this bias will contribute to $\mathrm{MSE}(\bar{y}_B)$.

Combining the expressions for v_1, v_2 and bias$(\bar{y}_B)$ above permits an analytical comparison of the MSE of the proposed approach with that of the approach of Bryant *et al.* (1960) using either $\bar{y}_U$ and $\bar{y}_B$. It is difficult, however, to make general statements about the relative performance of the different strategies and so we now consider introducing some model assumptions in order to approximate the different components of the MSE expressions, in some specific settings. We first consider the additive model:

$$y_{ijk} = \mu + \alpha_i + \beta_j + \epsilon_{ijk},$$

where y_{ijk} is the k-th observation in the ij-th cell, α_i and β_j are fixed effects and ϵ_{ijk} are independent errors with zero mean and common variance σ^2. Then $\mathrm{E}_m(S_{ij}^2) = \sigma^2$ and $\mathrm{E}_m(\bar{Y}_{ij}\bar{Y}_{i'j'}) \doteq (\mu + \alpha_i + \beta_j)(\mu + \alpha_{i'} + \beta_{j'})$. Thus the model-expected design-variance is given by replacing S_{ij}^2 by σ^2 and $\bar{Y}_{ij}$ by $\mu + \alpha_i + \beta_j$ in the formulas for v_1 and v_2 for the various estimators. In this case, $v_2(\bar{y}_B) = 0$. This point was realized by Bryant *et al.* when comparing $\bar{y}_U$ and $\bar{y}_B$. The bias term will be zero in this case unless there was rounding on the margins, that is bias$(\bar{y}_B) = 0$ provided $n_{i\cdot} = nP_{i\cdot}$ and $n_{\cdot j} = nP_{\cdot j}$ as is the case in their example. This easily follows from (4.8) and

$$\sum_i (P_{ij} - P_{i\cdot}P_{\cdot j}) = \sum_j (P_{ij} - P_{i\cdot}P_{\cdot j}) = 0.$$

This was also shown by Bryant *et al.* p. 119 equation (47). Using (4.7), it is easily shown that $v_2(\bar{y}) = 0$ as well. This combined with the unbiasedness of $\bar{y}$ and the fact that $v_1(\bar{y}_B) = v_1(\bar{y}) = \sigma^2/n$ in this case implies that for this situation MSE$(\bar{y}_B)$ = MSE$(\bar{y})$, that is the proposed procedure has the same MSE as the procedure of Bryant *et al.* using the biased estimator. We demonstrate in the sequel that even when this additive model is applicable ($\gamma = 0$ below), $v_2(\bar{y}_U)$ may be large while $v_1(\bar{y}_U) > v_1(\bar{y})$.

To compare the estimators further, let us consider the situation of Example 1. The above derivations allow us to obtain the MSE's of the three estimators for this example provided we have the S_{ij}'s, the $\bar{Y}_{ij}$'s and can calculate the Cov$(n_{ij}, n_{i'j'})$ for the Bryant *et al.* method as well as for our approach. The covariances for the Bryant *et al.* method are given in their paper in terms of the P_{ij}'s, while the covariances for our approach can be obtained from the solution in Table 3. We will consider non-additive departures from the above model, namely

$$y_{ijk} = \mu + \alpha_i + \beta_j + \gamma\alpha_i\beta_j + \epsilon_{ijk},$$

for various values of γ. For simplicity of presentation, let $\mu = 1$, $\alpha_i = i - 3$, $\beta_j = j - 2$ (note in fact that the MSE of each strategy is invariant to the choice of μ). Thus the model-expected design-variance is given by replacing S_{ij}^2 by 1 and $\bar{Y}_{ij}$ by $1 + (i-3) + (j-2) + \gamma(i-3)(j-2)$ in the formulas for v_1 and v_2 for the various estimators. Table 7 gives the resulting v_1, v_2, and MSE values for the three estimators (as well as the bias squared term for $\bar{y}_B$), for various values of γ. From Table 7, it can be seen that for an additive model, $\gamma = 0$, $\bar{y}_B$ and $\bar{y}$ perform equally well, while $\bar{y}_U$ is inferior. As the model becomes more non-additive, and $|\gamma|$ increases, the two estimators for the Bryant *et al.* strategy tend to perform similarly, both with MSE becoming increasingly greater than that of the proposed strategy. This pattern is primarily due to the v_2 component of the MSE of the three estimators. The bias term of $\bar{y}_B$ is of lesser importance, although it may be more important for larger n.

The greater increase in v_2 as $|\gamma|$ increases for the Bryant *et al.* design appears to reflect the greater variability of each n_{ij} for this design. It should be noted that it would have been possible to reduce this variability somewhat by applying a variant of the Bryant *et al.* method instead to Table 2, as was done for the proposed method, though one would need to derive adjusted G_{ij} weights for $\bar{y}_U$ and it would be difficult to handle the 0.0 cell entries in Table 2. However, even if this were accomplished, the $\tilde{n}_{ij}$ for this design may still take values other than just 0 and 1; for example n_{42} could take values 0, 1, or 2. This inflated n_c variability is inherent in the Bryant *et al.* method. For example, suppose $n_{1\cdot} = n_{\cdot 1} = 5$. Then using the Bryant *et al.* method, n_{11} can take values 0, 1, 2, 3, 4, or 5, while with the proposed method it can take only values $[nP_{11}]$ or $[nP_{11}] + 1$. If $nP_{11} < 1$, the technique used to go from Table 1 to Table 2 will not improve matters.

Table 7
Comparison of MSE for Three Estimators

	Bryant, Hartley, Jessen Design							Proposed Design		
	$\bar{y}_U$			$\bar{y}_B$				$\bar{y}$		
γ	v_1	v_2	MSE	v_1	v_2	Bias^2	MSE	v_1	v_2	MSE
0	.125	.105	.230	.100	.000	.000	.100	.100	.000	.100
±.5	.125	.063	.188	.100	.033	.002	.135	.100	.018	.118
±1	.125	.105	.230	.100	.131	.008	.239	.100	.071	.171
±2	.125	.440	.565	.100	.523	.032	.655	.100	.284	.384
±3	.125	1.111	1.236	.100	1.176	.073	1.349	.100	.638	.738

5. VARIANCE ESTIMATION

In this section, we will consider variance estimation for our proposed method. Using (4.1) and recalling constraint (2.1), it is clear that

$$\pi_{ck} = E[n_c(s)/N_c] = n/N.$$

The joint inclusion probability of two units k, k' in the same cell c is

$$\pi_{ck,ck'} = E[n_c(s)\{n_c(s) - 1\}/\{N_c(N_c - 1)\}].$$

Suppose $n_c(s) = I_c + \tilde{n}_c(s)$ when I_c is the fixed integer $[nP_c]$ and $\tilde{n}_c(s) = 0$ or 1.

If $nP_c \leq 1$ then $I_c = 0$ and $\pi_{ck,ck'} = 0$. Hence a necessary condition for unbiased variance estimation to be possible is that $nP_c > 1$ for all cells c. On the other hand if this condition holds then $n_c(s) \geq 1$ for all c and hence the probability of inclusion of any pair of units in different cells is also always positive. Hence this condition is necessary and sufficient for unbiased variance estimation to be possible.

When this condition holds we obtain

$$\pi_{ck,ck'} = I_c(I_c + 2r_c - 1)/[N_c(N_c - 1)] = A_c,$$

say, where $r_c = E[\tilde{n}_c(s)] = nP_c - I_c$.

The joint inclusion probability for pairs of units in different cells c and c' are

$$\begin{aligned}\pi_{ck,ck'} &= E[n_c(s)n_{c'}(s)/(N_cN_{c'})] \\ &= [I_cI_{c'} + r_{c'}I_c + r_cI_{c'} + r_{cc'}]/(N_cN_{c'}) = B_{cc'},\end{aligned} \tag{5.1}$$

say where $r_{cc'} = E[\tilde{n}_c(s)\tilde{n}_{c'}(s)]$.

Hence an unbiased estimator of $V(\bar{y}(s))$ of Sen-Yates-Grundy form may be constructed in the usual way.

In practice, however, we wish to consider situations where $nP_c \leq 1$ for some c. In this case one assumption we might make following Bryant *et al.* (1960, Sect. 7) in order to derive a variance estimator is that the population variance of Y is constant within each cell c, say S^2.

Let us first obtain the variance of $\bar{y}(s)$ in the general case

$$V(\bar{y}(s)) = \frac{1}{2n^2}\sum_c\sum\sum_{k \neq k'}\left(\frac{n^2}{N^2} - A_c\right)(y_{ck} - y_{ck'})^2$$

$$+ \frac{1}{2n^2}\sum\sum_{c \neq c'}\sum_{k,k'}\left(\frac{n^2}{N^2} - B_{cc'}\right)(y_{ck} - y_{c'k'})^2.$$

Now providing $B_{cc'} > 0 \;\forall\; c, c'$ we may estimate the second term unbiasedly by

$$\frac{1}{2n^2}\sum\sum_A \sum_{k=1}^{n_c(s)} \sum_{k'=1}^{n_{c'}(s)}\left(\frac{\frac{n^2}{N^2} - B_{cc'}}{B_{cc'}}\right)(y_{ck} - y_{c'k'})^2,$$

where $A = \{c,c' : n_c(s) \geq 1, n_{c'}(s) \geq 1, c \neq c'\}$.

The first term can be written as

$$\frac{1}{2n^2}\sum_c\left(\frac{n^2}{N^2} - A_c\right)2N_c^2S^2.$$

For any c s.t. $n_c(s) \geq 2$

$$E\left(\sum_{k=1}^{n_c(s)}\sum_{\substack{k'=1 \\ k \neq k'}}^{n'_c(s)} \frac{(y_{ck} - y_{ck'})^2}{2n_c(s)\{n_c(s) - 1\}} \,\middle|\, n_c(s)\right) = S^2.$$

Thus provided at least one $n_c(s)$ is ≥ 2 an unbiased estimator of the first term is

$$\frac{1}{2n^2D}\sum_{\{c:n_c(s) \geq 2\}}\left(\frac{n^2}{N^2} - A_c\right)2N_c^2\sum_{k=1}^{n_c(s)}\sum_{k'=1}^{n'_c(s)} \frac{(y_{ck} - y_{ck'})^2}{2n_c(s)\{n_c(s) - 1\}}$$

where $D =$ the number of cells, c, such that $n_c(s) \geq 2$.

The above requires $B_{cc'} > 0$. If

$$I_c = I_{c'} = 0,$$

by (5.1), we need

$$r_{cc'} = \sum \tilde{n}_c(s)\tilde{n}_{c'}(s)p(s) > 0, \tag{5.2}$$

which is linear in $p(s)$. The constraint (5.2) can be handled in linear programming if desired. There will be such a constraint for each pair c, c' s.t. $I_c = I_{c'} = 0$.

6. CONCLUDING REMARKS

We have proposed a linear programming approach to multi-way stratification, applying ideas of Rao and Nigam (1990, 1992). The approach is simple in conception and is very flexible in allowing for a range of different objectives via the loss function $w(s)$, as well as in permitting a variety of constraints such as that the joint inclusion probabilities of all cells be positive. The main practical constraint on the procedure is that it may rapidly become computationally expensive if not impossible as the number of cells in the multi-way classification increases. Some ideas on how to reduce the amount of computation have been considered. Further research on this question would be useful. For cases where the computational demands are prohibitive, the method of Causey *et al.* (1985) remains an alternative.

ACKNOWLEDGEMENTS

The authors wish to express their thanks to Wesley Yung for his computer programming assistance and J.N.K. Rao for helpful comments in the early stages of discussion. R.R. Sitter is supported by the Natural Sciences and Engineering Council of Canada. Our thanks are due to the referees and the Associate Editor for constructive suggestions and comments.

APPENDIX

Proof of (4.7) for Proposed Method

Note that

$$\mathrm{Cov}(n_{ij}(s),n_{i'j'}(s)) = \mathrm{E}(n_{ij}(s)n_{i'j'}(s)) - \mathrm{E}(n_{ij}(s))\mathrm{E}(n_{i'j'}(s)).$$

Equation (2.1) states that $\mathrm{E}(n_{ij}(s)) = nP_{ij}$. By definition

$$\mathrm{E}(n_{ij}(s)n_{i'j'}(s)) = \sum_s n_{ij}(s)n_{i'j'}(s)p(s).$$

Thus

$$\sum_j \mathrm{E}(n_{ij}(s))\mathrm{E}(n_{i'j'}(s)) = n^2P_{i'j'}\sum_j P_{ij} = n^2P_{i'j'}P_{i\cdot}\,, \tag{7.1}$$

and

$$\sum_j \mathrm{E}(n_{ij}(s)n_{i'j'}(s)) = \sum_j\sum_s n_{ij}(s)n_{i'j'}(s)p(s)$$
$$= \sum_s p(s)n_{i'j'}(s)\sum_j n_{ij}(s). \tag{7.2}$$

Assume that the solution to the linear optimization problem (2.2) equals zero, where $w(s)$ is given in (2.9). In this case, $\sum_j n_{ij}(s) = n_{i\cdot}(s) = nP_{i\cdot}$ and (7.2) implies

$$\sum_j \mathrm{E}(n_{ij}(s)n_{i'j'}(s)) = \sum_s p(s)n_{i'j'}(s)nP_{i\cdot}$$
$$= nP_{i\cdot}\sum_s n_{i'j'}(s)p(s)$$
$$= nP_{i\cdot}.\mathrm{E}(n_{i'j'}(s)) = nP_{i\cdot}.nP_{i'j'}\,. \tag{7.3}$$

Equations (7.1) and (7.3) together imply $\sum_j \mathrm{Cov}(n_{ij}(s), n_{i'j'}(s)) = 0$. It can be similarly shown that

$$\sum_i \mathrm{Cov}(n_{ij}(s),n_{i'j'}(s)) = \sum_{i'} \mathrm{Cov}(n_{ij}(s),n_{i'j'}(s))$$
$$= \sum_{j'} \mathrm{Cov}(n_{ij}(s),n_{i'j'}(s)) = 0.$$

REFERENCES

BRYANT, E.C., HARTLEY, H.O., and JESSEN, R.J. (1960). Design and estimation in two-way stratification. *Journal of the American Statistical Association*, 55, 105-124.

CAUSEY, B.D., COX, L.H., and ERNST, L.R. (1985). Applications of transportation theory to statistical problems. *Journal of the American Statistical Association*, 80, 903-909.

GOODMAN, R., and KISH, L. (1950). Controlled selection-a technique in probability sampling. *Journal of the American Statistical Association*, 45, 350-372.

JESSEN, R.J. (1970). Probability sampling with marginal constraints. *Journal of the American Statistical Association*, 65, 776-795.

RAO, J.N.K., and NIGAM, A.K. (1990). Optimal controlled sampling design. *Biometrika*, 77, 807-814.

RAO, J.N.K., and NIGAM, A.K. (1992). 'Optimal' controlled sampling: a unified approach. *International Statistical Review*, 60, 89-98.

WATERTON, J.J. (1983). An exercise in controlled selection. *Applied Statistics*, 32, 150-164.

Regression Weighting in the Presence of Nonresponse with Application to the 1987-1988 Nationwide Food Consumption Survey

WAYNE A. FULLER, MARIE M. LOUGHIN and HAROLD D. BAKER[1]

ABSTRACT

A regression weight generation procedure is applied to the 1987-1988 Nationwide Food Consumption Survey of the U.S. Department of Agriculture. Regression estimation was used because of the large nonresponse in the survey. The regression weights are generalized least squares weights modified so that all weights are positive and so that large weights are smaller than the least squares weights. It is demonstrated that the regression estimator has the potential for large reductions in mean square error relative to the simple direct estimator in the presence of nonresponse.

KEY WORDS: Non-negative weights; Consistency.

1. INTRODUCTION

In many sampling situations, the population means of auxiliary variables are known, but the particular values of the variables for individual elements are not used in the sample selection. Although the information is not used in the sampling design, it may be highly desirable to incorporate the information about population means into the estimation procedure. Common estimation procedures utilizing auxiliary information are ratio estimation, post-stratification, regression estimation, and raking. Regression estimation is the most general procedure in that the regression method can handle multiple auxiliary variables, continuous auxiliary variables, and discrete auxiliary variables. Post-stratification can be considered a special case of regression estimation in which the regression variables are indicator variables for the post strata. The raking procedure, also known as iterative proportional fitting, is restricted to auxiliary information in the form of discrete categories. Deming and Stephan (1940), Stephan (1942), El-Badry and Stephan (1955), Ireland and Kulblack (1968), Darroch and Ratcliff (1972), Brackstone and Rao (1979), and Oh and Scheuren (1987) are references on raking.

Early applications of regression estimation are Watson (1937), Cochran (1942) and Jessen (1942). Cochran (1977, Ch. 7) contains the basic theory. Regression estimation for survey samples has been discussed by numerous authors, including Mickey (1959), Fuller (1975), Royall and Cumberland (1981), Isaki and Fuller (1982), Wright (1983), Luery (1986), Alexander (1987), Bethlehem and Keller (1987), Copeland, Pritzmeier, and Hoy (1987), Lemaître and Dufour (1987), Särndal, Swensson and Wretman (1989), Deville and Särndal (1992), Zieschang (1990), and Rao (1992).

In much of the cited literature, regression estimation is described as a procedure for reducing variance in probability samples. In practice, one of the motivations for regression estimation is the potential for reducing bias associated with selective nonresponse. See, for example, Little and Rubin (1987, p. 55) for the special case of adjustment cells, and Bethlehem (1988) for the generalized regression estimator.

Nonresponse prompted the use of regression estimation in our application and we discuss regression estimation in the response adjustment context in Section 3. The standard regression estimator and the modified procedure that produces positive weights are introduced in Section 2. Application of the regression weighting procedure to the Nationwide Food Consumption Survey is described in Section 4.

2. REGRESSION ESTIMATOR

To introduce the regression estimator used in our study, assume that a sample containing n units has been selected and that the probability of selecting unit i is π_i. For our purposes, it is sufficient for π_i to be proportional to the selection probabilities. The sample might be a two-stage stratified sample, and the unit can be either the primary sampling unit or the observation unit. In our application, the unit is the observation unit. Assume that a k-dimensional vector of population means, denoted by $\bar{X} = (\bar{X}_1, \bar{X}_2, \ldots, \bar{X}_k)$ is known, that the vector $(y_i, x_{i1}, x_{i2}, \ldots, x_{ik})$ is observed for every unit in the sample and that an estimator of the mean of y is desired. We assume that the first element of x_i is one for all i. Hence, the first element of $\bar{X}$ is also one. The vector $x_i = (x_{i1}, x_{i2}, \ldots, x_{ik})$ is sometimes

[1] Wayne A. Fuller, Marie M. Loughin and Harold D. Baker, Iowa State University.

called the vector of control variables. A regression estimator of the mean of y is

$$\bar{y}_r = \bar{X}\hat{\beta}, \tag{2.1}$$

where

$$\hat{\beta} = \left(\sum_{i=1}^{n} x_i' \pi_i^{-1} x_i \right)^{-1} \sum_{i=1}^{n} x_i' \pi_i^{-1} y_i, \tag{2.2}$$

and we have assumed $\sum x_i' \pi_i^{-1} x_i$ to be nonsingular. This definition of the regression estimator follows Mickey (1959) who suggested restricting the term regression estimator to estimators that are location and scale invariant. The estimator (2.1) can also be written as

$$\bar{y}_r = \sum_{i=1}^{n} w_i y_i, \tag{2.3}$$

where

$$w_i = \bar{X}\left(\sum_{i=1}^{n} x_i' \pi_i^{-1} x_i \right)^{-1} x_i' \pi_i^{-1}, \tag{2.4}$$

and the weights have the property,

$$\sum_{i=1}^{n} w_i x_i = \bar{X}. \tag{2.5}$$

The weights of expression (2.4) are relatively easy to compute, and once computed, can be used for the estimation of any y-variable. If the vector x_j is replaced by the vector

$$(1, z_j) = (1, x_{j2} - \bar{X}_2, x_{j3} - \bar{X}_3, \ldots, x_{jk} - \bar{X}_k), \tag{2.6}$$

the estimator can be written in the form

$$\bar{y}_r = \bar{y}_\pi + (\bar{Z} - \bar{z}_\pi)\hat{\beta}_z = \bar{y}_\pi - \bar{z}_\pi \hat{\beta}, \tag{2.7}$$

where $\bar{Z} = 0$ is the population mean of z_j, $\bar{z}_\pi = \bar{x}_\pi - \bar{X}$,

$$(\bar{y}_\pi, \bar{z}_\pi) = \left(\sum_{i=1}^{n} \pi_i^{-1} \right)^{-1} \sum_{i=1}^{n} \pi_j^{-1} (y_i, z_i)$$

and

$$\hat{\beta}_z = \left[\sum_{j=1}^{n} (z_j - \bar{z}_\pi)' \pi_i^{-1} (z_j - \bar{z}_\pi) \right]^{-1} \sum_{j=1}^{n} (z_j - \bar{z}_\pi)' \pi_i^{-1} y_j .$$

In the form (2.7), $\bar{y}_r$ is the intercept in the regression of y on z. Thus, the theory given by Fuller (1975) for regression coefficients is applicable to the regression estimator of the mean. If the population total of units is known and denoted by N, the estimated population total is $N\bar{y}_r$.

By defining a sequence of populations and samples, it is possible to show that the estimator (2.1) is a consistent estimator of the mean of y. See, for example, Fuller (1975). The estimator of the variance of the regression estimator is a function of the joint probabilities. Consider a stratified two-stage sample and replace our single subscript i with the triple ℓjt. Then, omitting the finite correction term, a variance estimator is

$$\hat{V}\{\bar{y}_r\} = (n - k)^{-1} n \sum_{\ell=1}^{L} (n_\ell - 1)^{-1} n_\ell \sum_{j=1}^{n_\ell} (d_{\ell j.} - d_{\ell ..})^2, \tag{2.8}$$

where

$$d_{\ell j.} = \sum_{t=1}^{m_{\ell j}} w_{\ell jt} (y_{\ell jt} - x_{\ell jt} \hat{\beta}),$$

$$d_{\ell ..} = n_\ell^{-1} \sum_{j=1}^{n_\ell} d_{ij.},$$

n_ℓ is the number of sample primary sampling units in stratum ℓ, $m_{\ell j}$ is the number of sample elements in primary sampling unit j of stratum ℓ, $\hat{\beta}$ is the vector of coefficients defined in (2.2), n is the total number of elements in the sample, and $w_{\ell jt}$ is the weight for element t in primary sampling unit j of stratum ℓ. The factor $n - k$ is used by analogy to the divisor for the unbiased estimator of the error variance in ordinary regression. When the vector of control variables is coded as in (2.6), the estimator (2.8) is the estimated variance of the first element of $\hat{\beta}$, the estimated intercept. The estimator (2.8) was suggested in Hidiroglou, Fuller and Hickman (1976) and the consistency of the estimator was established by Fuller (1975). Also see Särndal, Swensson and Wretman (1989).

The estimators, constructed with weights (2.4), have good large sample properties. However, they may have undesirable behavior in small samples. Because the weights are linear functions of x_i, it is possible for some of the weights to be negative. Negative weights make it possible for estimates of positive parameters to be negative. Early research on methods of constructing nonnegative regression weights was conducted by Husain (1969). Huang (1978) designed a computer program to produce nonnegative regression weights. Huang and Fuller (1978) described the weight generation procedure and showed that the large sample distribution of the modified estimator is the same as that of the ordinary regression estimator. Also see Goebel (1976).

The computer algorithm of Huang (1978) is an iterative procedure based upon the ideas of generalized least squares. The goal of the Huang algorithm is a set of weights w_i, $i = 1, 2, \ldots, n$, satisfying (2.5) that do not differ greatly from the initial weights, where difference is a function of the initial weight. The Huang algorithm attempts to compute weights w_i satisfying

$$(1 - M) \max_{1 \le i \le n} w_i \pi_i^{-1} \le (1 + M) \min_{1 \le i \le n} w_i \pi_i^{-1},$$

where the parameter M, $0 < M \le 1$, is specified by the user and is generally chosen in the interval [0.8, 1.0]. If the first round regression weights defined by (2.4) do not satisfy the requirements, a second round of regression weights is computed. The second round weights are weighted regression weights in which a control factor is assigned to each observation. Small control factors are assigned to observations with large or small first round weights. Relatively large control factors are assigned to observations with first round weights close to π_i^{-1}. The second round regression weights are checked and if they fail to satisfy the criteria, the control factors are modified, and so on. The algorithm is given in the Appendix.

The control weighting used in the Huang algorithm has much in common with bounded-influence and robust regression methods. That is, in the final estimator, the contribution to the estimation of the slope vector is reduced for observations that are far from the mean. See Hampel (1978), Krasker (1980), and Mallows (1983). Recent research on this type of estimator for survey samples is that of Deville and Särndal (1992), Akkerboom, Sikkel, and van Herk (1991), Hulliger (1993) and Singh (1993).

It is not always possible to construct weights meeting the criteria and also satisfying (2.5). For example, if all of the observations on x_{t2} exceed the mean, there is no set of positive weights summing to one that also satisfy $\sum_{t=1}^{n} x_{t2} w_t = \bar{X}_2$. Therefore, the weight generation program will terminate if weights meeting the specified criteria cannot be constructed after a specified number of iterations.

In some situations it is desirable to restrict the weights to the nonnegative integers. This is true when estimates of totals are being constructed and the population contains well defined units, such as people. Nonnegative integer weights then provide more comfortable estimates, in that the estimates are physically attainable. Integer weights can be constructed so that no rounding is necessary when building tables. With such integer weights, all multiple way tables will automatically be internally consistent.

The Huang program contains an option to round the real weights to integer weights in a manner that maintains the sum of the weights. After rounding, the equalities (2.5) will generally no longer hold exactly. We have found that by iterating the Huang algorithm using the first-round integer weights as initial weights, integer weights can be constructed such that there is a very modest deviation from equality for expression (2.5). Cox (1987), Cox and Ernst (1982), and Fagan, Greenberg and Hemmig (1988) discuss rounding.

3. REGRESSION ESTIMATION WITH NONRESPONSE

The early theoretical developments for regression estimation assumed the sample to be a probability sample from the population. However, it has long been recognized that regression estimation can be used to reduce the bias that arises from imperfections in the data collection procedure. The most obvious of these imperfections is nonresponse. In all large samples of human subjects, some of the subjects fail to provide information. If the nonrespondents differ from the respondents, direct estimates constructed from the respondents will be biased. Given auxiliary information, regression estimation provides a method or reducing the bias. The degree to which the bias is reduced depends upon the relationship between the control variables, the variables of interest, and the response probabilities. See Little and Rubin (1987) for a general discussion of these issues.

Let π_i^* denote the inclusion probability equal to the product of π_i and the conditional probability of observing the unit given that the unit is selected. Then

$$E\left\{ \sum_{i=1}^{n} x_i' \pi_i^{-1} x_i \mid \xi_N \right\} = \sum_{i=1}^{N} x_i' \pi_i^{-1} \overset{*}{\pi}_i x_i \qquad (3.1)$$

and

$$E\left\{ \sum_{i=1}^{n} x_i' \pi_i^{-1} y_i \mid \xi_N \right\} = \sum_{i=1}^{N} x_i' \pi_i^{-1} \overset{*}{\pi}_i y_i, \qquad (3.2)$$

where the expectations are conditional on the given finite population ξ_N, and n is the realized sample size. In the case of nonresponse, the ratio $p_i = \overset{*}{\pi}_i \pi_i^{-1}$ is the response probability for individual i. Therefore, under conditions such as those used by Fuller (1975),

$$\operatorname*{plim}_{\substack{n \to \infty \\ N \to \infty}} (\hat{\beta} - \gamma) = 0, \qquad (3.3)$$

where $\hat{\beta}$ is defined in (2.2) and

$$\gamma = \left(\sum_{i=1}^{N} x_i' \pi_i^{-1} \overset{*}{\pi}_i x_i \right)^{-1} \sum_{i=1}^{N} x_i' \pi_i^{-1} \overset{*}{\pi}_i y_i. \qquad (3.4)$$

Then

$$\bar{Y} = \bar{X}\gamma + \bar{A}, \tag{3.5}$$

where $\bar{A} = N^{-1}\sum_{i=1}^{N} a_i$ and $a_i = y_i - x_i\gamma$. Thus, the regression estimator (2.1) will be a consistent estimator of $\bar{Y}$ if $\operatorname{plim}_{N\to\infty} \bar{A} = 0$. The probability limit of $\bar{A}$ will be zero if the finite population is a random sample from an infinite population in which the linear model

$$y_i = x_i\beta + e_i, \quad E\{e_i\} = 0$$

holds for all i.

The mean $\bar{A}$ is zero when $\overset{*}{\pi}_i = \pi_i$ for all i and an element of x_i is one for all i because then

$$\gamma = \beta = \left(\sum_{i=1}^{N} x_i'x_i\right)^{-1} \sum_{i=1}^{N} x_i'y_i \tag{3.6}$$

and $\sum_{i=1}^{N}(y_i - x_i\beta) = 0$. A sufficient condition for $\bar{A}$ to be zero is the existence of a row vector c such that

$$cx_i' = \overset{*}{\pi}_i^{-1}\pi_i = p_i^{-1}, \tag{3.7}$$

for $i = 1, 2, \ldots, N$. Thus, if the ratio of nominal probabilities to true probabilities is a linear function of the control variables, the regression estimator is a consistent estimator of the mean of y, where the limit is for sequences as defined in Fuller (1975). One way in which (3.7) can be satisfied is for the elements of x_i to be dummy variables that define subgroups and for the response probabilities to be constant in each subgroup. This situation is sometimes described by saying that elements are missing at random in each subgroup. We take the assumption that $\bar{A} = 0$ as our working assumption in the empirical analysis.

In any regression problem, it is impossible to use the sample to verify some of the assumptions. For example, in ordinary least squares regression, the residuals $\hat{e}_i = y_i - x_i\hat{\beta}$ are uncorrelated with x_i and, hence, the residuals cannot be used to check the assumption that the true errors are uncorrelated with x. Thus, in a survey with nonresponse, one searches for control variables that are correlated with y and (or) that one believes are correlated with the response probabilities. But one cannot guarantee that all bias has been removed by regression estimation based on a particular set of control variables.

In practice, one can often identify x-variables that are correlated with the probability of response and (or) correlated with the y-variables. For example, in the 1987-1988 Nationwide Food Consumption Survey, the response rate was low among high-income households. Therefore, use of variables for household income in a regression estimator is expected to reduce the bias in the estimated mean for characteristics that are correlated with income.

The error in $\hat{\beta}$ as an estimator of γ can be approximated by

$$\hat{\beta} - \gamma \doteq G^{-1}T^{-1}\sum_{i=1}^{n} x_i'\pi_i^{-1}a_i,$$

where a_i is defined in (3.5),

$$T = \sum_{i=1}^{N} \pi_i^{-1}\overset{*}{\pi}_i$$

and

$$G = T^{-1}\sum_{i=1}^{N} x_i'\pi_i^{-1}\overset{*}{\pi}_i x_i.$$

Under reasonable assumptions

$$\hat{T} = \sum_{i=1}^{n} \pi_i^{-1}$$

and

$$\hat{G} = \hat{T}^{-1}\sum_{i=1}^{n} x_i'\pi_i^{-1}x_i$$

are consistent estimators of T and G. Thus, the variance of the regression estimator can be estimated by estimating the variance of $\sum_{i=1}^{n} x_i'\pi_i^{-1}a_i$. If we assume that the conditional probabilities of response in one primary sampling unit are independent of those in all other primary sampling units and that at least one observation unit is observed in each selected primary sampling unit, then (2.8) remains an appropriate estimator of the variance of the regression estimated mean of y.

The estimator of variance (2.8) also remains appropriate if the regression weights are constructed by a procedure other than (2.4). For example, let the weights be defined by

$$w_{gi} = \bar{X}\left[\sum_{i=1}^{n} x_i'\pi_i^{-1}g_i x_i\right]^{-1} x_i'\pi_i^{-1}g_i,$$

where the g_i are functions of the x_i. Assume

$$\operatorname{plim}\hat{\beta}_g = \gamma_g,$$

where

$$\hat{\beta}_g = \left[\sum_{i=1}^{n} x_i'\pi_i g_i x_i\right]^{-1} \sum_{i=1}^{n} x_i'\pi_i^{-1}g_i y_i.$$

Also assume

$$\operatorname*{plim}_{N\to\infty} N^{-1}\sum_{i=1}^{N}(y_i - x_i\gamma_g) = 0.$$

Then expression (2.8) with w_{gi} replacing $w_{\ell jt}$ is a consistent estimator of the variance of the estimator. The estimator (2.8) will be used in our empirical analyses.

Formula (2.8) identifies the two effects of regression estimation on the variance of an estimated mean. The correlation effect reduces the variance of the estimated mean while the increase in the sum of squares of the weights increases the variance of the estimated mean. To understand these effects, consider a simple random sample. If the y variable is correlated with x, the correlation tends to reduce the variance of the regression estimator relative to that of the simple estimator because

$$E\{(y_i - x_i\beta)^2\} \leq E\{[y_i - E(y_i)]^2\}.$$

If the sample means of the control variables differ from the population means, then

$$\sum_{i=1}^{n} w_i^2 > n^{-1},$$

where n^{-1} is the sum of squares of the simple weights for a simple random sample.

When comparing the variance of the sample mean with the variance of the regression estimator, one should not forget that one of the reasons for using regression estimation in samples with nonresponse is to produce an estimator with less bias than that of the direct estimator. Thus, in the next section we compare an estimator of the mean square error of the simple estimator to an estimator of the variance of the regression estimator.

4. APPLICATION TO THE NATIONWIDE FOOD CONSUMPTION SURVEY

The 1987-1988 Nationwide Food Consumption Survey was conducted by the Human Nutrition Information Service of the U.S. Department of Agriculture. The original sample was a self-weighting stratified sample of area primary sampling units within the 48 conterminous states. Primary sampling units were divided into secondary units called area segments. Households within the sample segments were contacted by personal interview. The field operation was conducted during the period April 1987 through August 1988 by a contractor under contract to the Human Nutrition Information Service.

Approximately 37% of the housing units identified as occupied provided complete household food use information. The realized household sample contains 4,495 households. Because of the low response rate, the Human Nutrition Information Service decided to use regression weighting in the estimation. Population totals for all characteristics except urbanization were estimated by the Human Nutrition Information Service from the March 1987 Current Population Survey. See Bureau of the Census (1987). The population totals for urbanization classes were furnished by the contractor. In our analysis, we treat the estimated population totals as if they were known population totals.

Table 1

Sample and population characteristics of households

Characteristic	Category	Household Sample Frequency	Household Sample Percent	Population Percent
Season	Spring	1,828	40.7	25.0
	Summer	678	15.1	25.0
	Fall	717	16.0	25.0
	Winter	1,272	28.3	25.0
Region	Northeast	905	20.1	21.2
	Midwest	1,172	26.1	24.7
	South	1,567	34.9	34.4
	West	851	18.9	19.6
Urbanization	Central Cities	1,064	23.7	31.2
	Suburban	2,122	47.2	46.0
	Nonmetro	1,309	29.1	22.9
Household Income as % of Poverty	< 131%	1,041	23.2	20.0
	131-300%	1,564	34.8	32.2
	301-500%	1,108	24.6	25.9
	> 500%	782	17.4	21.8
Household Food Stamps	Yes	314	7.0	7.4
	No	4,181	93.0	92.6
Ownership of Domicile	Yes	2,998	66.7	64.1
	No	1,497	33.3	35.9
Race of Household Head	Black	519	11.5	11.1
	Nonblack	3,976	88.5	88.9
Age of Household Head	< 25	338	7.5	7.9
	25-39	1,588	35.3	36.1
	40-59	1,369	30.5	30.5
	60-69	660	14.7	13.0
	70+	540	12.0	12.6
Household Head Status	Both Male and Female	3,057	68.0	60.8
	Female Only	1,044	23.2	26.0
	Male Only	394	8.8	13.2
Female Head Worked	Yes	1,792	39.9	41.5
	No	2,703	60.1	58.5
Exactly One Adult in Household	Yes	1,211	26.9	29.7
	No	3,284	73.1	70.3
Exactly Two Adults in Household	Yes	2,616	58.2	54.2
	No	1,879	41.8	45.8
Presence of Child < 7 Years Old	Yes	1,009	22.4	20.1
	No	3,486	77.6	79.9
Presence of Child 7-17 Years Old	Yes	1,309	29.1	26.5
	No	3,186	70.9	73.5
Household Size	(Mean)		2.731	2.642
Household Size, Squared	(Mean)		9.546	9.125

Characteristics of the population and of the household sample are given in Table 1. The original sample was unbalanced with respect to time of interview with nearly 41% of the interviews in the spring quarter and about 16% of the interviews in each of the summer and fall quarters. Interviews for the spring and summer quarters were done in both 1987 and 1988.

The sample was also unbalanced with respect to urbanization. There was a lower fraction of central city households in the sample than in the population (24% versus 31%), and a higher fraction of nonmetropolitan households in the sample than in the population (29% versus 23%).

The fraction of high income households was smaller in the sample than in the population. The sample contained a higher fraction of households with both a male and female head than the population (68% versus 61%). A higher fraction of the sample than of the population consisted of households with children. The sample was only mildly unbalanced with respect to several other sociodemographic characteristics.

The characteristics listed in Table 1 are believed by the staff of the Human Nutrition Information Service to be related to food consumption behavior. Therefore, variables based on those characteristics were used in the regression weighting procedure. To implement the weight generation program, each of the categorical variables of Table 1 was converted to a set of indicator variables. For example, three variables were created for the characteristic, household income as a percent of poverty. These were

Z_{t1} = 1 if income <131% for t-th household
= 0 otherwise,

Z_{t2} = 1 if income is 131-300% for t-th household
= 0 otherwise,

Z_{t3} = 1 if income is 301-500% for t-th household
= 0 otherwise.

Using this procedure, 25 indicator variables were created. In addition, household size and the square of household size were used as continuous variables.

The twenty-seven variables were used to generate regression weights using Huang's program. The parameter M of the weight generation program was set equal to 0.9 in the computation. The weights were rounded to integers, where each integer weight is a weight in thousands. The sum of the weights is 88,942, which is the number of households in the population in thousands. The average weight is 19.787, the smallest weight is 6, and the largest weight is 47. Thus, the largest weight is 2.38 times the average weight. The sum of squares of the weights is 2,317,930. The average weight squared and multiplied by the sample size is 1,759,884. Thus, if a variable has zero multiple correlation with the 27 variables, the variance of an estimate computed with the weights will be about 1.32 times the variance of the simple unweighted estimator.

Figure 1 shows the regression weights computed by the Huang algorithm plotted against the ordinary least squares weights. Because there are 4,495 households, many points are hidden. Both weights are standardized by dividing by the average weight. Thus, the average for weights coded in this manner is one. Because there are 27 control variables used in the construction, the Huang weights tend to form a swarm of points about an S-shaped function of the original weights. If there were only one control variable, the points would fall on an S-shaped curve. The original weights for observations to the left of zero were negative.

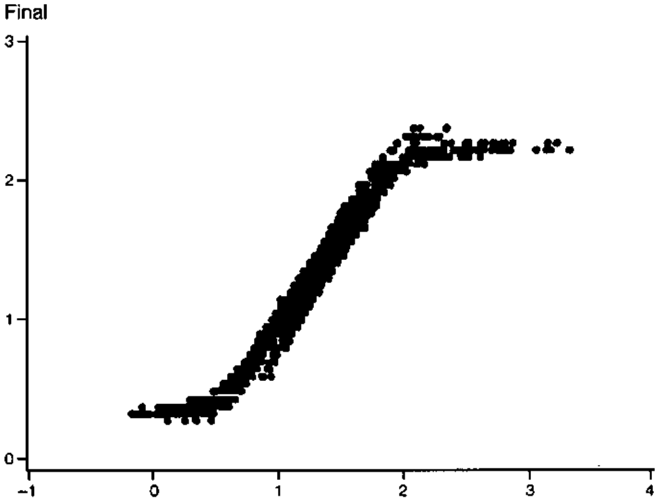


Figure 1. Plot of final weights against the ordinary least squares weights, both expressed relative to the average weight.

To compare estimates constructed with weights to unweighted estimates, we use the variables

Y_1 = adjusted total number of meals away from home (meals away),

Y_2 = total money value of food used at home (home food),

Y_3 = household size in 21-meal-equivalent persons (meal persons),

Y_4 = indicator to identify housekeeping households (housekeeping).

The household size in 21-meal-equivalent persons is the total adjusted meals eaten from household food supplies in the past 7 days divided by 21. "Meal persons" is the sum of two terms. The first term is the sum of the proportions of meals eaten at home in the interview week by each household member. The second term is the number of meals served to guests, boarders, and employees during the interview week, divided by 21. In other words:

$$\text{meal persons for } j\text{-th household} = \sum_i (h_{ij} + a_{ij})^{-1} h_{ij} + (21)^{-1} b_j ,$$

where h_{ij} = meals eaten at home by the i-th individual in the j-th household during the interview week, a_{ij} = meals eaten away from home by the i-th individual in the j-th household during the interview week, b_j = number of meals eaten by nonhousehold members in the j-th household during the interview week.

The adjusted total number of meals bought and eaten away from home is the sum of the proportions of meals eaten away from home in the interview week by household members, multiplied by 21. In the notation used to define meal persons,

$$\begin{array}{l}\text{meals away for}\\ j\text{-th household}\end{array} = 21 \sum_i (h_{ij} + a_{ij})^{-1} a_{ij}.$$

The total value of food used at home is the expenditures for purchased food plus the money value of home-produced food and food received free-of-cost that was used during the survey week. Expenditures for purchased food were based on prices reported as paid regardless of the time of purchase. Sales tax was excluded. Purchased food with unreported prices, food produced at home, food received as a gift, and food received instead of pay were valued at the average price per pound paid for comparable food by survey households in the same region and season.

A housekeeping household is a household with at least one person having ten or more adjusted meals from the household food supply during the seven days before the interview. Household food-use analyses generally consider only housekeeping households.

Table 2

Properties of alternative estimators

Variable	Un-weighted Mean	Weighted Mean	Differ-ence	Relative Efficiency of Regression
Meals away				
Housekeeping	7.75 (0.22)	7.93 (0.17)	−0.18 (0.09)	2.52
Nonhousekeeping	18.31 (1.14)	18.12 (1.19)	0.19 (0.68)	0.92
All	8.27 (0.22)	8.57 (0.22)	−0.30 (0.12)	2.56
Home food				
Housekeeping	61.10 (1.14)	59.56 (0.98)	1.54 (0.41)	3.65
Nonhousekeeping	25.99 (1.25)	26.39 (1.46)	−0.40 (1.00)	0.73
All	59.37 (1.12)	57.49 (0.91)	1.88 (0.39)	5.60
Meal persons				
Housekeeping	2.42 (0.03)	2.33 (0.01)	0.09 (0.01)	89.00
Nonhousekeeping	0.51 (0.03)	0.49 (0.03)	0.02 (0.02)	1.00
All	2.33 (0.03)	2.22 (0.01)	0.11 (0.01)	129.00
Housekeeping (%)	95.06 (0.40)	93.77 (0.58)	1.29 (0.10)	5.30

The means of the variables computed using unweighted data are given in Table 2 in the column headed, "Unweighted mean". Three means are given for meals away, home food, and meal persons. Two means are computed for the two subpopulations defined by the housekeeping variables. The third mean, designated by "all" in the table, is the estimated mean for the entire population. The standard errors of the estimates are given in parentheses below the estimates. The estimates and standard errors for the unweighted estimates were computed in PC CARP. See Fuller *et al.* (1986). The computations accounted for the fact that the sample is an area stratified cluster sample.

Because the sample is a two-stage sample, the estimated variances are larger than the variance of a simple random sample containing the same number of households. The ratio of the variance for a sample estimate to the variance of a simple random sample containing the same number of individuals is called the design effect. The estimated design effect is about 2.5 for meals away and meal persons, is about 4.1 for home food, and is about 1.5 for housekeeping for the "all" means for the unweighted sample.

The column headed "Weighted mean" contains the estimates computed with the regression weights. The standard errors were computed in PC CARP using formula (2.8) with the regression weights replacing the π_i^{-1}. The variance calculation requires computing a regression for every y-variable. The estimated means for the subpopulations are ratios of regression estimators. The variances for the subpopulation means were computed by calculating the variances of the Taylor deviates for the ratio using formula (2.8). The standard errors for unweighted and weighted estimates are similar for meals away and home food. However, the standard errors for the regression estimate of the population mean of meal persons is about one third of the standard error of the unweighted estimate. The standard error of the regression estimator is smaller because meal persons is highly correlated with the household size variables used as controls in the regression procedure.

The estimated squared multiple correlation between the variables of the table and the 27 control variables is 0.29, 0.44, 0.82, and 0.12 for meals away, home food, meal persons, and housekeeping, respectively. If the sample means of the control variables were nearly equal to the population means, the standard error of the regression estimate of meals away would be about $(1 - 0.29)^{1/2} = 0.84$ times the standard error of the unweighted estimate. In fact, the estimated standard error of the regression is about 0.97 times the standard error of the unweighted estimate. The difference is due to the fact that $\sum_{i=1}^{n} w_i^2$ is considerably bigger than n^{-1} because the sample is unbalanced on a number of items. Note that

$$0.97 \doteq [(0.71)(1.32)]^{1/2},$$

where $0.71 = (1 - 0.29)$ is one minus the squared correlation and $1.32 = n \sum_{i=1}^{n} w_i^2$. The situation for housekeeping is more extreme. The correlation is not large, and, apparently, the large deviations from the regression line are associated with large weights. The estimated variance for the regression estimator is about twice the estimated variance of the unweighted estimator.

Table 2 also contains the estimated differences between the unweighted and weighted estimators. The difference between the unweighted and the weighted estimated total is

$$\sum_{t=1}^{n} Nn^{-1}y_t - \sum_{t=1}^{n} w_t y_t = \sum_{t=1}^{n} (n^{-1}N - w_t)y_t .$$

The difference between the estimated means is the difference between the totals divided by the population size. To compute the variance of the difference between the means, we note that the hypothesis of a zero difference is equivalent to the hypothesis that the correlation between w_t and y_t is zero. Therefore, we computed the unweighted regression of y_t on w_t and computed the variance of the regression coefficient under the design using PC CARP. The standard errors for the difference in Table 2 are such that the "t-statistic" for the hypothesis of zero difference is equal to the "t-statistic" for the coefficient of w_t in the regression of y_t on w_t.

For all four characteristics, the difference between the weighted and unweighted estimators of the population mean is significant at traditional levels. Thus, under the assumption that the regression estimators are unbiased, there are significant biases in the unweighted estimators.

The bias picture is mixed for the estimates of the subpopulation means. The difference between the two estimators is significant for the three means for the housekeeping subpopulation, which is the population of interest. The difference is nonsignificant for the three means for the nonhousekeeping subpopulation. The sample contains only 222 nonhousekeeping households. Therefore, the variance of the difference between the weighted and unweighted estimates is much larger for the nonhousekeeping households than for the housekeeping households.

The differences between the two estimates of the population means are a function of the differences between the two estimates of the subpopulation means and the two estimates of the fraction of households in the two categories. This explains why the difference for "all" can be larger than both the "housekeeping" and "nonhousekeeping" differences.

The last column of Table 2 contains the ratio of the estimated mean square error of the unweighted estimator to the variance of the regression estimator. The estimated mean square errors for the unweighted estimators were computed as

$$\widehat{\mathrm{MSE}}_u = \hat{V} + \max\{0, (\mathrm{Diff})^2 - (\mathrm{s.e.\ diff})^2\},$$

where $\hat{V}$ is the estimated variance of the unweighted estimate, Diff is the difference between the two estimates from Table 2, and s.e. diff is the standard error of the difference from Table 2. The estimated variance $\hat{V}$ for the unweighted estimator is variance formula (2.8) with constant w_{tji}, and with $x_{tji}\hat{\beta}$ replaced by $\bar{y}_{t..}$. The second term of the estimated mean square error is the estimated squared bias. Under the assumption that the regression estimator is unbiased, the expected value of (Diff)^2 is the squared bias plus the variance of the difference. Hence, under the assumption that the regression estimator is unbiased, the estimated mean square error of the unweighted estimator is a consistent estimator. The estimated mean square errors of the weighted estimators are the variances of the weighted estimators computed as the squares of the standard errors of Table 2.

Of the four characteristics for which the population mean was estimated, the estimated relative efficiency of the regression estimator to the simple mean ranges from 2.5 to 129. The regression estimator for meals away has the smallest estimated efficiency. The variances of the two estimators are similar, but because of the estimated bias, the regression estimate for meals away is estimated to have a mean square error that is about 40% of that of the unweighted estimate. The mean square error of the regression estimate for home food is less than 20% of that of the unweighted estimate, that for meal persons is about 1% of that of the unweighted estimate, and that for housekeeping is about 20% that of the unweighted estimator. In all cases, the squared bias is a very important component of the estimated mean square error.

Because the unweighted subpopulation estimates for the nonhousekeeping households showed little bias, the unweighted estimates are estimated to be somewhat more efficient than the regression estimates. The nonhousekeeping subpopulation is only about 6% of the population and the deviations from the subpopulation mean show little correlation with the control variables. On the other hand, the regression estimates for the housekeeping subpopulation are estimated to be much more efficient than the unweighted estimates. The relative efficiencies for the housekeeping subpopulation are close to those of the total population estimates.

Even after allowing for the fact that the population totals from the Current Population Survey are not known population totals, it is clear that large gains are associated with regression estimation for the population means. Although the regression estimator for the means of the small subpopulation is estimated to be less efficient than the unweighted estimators, the loss in efficiency is small relative to the large gains in efficiency estimated for the other variables.

ACKNOWLEDGEMENTS

This research was partly supported by Research Support Agreement 58-3198-9-032 with the Human Nutrition Information Service, U.S. Department of Agriculture. We thank Phil Kott, Patricia Guenther, and the referees for useful comments.

APPENDIX
WEIGHT GENERATION PROGRAM

In this appendix, we present the regression weight generation procedure of Huang and Fuller (1978). The procedure we describe contains the option of specifying maximum and minimum weights. This option was not part of the original program. For a discussion of related weight generation procedures, see Singh (1993).

Suppose that the population means $(\bar{X}_1, \bar{X}_2, \ldots, \bar{X}_k)$ of the k auxiliary variables $(X_1, X_2, \ldots, X_k)$ are known. Let a sample of n observations be available and let

$$X = \begin{pmatrix} X_{11} & X_{12} & \ldots & X_{1k} \\ X_{21} & X_{22} & \ldots & X_{2k} \\ \vdots & \vdots & & \vdots \\ X_{n1} & X_{n2} & \ldots & X_{nk} \end{pmatrix}, \qquad \text{(A.1)}$$

where X_{ij} is the observation on variable j for individual i.

In addition to the array of sample observations and the populations means, two initial factors v_i and $g_i^{(0)}$, $i = 1, 2, \ldots, n$, are required to initiate the computations. The v_i are typically inversely proportional to the probabilities of selection. The default values for $g_i^{(0)}$ are $g_i^{(0)} = 1$. For stratified samples or data with unequal variances, the user may choose other values for $g_i^{(0)}$. (See Huang 1978 or Goebel 1976.) The program input includes the sample size n, the population size N, the parameter M, the maximum number of iterations LI, the upper bound of the ratios of weights to the average weight $\mathrm{U_B}$, and the lower bound of the ratios of weights to the average weight $\mathrm{L_B}$. It is required that $0 \le \mathrm{L_B} < 1 < \mathrm{U_B}$. In our description, we assume $\sum_{i=1}^{n} v_i = n$. The program normalizes the v_i so that the sum is n.

The program can be used to construct weights to estimate means or to estimate totals. The weights for totals are the weights for the means multiplied by N. For means, the program attempts to construct weights w_i such that

$$\sum_{i=1}^{n} w_i(1, X_i) = (1, \bar{X}), \qquad \text{(A.2)}$$

$$\mathrm{L_B} < n w_i < \mathrm{U_B}, \qquad \text{(A.3)}$$

$$(1 - M) \max_{1 \le i \le n} w_i v_i \le (1 + M) \min_{1 \le i \le n} w_i v_i, \qquad \text{(A.4)}$$

for $i = 1, 2, \ldots, n$.

The program is iterative, where an iteration consists of computing the generalized least squares weights, where a control factor h_i is applied to each observation. The h_i is a product of v_i and g_i, where g_i for iterations after the first is a "bell" shaped function of the distance (in a suitable metric) that the observation is from the population mean. At each iteration, the weights satisfy (A.2) but may fail (A.3) or (A.4).

It will not always be possible to construct weights satisfying the specified restrictions in the specified number of iterations. If the sample is such that the restriction cannot be met, the program outputs the weights of the last iteration. In the single x case, when the criterion cannot be satisfied, there will be two weights, one for those greater than the population mean, and one for those less than the population mean.

To describe the algorithm, let

$$Z_{ij} = X_{ij} - \bar{X}_j,$$

$$Z = \begin{pmatrix} Z_{11} & Z_{12} & \ldots & Z_{1p} \\ \vdots & \vdots & & \vdots \\ Z_{n1} & Z_{n2} & \ldots & Z_{np} \end{pmatrix},$$

$$V = \mathrm{diag}(v_1, v_2, \ldots, v_n),$$

$$J_n = (1, 1, \ldots, 1)',$$

$$A^{(0)} = Z' H^{(0)} Z,$$

$$G^{(0)} = \mathrm{diag}(g_1^{(0)}, \ldots, g_n^{(0)})$$

and

$$H^{(0)} = V G^{(0)}.$$

The algorithm consists of iterating three steps.

1. The initial calculation is for $\alpha = 0$. At iteration α, the vector of regression weights, denoted by $w^{(\alpha)}$, is

$$\begin{aligned} w^{(\alpha)} &= [1 + n\bar{u}_v^{(\alpha)}]^{-1} V(n^{-1} J_n + u^{(\alpha)}) \\ &= (w_1^{(\alpha)}, \ldots, w_n^{(\alpha)})', \qquad \text{(A.5)} \end{aligned}$$

where

$$u^{(\alpha)} = G^{(\alpha)} Z (A^{(\alpha)})^{\dagger} (\bar{X} - \bar{x}_v) = (u_1^{(\alpha)}, \ldots, u_n^{(\alpha)})',$$

$$\bar{x}_v = \left(\sum_{i=1}^{n} v_i \right)^{-1} \sum_{i=1}^{n} v_i x_i,$$

$(A^{(\alpha)})^{\dagger}$ is a symmetric generalized inverse of $A^{(0)}$,

$$n\bar{u}_v^{(\alpha)} = \max\{ J_n' V u^{(\alpha)}, n^{-1} - 1 \}, \qquad \text{(A.6)}$$

and

$$A^{(\alpha)} = Z' H^{(\alpha)} Z.$$

2. The weights of Step 1 are checked to see if they satisfy the criteria.

   (a) Is $| nu_i^{(\alpha)} | \leq M$ for all i?

   (b) Is
   $$\mathrm{L_B} \leq nw_i^{(\alpha)} \leq \mathrm{U_B}$$
   for all i?

   If either (a) or (b) fails for any i and LI iterations have not been completed, go to Step 3. If (a) and (b) are satisfied, or if LI iterations have been completed, the weights are output.

3. The control factors $h_i^{(\alpha)}$, $i = 1, 2, \ldots, n$, are modified. Set
   $$H^{(\alpha)} = H^{(\alpha-1)}G^{(\alpha)},$$
   where

$$G^{(\alpha)} = \mathrm{diag}(g_1^{(\alpha)}, g_2^{(\alpha)}, \ldots, g_n^{(\alpha)}),$$

$$\begin{aligned} g_i^{(\alpha)} &= 1 && 0 \leq d_i^{(\alpha)} < 0.5 \\ &= 1 - 0.8(d_i^{(\alpha)} - 0.5)^2 && 0.5 \leq d_i^{(\alpha)} \leq 1 \\ &= 0.8(d_i^{(\alpha)})^{-1} && d_i^{(\alpha)} > 1, \end{aligned}$$

$$d_i^{(\alpha)} = 1.33[D_i^{(\alpha-1)}]^{-1}n\,| u_i^{(\alpha-1)} |,$$

$$\begin{aligned} D_i^{(\alpha-1)} &= \min\{M, C_{\mathrm{Li}}^{(\alpha-1)}\} && \text{if } w_i^{(\alpha-1)} < v_i \\ &= \min\{M, C_{\mathrm{Bi}}^{(\alpha-1)}\} && \text{if } w_i^{(\alpha-1)} \geq v_i, \end{aligned}$$

$$C_{\mathrm{Li}}^{(\alpha-1)} = \max\{| v_i^{-1}(1 + n\bar{u}_v^{(\alpha-1)})\mathrm{L_B} - 1 |, 0.1\,M\},$$

$$C_{\mathrm{Bi}}^{(\alpha-1)} = \max\{| v_i^{-1}(1 + n\bar{u}_v^{(\alpha-1)})\mathrm{U_B} - 1 |, 0.1\,M\}.$$

Go to Step 1 to compute new regression weights.

The constant 1.33 in the definition of $d_i^{(\alpha)}$ and the constant of 0.8 in the definition of $g_i^{(\alpha)}$ were chosen to speed convergence. The control factors $g_i^{(\alpha)}$ are chosen to downweight observations on the basis of a distance from the population mean.

The definition of $w^{(\alpha)}$ in (A.5) is an alternative way of writing the vector of generalized least squares weights of (2.4) when $\pi_i^{-1} = h_i^{(\alpha)}$.

REFERENCES

AKKERBOOM, J.C., SIKKEL, D., and van HERK, H. (1991). Robust weighting of financial survey data. Contributed paper presented at meeting of the International Statistical Institute, Cairo, Egypt.

ALEXANDER, C.H. (1987). A model based justification for survey weights. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 183-188.

BETHLEHEM, J.G. (1988). Reduction of nonresponse bias through regression estimation. *Journal of Official Statistics*, 4, 251-260.

BETHLEHEM, J.G., and KELLER, W.A. (1987). Linear weighting of sample survey data. *Journal of Official Statistics*, 3, 141-153.

BRACKSTONE, G.J., and RAO, J.N.K. (1979). An investigation of raking ratio estimators. *Sankhyā*, Series C, 97-114.

BUREAU OF THE CENSUS (1987). Current Population Survey, March 1987: Technical Documentation. Washington, D.C.

COCHRAN, W.G. (1942). Sampling theory when the units are of unequal sizes. *Journal of the American Statistical Association*, 37, 199-212.

COCHRAN, W.G. (1977). *Sampling Techniques*, 3rd Ed. New York: John Wiley.

COPELAND, K.R., PEITZMEIER, F.K., and HOY, C.E. (1987). An alternative method of controlling current population survey estimates of population counts. *Survey Methodology*, 13, 173-182.

COX, L.H. (1987). A constructive procedure for unbiased controlled rounding. *Journal of the American Statistical Association*, 82, 520-524.

COX, L.H., and ERNST, L.R. (1982). Controlled rounding. *INFOR*, 20, 423-432.

DEVILLE, J.-C., and SÄRNDAL, C.-E. (1992). Calibration estimators in survey sampling. *Journal of the American Statistical Association*, 87, 376-382.

DEMING, W.E., and STEPHAN, F.F. (1940). On a least squares adjustment of a sampled frequency table when the expected marginal totals are known. *Annals of Mathematical Statistics*, 11, 427-444.

DARROCH, J.N., and RATCLIFF, D. (1972). Generalized iterative sealing for log-linear models. *Annals of Mathematical Statistics*, 43, 1470-1480.

EL-BADRY, M.A., and STEPHAN, F.F. (1985). On adjusting sample tabulations to census counts. *Journal of the American Statistical Association*, 50, 738-762.

FAGAN, J.T., GREENBERG, B.V., and HEMMIG, B. (1988). Controlled rounding of three dimensional tables. Statistical Research Division Report Census/SRD/RR-88/02. U.S. Bureau of the Census, Washington, D.C.

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

FULLER, W.A., KENNEDY, W., SCHNELL, D., SULLIVAN, G., and PARK, H.J. (1986). PC CARP. Statistical Laboratory, Iowa State University, Ames Iowa.

GOEBEL, J.J. (1976). Application of an iterative regression technique to a national potential cropland survey. *Proceedings of the Social Statistics Section, American Statistical Association*, 350-353.

HAMPEL, F.R. (1978). Optimally bounding the gross-error-sensitivity and the influence of position in factor space. *Proceedings of the Statistical Computing Section, American Statistical Association*, 59-64.

HIDIROGLOU, M.A. (1974). Estimation of regression parameters for finite populations. Unpublished Ph.D. thesis, Iowa State University, Ames, Iowa.

HIDIROGLOU, M.A., FULLER, W.A., and HICKMAN, R.D. (1976). SUPER CARP, Statistical Laboratory, Iowa State University, Ames, Iowa.

HUANG, E.T. (1978). Nonnegative regression estimation for sample survey data. Unpublished Ph. D. thesis. Iowa State University, Ames, Iowa.

HUANG, E.T., and FULLER, W.A. (1978). Nonnegative regression estimation for survey data. *Proceedings of the Social Statistics Section, American Statistical Association 1978*. 300-303.

HULLIGER, B. (1993). Robustification of the Horvitz-Thompson estimator. Contributed paper 49th Session of the International Statistical Institute. Book 1, 583-584.

HUSAIN, M. (1969). Construction of regression weights for estimation in sample surveys. Unpublished M.S. thesis, Iowa State University, Ames, Iowa.

IRELAND, C.T., and KULLBACK, S. (1968). Contingency tables with given marginals. *Biometrika*, 55, 169-188.

JESSEN, R.J. (1942). Statistical investigation of a sample survey for obtaining farm facts. Iowa Experiment Station Research Bulletin, 304.

KRASKER, W.A. (1980). Estimation in linear regression models with disparate data points. *Econometrica*, 48, 1333-1346.

LEMAÎTRE, G., and DUFOUR, J. (1987). An integrated method for weighting persons and families. *Survey Methodology*, 13, 199-207.

LITTLE, R.J.A., and RUBIN, D.B. (1987). *Statistical Analysis with Missing Data*. New York: John Wiley.

LUERY, D. (1986). Weighting survey data under linear constraints on the weights. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 325-330.

MALLOWS, C.L. (1983). Discussion of Huber: Mimimax aspects of bounded-influence regression. *Journal of the American Statistical Association*, 78, 77.

MICKEY, M.R. (1959). Some finite population unbiased ratio and regression estimators. *Journal of the American Statistical Association*, 54, 594-612.

OH, H.L, and SCHEUREN, F. (1987). Modified raking ratio estimation. *Survey Methodology*, 13, 209-219.

RAO, J.N.K. (1992). Estimating totals and distribution functions using auxiliary information at the estimation stage. Paper presented at the Workshop on Users of Auxiliary Information in Surveys, Örebo, Sweden, October, 1992.

ROYALL, R.M., and CUMBERLAND, W.G. (1981). The finite-population linear regression estimator and estimators of its variance - an empirical study. *Journal of the American Statistical Association*, 76, 924-930.

SÄRNDAL, C.-E., SWENSSON, B., and WRETMAN, J.H. (1989). The weighted residual technique for estimating the variance of the general regression estimator of the finite population total. *Biometrika*, 76, 527-537.

SAS INSTITUTE INC. (1989). SAS/STAT User's Guide, Version 6 Fourth Edition, Volume 1. Cary, NC: SAS Institute Inc.

SINGH, A.C. (1993). On weight adjustment in survey sampling. Unpublished manuscript. Statistics Canada, Ottawa, Canada.

STEPHAN, F.F. (1942). An iterative method of adjusting sample frequency tables when expected marginal totals are known. *Annals of Mathematical Statistics*, 13, 166-178.

WATSON, D.J. (1937). The estimation of leaf areas. *Journal of Agricultural Science*, 27, 474-483.

WRIGHT, R.L. (1983). Finite population sampling with multivariate auxiliary information. *Journal of the American Statistical Association*, 78, 879-884.

ZIESCHANG, K.D. (1990). Sample weighting methods and estimation of totals in the consumer expenditure survey. *Journal of the American Statistical Association*, 85, 986-1001.

Survey Methodology, June 1994
Vol. 20, No. 1, pp. 87–94
Statistics Canada

Estimating the Rate of Rural Homelessness: A Study of Nonurban Ohio

ELIZABETH A. STASNY, BEVERLY G. TOOMEY and RICHARD J. FIRST[1]

ABSTRACT

Recently, much effort has been directed towards counting and characterizing the homeless. Most of this work, however, has focused on homeless persons in urban areas. In this paper, we describe efforts to estimate the rate of homelessness in nonurban counties in Ohio. The methods for locating homeless persons and even the definition of homelessness are different in rural areas where there are fewer institutions for sheltering and feeding the homeless. There may also be a problem with using standard survey sampling estimators, which typically require large population sizes, large sample sizes, and small sampling fractions. We describe a survey of homeless persons in nonurban Ohio and present a simulation study to assess the usefulness of standard estimators for a population proportion from a stratified cluster sample.

KEY WORDS: Biased estimator; Regression estimator; Small sample size; Stratified cluster sample; Simulation.

1. INTRODUCTION

When we think of the homeless, we often think of "street people" and "bag ladies". We picture people sleeping on park benches, on heating grates, and in homeless shelters. These stereotypes of the homeless originated in large cities, however, and do not necessarily provide an accurate picture of homeless persons in rural areas.

Many of the studies of homeless persons have been carried out in larger cities. For example, the 1987 Urban Institute Study counted homeless persons in 20 major cities in the U.S. Another major study by Rossi was carried out in Chicago. (See Burt and Taeuber (1991) for an overview of survey methods for these and other studies that counted homeless populations.)

During the 1990 United States Population Census, a special attempt was made to include homeless persons in the population count through the S-Night (Shelter and Street Night) count. For this effort, a special national list of shelters and locations in which homeless persons sleep was compiled. The highest elected official of over 39,000 rural and urban local governments was asked to provide a list of shelters, street locations, and open public locations where the homeless stay at night. The homeless were counted by Census enumerators during a single night, March 20. Note that the main goal of S-Night was to include homeless persons in the Census count; relatively little information on characteristics of the homeless is available in the Census data. Details on the S-Night procedures are provided by Taeuber and Siegel (1990).

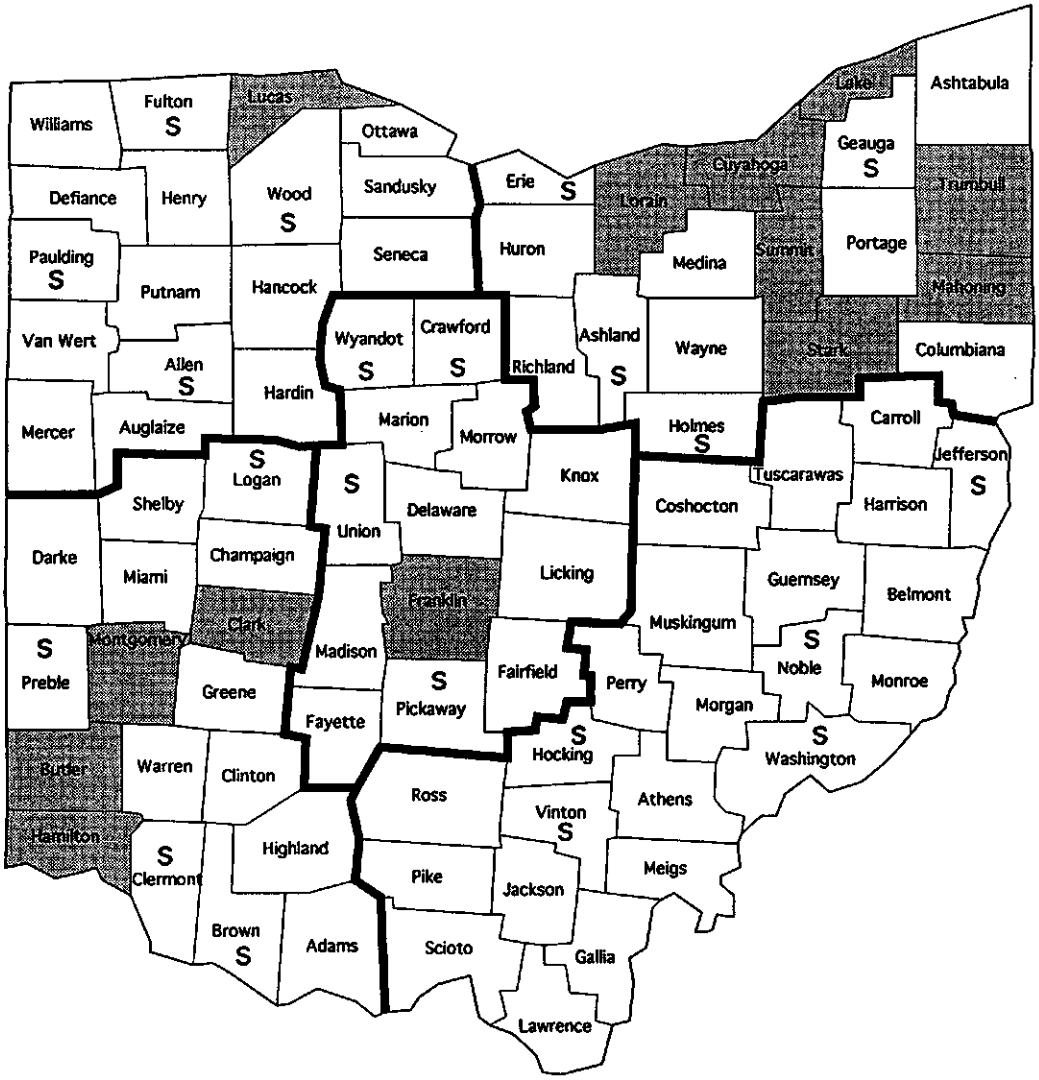
In contrast tp surveys of homeless persons in urban areas and to the Census S-Night, the goal of the survey described here was to locate and count the nonurban homeless wherever they might be, and to collect information to describe these homeless persons. In Section 2 of this paper, we describe the design of the 1990 survey of rural homeless persons in Ohio. We present our definition of rural homelessness and we describe the methods used to locate and survey the homeless. In Section 3 we present our estimates of the rates of rural homelessness obtained using the standard estimator of a proportion from a stratified cluster sample. Since these estimates are likely to be biased, we also present the results of a simulation study conducted to assess the likely size of the bias. In Section 4 we consider a regression estimator for the rate of homelessness and compare the regression estimator to the standard estimator of Section 3. In Section 5 we present our conclusions.

2. THE SURVEY

There are 88 counties in Ohio. Of these, 13 are urban counties with large cities and 75 are defined as rural or nonurban. These 75 counties of interest include counties that are completely rural, counties that are not adjacent to urban counties and that have moderately populated county seats, and suburban counties that border counties with large metropolitan areas.

The design used in this 1990 survey was selected to facilitate comparisons with a 1984 study of Ohio rural homeless persons (Roth *et al.* 1985). In the earlier study, Ohio's counties were divided into five regions, northeast, northwest, central, southeast, and southwest, and a stratified random sample of 16 rural counties was selected. The 21 counties selected for the 1990 study included the

[1] Elizabeth A. Stasny, Department of Statistics; Beverly G. Toomey and Richard J. First, College of Social Work, The Ohio State University, Columbus, Ohio 43210.



Note: Shaded counties are urban counties that were excluded from the study. An "S" indicates a county in the sample. The heavy boundaries divide the state into the five geographical strata: northeast, northwest, central, southeast, and southwest.

Figure 1. County Map of Ohio

16 counties from the original study and one additional county selected at random from within each region. (We should note that analysis of data from the present study suggests that stratification of Ohio into the five regions is not useful for improving the estimate of rural homelessness.) A map of Ohio showing the five regions, the urban counties, and the sampled counties is provided in Figure 1.

The following is a brief description of the 1990 survey methodology. More detailed descriptions are given by First *et al.* (1994) and Toomey *et al.* (1993).

2.1 Survey Personnel

A census of all homeless persons within the 21 sampled counties was attempted. Because there are not typically homeless shelters or other gathering places for the homeless in nonurban areas, the survey was conducted over a six-month period and made use of a network of advisors to locate the homeless. The survey period began with the first full week of February 1990. Homeless persons were identified and located by a referral network within each sampled county. Each network was supervised by a local county coordinator. The principal investigators supervised the county coordinators and the central office staff. They monitored the data collection, through bi-weekly phone calls and field visits, to assure uniformity and to control quality.

Advisors and interviewers, selected for their knowledge of the counties in which they worked, identified people who met the criteria for homelessness. Advisors included church leaders, hospital staff, civic club leaders, elected community officials, informal community leaders, bartenders, hotel clerks, laundromat attendants, and professional service providers such as health department staff, librarians, agricultural extension agents, postal workers, ministers, park rangers, neighborhood action groups, human service case workers, mental health workers, and law enforcement officers. One hundred interviewers conducted the interviews with the homeless. Interviewers attended a four-hour training session and were provided with a training manual of field guidelines. Interviews took place in offices, diners, motel rooms, cars, state parks, barns, laundromats, bars, and under railroad trestles. Interviewers were trained to know about available community resources and to make referrals for respondents who wanted services. In addition, interviewers had access to funds to offer a meal or minor assistance if necessary (less than $600 was spent on such purchases). Assistance provided through interviewers was limited so that people would not have an incentive to falsely identify themselves as homeless.

2.2 Definition of Rural Homeless

Screening questions were used to identify homeless persons. The definition of homelessness used in this study was necessarily somewhat different from the definition used for studies in urban areas. In rural areas there are fewer public shelters and housing alternatives specifically for the homeless. Respondents were counted as homeless if they did not have a permanent residence they considered home and if, on the previous night, they had slept in (1) limited or no shelter, (2) shelters or missions that serve homeless persons, (3) cheap hotels or motels when the actual stay or intent to stay was 45 days or less, or (4) other unique situations when the actual stay or intent to stay was 45 days or less. Included in the fourth category were people who stayed in sheds, barns, old buses, and old trailers without water or power, provided the person did not own the property and was not paying rent to stay there. Also included as homeless were people who were temporarily staying with friends or relatives, had not been staying in that household more than 45 days, were not a part of the household, and were planning on moving out in 45 days or less. Persons who were staying in battered women's shelters, hospitals, prisons, migrant workers camps, *etc.* were not counted as homeless unless they were leaving the facility and had nowhere to go.

Our definition of homelessness may be contrasted with that used in studies of homeless persons in urban areas. The common criteria of the definition of homelessness for such studies is based on the Stewart B. McKinney Homeless Assistance Act (1987). The Act defines a homeless person as "an individual who lacks a fixed, regular, and adequate nighttime residence and an individual who has a primary nighttime residence that is (a) a supervised publicly or privately operated shelter designed to provide temporary living accommodations (including welfare hotels, congregate shelters, and transitional housing for the mentally ill); (b) an institution that provides a temporary residence for individuals intended to be institutionalized; or (c) a public or private place not designed for, or ordinarily used as, a regular sleeping accommodation for human beings." From this definition comes the notion of "literally homeless" as suggested by Rossi *et al.* (1987). These standard definitions do not include, for example, those homeless persons who double up with family or friends. We did include such persons in our count of the rural homeless. Our analysis indicates that about a third of the persons counted in our census would not be counted under the urban definition of homelessness. It is not known how much counting those doubling up would increase estimates in urban areas.

2.3 The Interview Period

The use of a six-month survey period for counting the rural homeless is different from the typical one-day survey period used most often in surveys conducted in urban areas. In a review of seven studies of the homeless, Burt and Taeuber (1991) report that these studies used single nights, or one or two weeks as the interview period at a single location. Most of these studies relied on locating the homeless in shelters, soup kitchens, abandoned buildings, or similar locations. Since the homeless in rural areas are less likely to have shelters or soup kitchens available to them, they are harder to find and a longer survey period is recommended.

To facilitate comparisons with single-day or single-week surveys, homeless persons found in this study were asked how long they had been homeless. Using this information we were able to determine the number of persons in the sampled counties who were homeless during the first week of the survey, the first full week of February 1990.

In Section 3 we present estimates of the homeless rate for both the six-month period and the single week. The six-month rate includes anyone who met the definition of homelessness at any time during the six-month interview period. The one-week rate includes those interviewed throughout the six months who reported being homeless during the first full week of February.

To avoid duplication of respondents over the six-month period, each subject was assigned a unique identification number which included the subject's birth date, gender, and first three letters of the last name. Only a single duplicate interview was found in the data base; it was removed from the data base. (We do not have information on duplicates found in the field.) Because of this control for duplicate counting, we feel that any bias in our data collection procedures would be in the direction of an undercount of the rural homeless.

During the six-month interviewing period, 1,100 adults and 480 accompanying children were identified as homeless in the 21 sampled counties.

2.4 The Survey Questionnaire

If the responses to the screening questions indicated that a person was homeless, that subject was asked to respond to a questionnaire designed to obtain information about the person and his or her life experiences. Of the 1,100 adults identified as homeless, 919 completed the full interview. Although the focus of this paper is on estimating the number of rural homeless, we will describe briefly the questionnaire used to collect information to characterize the homeless.

The full questionnaire contained three sections. The first included questions on demographics and life experiences (for example, reasons for being homeless, use of mental health and other human services, employment history, drug and alcohol usage, family structure, and general well-being). The second section contained ten scales (including, for example, depression-anxiety, disorientation-memory impairment, and retardation-lack of emotion) from the Psychiatric Status Schedule developed by Spitzer, *et al.* (1970). The final section was an interview post-mortem which was completed by the interviewer and included information on where the interview occurred, respondent characteristics (for example, gender and unusual behaviors), and an assessment of the accuracy of the respondent's answers. The findings from this portion of the study are summarized by First *et al.* (1994).

3. THE ESTIMATES OF RATE OF HOMELESSNESS

3.1 The Estimator

The regional estimate of the rate of rural homelessness was obtained using the standard estimator for a proportion from a stratified cluster sample with unequal cluster sizes. In this case, the cluster is the county, the cluster size is the population within the county, and the strata size is the population within a region. The estimator is as follows:

For the i-th region, the estimated rate of homelessness is r_i where

$$r_i = \frac{\begin{array}{c}\text{number of homeless in sampled rural counties}\\ \text{in the } i\text{-th region}\end{array}}{\begin{array}{c}\text{total population in sampled rural counties}\\ \text{in } i\text{-th region}\end{array}} .$$

Then the estimated homeless rate for the state is

$$r_{\text{state}} = \frac{\sum_i [r_i \times \text{rural county population in } i\text{-th region}]}{\text{total rural county population in Ohio}},$$

where the summation is over the five geographical regions shown in Figure 1. The population totals for the 75 nonurban counties were obtained from 1990 Census data.

The estimated one-week and six-month rates of homelessness, given as number of homeless persons per 10,000 population, are shown in Table 1.

Because the above estimator involves the ratio of two random variables, the number of homeless and the population size for sampled clusters, the estimator is biased (see, for example, Cochran 1977). The bias decreases as sample size (number of counties sampled in this case) increases. Since our sample size is small, we recognize that our estimates are likely to be biased. On the other hand, our sampling fraction is large because the number of rural counties is small. Hence, we wish to assess the likely amount of bias in our estimates. (Note that the small sample sizes could also make the standard errors given in Table 1 inaccurate.)

Table 1

Estimated Rates of Homelessness per 10,000 in Rural Ohio

Area	One-Week Rate (February 4 - February 10, 1990)	Six-Month Rate (February - July 1990)
State	5.68 (0.99)	14.00 (2.09)
Northeast	3.44 (0.79)	12.00 (2.19)
Northwest	5.21 (3.51)	12.77 (5.18)
Central	5.85 (1.86)	12.11 (3.05)
Southeast	6.89 (1.93)	15.90 (5.91)
Southwest	7.25 (2.44)	16.78 (5.32)

Note: Standard errors are given in parentheses after each estimate.

3.2 The Simulation Study

We conducted a simulation study to help us assess the likely amount of bias in our estimates. We first generated five "populations" each with counts of the homeless for all 75 nonurban counties in Ohio. For all five simulated populations, the observed numbers of homeless persons for the 21 sampled counties were used as the counts in those counties. Counts for the remaining 54 counties were generated randomly as described below. Note that the simulated counts represent the six-month counts of the homeless.

The first simulated population was created by generating the natural log of the rate of homelessness from a single normal distribution. The log of the rate was used because the observed rates for the 21 sampled counties have a highly skewed histogram but the histogram for the log of the rates is approximately mound shaped. The mean of the observed log rates is 2.465 with a standard deviation of 0.7154. Thus, the generated log rates of homelessness were randomly sampled (using the statistical package S) from a normal distribution with this mean and standard deviation. After the log rates were generated for the 54 nonsampled counties, they were used along with the population counts from the 1990 Census for each county to obtain the simulated numbers of homeless persons for those counties.

The second simulated population was created in a manner similar to the first except that separate normal distributions were used within each of the five geographic regions of Ohio. The means and standard deviations of the log rates of homelessness for the sampled counties within each region were used as the parameters of the normal distributions from which the simulated values were generated. Again the simulated log rates were used to obtain the numbers of homeless persons for the 54 nonsampled rural counties.

The third simulated population was generated using the regression of rate of homelessness per 10,000 on the percent elderly in each sampled county. (This choice of predictor variable is based on the selection of a regression estimator as described in Section 4.) The fitted regression model is

$$\widehat{\text{rate}} = -9.02 + 2.32\%\text{elderly},$$

with $R^2 = 0.197$, $\sqrt{\text{MSE}} = 9.03$, and p-value $= 0.044$ for the overall F-test for the regression line. The simulated population was created by estimating the rate of homelessness in each nonsampled county from the percent elderly in the county and then adding a random normal error term. Because a plot of the residuals from the regression line suggested that the variance in the residuals is larger for counties with higher percentages of elderly, the random error terms were generated from two different normal distributions depending on whether the percent elderly in the county was more or less than 10%. The standard deviations used for the two normal distributions were the standard deviations in the residuals for the counties with 10% or more elderly and with less than 10% elderly.

The fourth simulated population was generated using the regression of rate of homelessness per 10,000 on the percent elderly in each sampled county and on the indicators of the region of the state to which the county belongs. Using I_{NE}, I_{NW}, I_C, and I_{SE} to represent indicator variables for the northeast, northwest, central, and southeast regions respectively, the fitted regression model is

$$\widehat{\text{rate}} = -10.40 + 3.23\%\text{elderly} - 6.47 I_{NE} - 8.55 I_{NW} - 8.64 I_C - 14.25 I_{SE},$$

with $R^2 = .407$ (R^2-adjusted $= .210$), $\sqrt{\text{MSE}} = 8.73$, and p-value $= 0.127$ for the overall F-test for the regression line. The simulated population was created by estimating the rate of homelessness in each nonsampled county from the regression equation and then adding a random normal error term. A residual plot again suggested that the variance in the residuals is larger for counties with higher percentages of elderly. Thus the random error terms were generated from two different normal distributions depending on whether the percent elderly in the county was more or less than 10%. Again, the standard deviations for the two normal distributions were the standard deviations of the appropriate subsets of residuals.

The fifth simulated population was generated to be somewhat different from the other populations. It was generated using a regression model to predict number of homeless directly from the population size within each county. The fitted regression model is

$$\widehat{\text{homeless}} = 13.23 + 0.001154\text{population},$$

with $R^2 = 0.386$, $\sqrt{\text{MSE}} = 54.29$, and p-value $= 0.003$ for the overall F-test for the regression line. The simulated population was created by estimating the number of homeless persons in each nonsampled county from the fitted regression equation and then adding a random normal error term. Because a plot of the residuals suggested that the variance in the residuals is larger for counties with larger populations, the random error terms were generated from two different normal distributions depending on whether the county population was more or less than 30,000. The standard deviations for the two normal distributions were the standard deviations of the appropriate subsets of residuals.

After the five populations had been generated, they were each used to assess the amount of bias in the estimates of the rate of rural homelessness. Since we had created the entire "population", we could compute the "true" rate of homelessness within the entire state and the five geographical regions for each of the five populations.

In the simulation, samples of 21 rural counties were selected using the stratified sampling scheme that was used for the actual study. That is, four counties were sampled at random without replacement from each of the northeast, northwest, central, and southwest regions; five were sampled from the southeast region. The estimated rates of homelessness were computed for the five regions and for the state using the formulas given in Section 3.1. These estimates were compared to the population rates of homelessness for the simulated population to determine the bias in the estimate. This process of selecting a sample, computing estimates, and determining the bias was repeated 1 million times with replacement for each simulated population. (The number of possible samples is more than 7.15 × 10^{15}.) The same stream of random numbers was used to select the samples for each of the five populations. The results of the simulation are presented in Table 2.

Table 2

Bias in the Estimate of the Homeless Rate per 10,000 for Five Simulated Populations (Based on 1,000,000 simulated samples)

	Population				
	1	2	3	4	5
STATE	0.0406	0.1308	0.2618	0.2433	0.2547
	(2.056)	(1.759)	(2.144)	(1.807)	(1.605)
REGION					
NE	−0.0406	−0.0379	0.1538	0.0317	0.0993
	(3.333)	(2.923)	(3.748)	(4.034)	(1.937)
NW	−0.0578	−0.2948	0.0529	0.0254	0.3234
	(3.591)	(3.194)	(3.474)	(3.460)	(4.249)
C	−0.2442	0.2700	0.3974	0.1362	0.1901
	(3.122)	(3.762)	(3.426)	(2.260)	(2.869)
SE	−0.1034	−0.0279	−0.1132	−0.1798	0.0427
	(6.512)	(4.298)	(6.600)	(3.892)	(3.973)
SW	0.6184	0.8093	0.9196	1.277	0.6716
	(4.215)	(4.990)	(4.610)	(5.173)	(4.274)

Note: The standard deviation of the simulated sampling distribution of the estimator is given in parentheses below each value.

From Table 2 we see that the size of the bias in the overall state estimate of homelessness is about 1/100th of the size of the estimate itself. (Recall that the actual estimated six-month rate of homelessness for the state is about 14 per 10,000 population. The simulated populations have state rates between about 13 and 15 per 10,000.) At the regional level, the size of the bias is also about 1/100th of the size of the regional estimates even though the regional estimates are based on much smaller sample sizes. These results suggest that the size of the bias in our actual estimate is likely to be relatively small.

As would be expected from the small number of counties in the sample, the variance of the sampling distribution of the estimator is fairly large. The standard deviation in the estimates from the simulation study was about 10 times the size of the bias. (The standard deviations of the 1,000,000 estimates in each of the five simulations are of the same order of magnitude as the standard error of the actual estimate shown in Table 1.) This result suggests that the bias in the actual estimate is likely to be rather unimportant when compared to the standard error of the estimate.

Finally, we assessed the shape of the sampling distribution of our estimator by looking at histograms of the 1,000,000 estimates from each of our five simulation studies. The histograms appeared symmetric, mound shaped, and remarkably like histograms of normal data. Thus, confidence intervals based on the normal approximation are likely to be fairly accurate.

4. A REGRESSION ESTIMATOR

There is a great deal of information available, for example from the Bureau of the Census, on the economic conditions in a county. We hoped to be able to use some of this information to improve our estimate for the rate of homelessness by using a regression estimator. To this end, we searched for a regression model relating either the number of homeless persons in a county or the rate of homelessness with a variety of predictor variables which we thought might be useful in explaining homelessness. These possible predictor variables included county population, percentage change in population from 1980 to 1990, unemployment rate, percent elderly, public welfare expenditures, average weekly earnings, percent of rental property, median rent, poverty rate, percent female head of household, percentage of land in farming, average value of farms, average income per farm, ratio of manufacturing to farm jobs, indicator of Beale scores - a classification system for degree of ruralness (see Thomas 1977), and regional indicators.

None of these possible predictors individually or in combination provided a good predictor of the number of homeless persons or rate of homelessness. The best single predictor was percent elderly, the model which was used in generating the third simulated population described in Section 3.2, but it explained less than 20% of the variability in the rate of homelessness. No other variable was useful in addition to percent elderly and we could not find another reasonable regression model. Thus we used percent elderly in a regression estimator for the state rate of rural homelessness. Note that percent elderly is a plausible predictor of the rate of homelessness because poor economic conditions in a rural county appear to result in out-migration of the young; the elderly remain behind making up a greater proportion of the population. Therefore, it is possible that the percentage of elderly in a county is a proxy for poor economic conditions and out-migration. We cannot, however, rule out the possibility that percent elderly appears to be related to rate of homelessness in our data due to chance. We also realize that unavoidable errors in the county-based data collection procedures, such as interviewer effect, amount of services available, and geographic factors, may contribute to the lack of association between rate of homelessness and theoretically relevant variables.

We used the combined regression estimator (see, for example, Cochran 1977) to obtain the state estimate of 14.85 rural homeless per 10,000 with a standard error of 1.64. This compares with the original estimate of 14.00 with a standard error of 2.09 as shown in Table 1. Because the regression estimator is also biased with the bias decreasing for larger sample sizes, we again used a simulation study to assess the bias in this regression estimator.

The simulation study for the regression estimator was carried out using the third and fourth simulated populations described in Section 3.2 because those populations were generated using a regression model involving percent elderly. The simulation again computed the bias in the estimate for 1 million stratified cluster samples chosen with replacement from each population. The same stream of random numbers was used to generate the samples in both cases. A summary of the results of the simulation study for both the original estimator and the regression estimator is given in Table 3.

Table 3

Comparison of Estimators of State Homeless Rate per 10,000 (Summary for 1,000,000 repetitions from two simulated populations)

	Original Estimator		Regression Estimator	
	Population		Population	
	3	4	3	4
Average Bias	0.2618	0.2433	1.7115	0.8360
Standard Deviation	2.144	1.807	1.820	1.246
MSE	4.664	3.325	6.242	2.250

Note that the average bias is larger for the regression estimator than for the standard estimator for a rate from a stratified cluster sample. The standard deviation of the sampling distribution for the regression estimator, however, appears to be slightly smaller than that of the original estimator for each of the two simulated populations. The mean squared errors for the regression estimator fell above and below those of the original estimator. Thus, the choice of which estimator to use was unclear from the summary information in Table 3.

Because the regression estimator does not provide a clear improvement over the original estimator, the bias on average appears to be larger for the regression estimator, and the percent elderly variable may have been selected out of the many variables we tried due to chance, we chose to use the standard estimator of Section 3 for estimating the rate of rural homelessness.

5. CONCLUSIONS

The most often quoted national figures on homelessness were published by Burt and Cohen (1989) who estimated rates of homelessness in urban areas at 37.4 per 10,000 population in cities of more than 100,000 and 9 per 10,000 outside of SMAs. This current study of homeless persons in nonurban Ohio gives a six-month rate of about 14 homeless per 10,000 population and a one-week rate of 5.68 per 10,000 population.

The results of our simulation study suggest that the bias in the usual estimate of a rate based on our small cluster sample is not likely to be important, particularly in comparison to the size of the standard error of the estimate. The bias in the estimates for the five geographic regions in Ohio was found to be of a similar, relatively small size. The simulation study suggests that statistical biases and errors are not likely to discredit the substantive results of the survey of rural homeless.

Our regression analysis of the numbers of homeless persons from sampled counties suggests that it is difficult to explain the numbers of homeless persons in nonurban counties using economic and demographic variables that might be thought to be related to homelessness. It may be that each county is so different from the others, because of its location relative to population centers and related economic characteristics, that it is impossible to find a suitable stratification of the nonurban counties within Ohio. The use of a geographically stratified sample in Ohio did not appear to reduce the variance of the estimate and no other stratification variable was suggested by our regression analysis. This may be the case for other states as well, although stratification by some variable may be possible over, say, the entire United States.

ACKNOWLEDGEMENTS

This research was supported in part by Grant #R01MH46111 from the National Institute of Mental Health and by the Ohio Departments of Health and Mental Health. The authors wish to recognize the excellent work of Ms. H. C. Tsai who did most of the computing for this research. The authors thank two referees for their useful comments on an earlier version of this paper.

REFERENCES

COCHRAN, W.G. (1977). *Sampling Techniques*, (3rd edition). New York: Wiley.

BURT, M.R., and COHEN, B. (1989). *America's Homeless: Numbers, Characteristics and Programs that Serve Them.* Urban Institute Report 89-3. Washington, DC: Urban Institute Press.

BURT, M.R., and TAEUBER, C. M. (1991). Overview of seven studies that counted or estimated homeless populations in *Enumerating Homeless Persons: Methods and Data Needs.* Proceedings of Census Bureau/ Interagency Council/ Department of Housing and Urban Development Conference, November 29-30, 1990, edited by C.M. Taeuber, Washington, D.C.: US Department of Commerce, 30-76.

FIRST, R.J., TOOMEY, B.G., RIFE, J.C., and STASNY, E.A. (1994). Outside of the City: A Statewide Study of Homelessness in Nonurban/Rural Areas. Final report for NIMH Grant #R01MH46111. Columbus, OH: College of Social Work, The Ohio State University.

ROSSI, P.H., WRIGHT, J.D., FISCHER G.A., and WILLIS, G. (1987). The urban homeless: estimating composition and size. *Science*, 235, 1336-1341.

ROTH, D., BEAN, J., LUST, N., and SAVEANU, T. (1985). *Homelessness in Ohio: A Study of People in Need.* Ohio Department of Mental Health, Columbus, Ohio.

SPITZER, R., ENDICOTT, J., and COHEN, J. (1970). The psychiatric status schedule: A technique for evaluation psychopathology and impairment in role functioning. *Archives of General Psychiatry*, 23, 41-55.

TAEUBER, C.M., and SIEGEL, P.M. (1990). Counting the Nation's Homeless Population in the 1990 Census. Paper presented at the Conference on Enumerating Homeless Persons: Methods and Data Needs, Washington, D.C., November 29, 1990.

THOMAS, D.W. (1977). Beale Code Revisions Based on Those Devised by Fuguitt and Beale. The Rural Turnaround in Ohio: 1970-1975. E.S.S. #560. Columbus, OH: Department of Agricultural Economics and Rural Sociology, The Ohio State University.

TOOMEY, B.G., FIRST, R.J., GREENLEE, R., and CUMMINS, L. (1993). Counting the rural homeless population: Methodological dilemmas. *Social Work Research and Abstracts*, 24, 23-27.

GUIDELINES FOR MANUSCRIPTS

Before having a manuscript typed for submission, please examine a recent issue (Vol. 19, No. 1 and onward) of *Survey Methodology* as a guide and note particularly the following points:

1. Layout

1.1 Manuscripts should be typed on white bond paper of standard size (8½ × 11 inch), one side only, entirely double spaced with margins of at least 1½ inches on all sides.
1.2 The manuscripts should be divided into numbered sections with suitable verbal titles.
1.3 The name and address of each author should be given as a footnote on the first page of the manuscript.
1.4 Acknowledgements should appear at the end of the text.
1.5 Any appendix should be placed after the acknowledgements but before the list of references.

2. Abstract

The manuscript should begin with an abstract consisting of one paragraph followed by three to six key words. Avoid mathematical expressions in the abstract.

3. Style

3.1 Avoid footnotes, abbreviations, and acronyms.
3.2 Mathematical symbols will be italicized unless specified otherwise except for functional symbols such as "exp(·)" and "log(·)", etc.
3.3 Short formulae should be left in the text but everything in the text should fit in single spacing. Long and important equations should be separated from the text and numbered consecutively with arabic numerals on the right if they are to be referred to later.
3.4 Write fractions in the text using a solidus.
3.5 Distinguish between ambiguous characters, (e.g., w, ω; o, O, 0; l, 1).
3.6 Italics are used for emphasis. Indicate italics by underlining on the manuscript.

4. Figures and Tables

4.1 All figures and tables should be numbered consecutively with arabic numerals, with titles which are as nearly self explanatory as possible, at the bottom for figures and at the top for tables.
4.2 They should be put on separate pages with an indication of their appropriate placement in the text. (Normally they should appear near where they are first referred to).

5. References

5.1 References in the text should be cited with authors' names and the date of publication. If part of a reference is cited, indicate after the reference, e.g., Cochran (1977, p. 164).
5.2 The list of references at the end of the manuscript should be arranged alphabetically and for the same author chronologically. Distinguish publications of the same author in the same year by attaching a, b, c to the year of publication. Journal titles should not be abbreviated. Follow the same format used in recent issues.