SURVEY METHODOLOGY

Catalogue No. 12-001-XPB

A JOURNAL PUBLISHED BY STATISTICS CANADA

DECEMBER 2004

•

VOLUME 30

•

NUMBER 2

Statistics Canada Statistique Canada

Canada

SURVEY METHODOLOGY

A JOURNAL PUBLISHED BY STATISTICS CANADA

DECEMBER 2004 • VOLUME 30 • NUMBER 2

Published by authority of the Minister
responsible for Statistics Canada

January 2005

Catalogue no. 12-001-XPB

Frequency: Semi-annual

ISSN 0714-0045

Ottawa

Statistics Canada Statistique Canada

Canada

SURVEY METHODOLOGY

A Journal Published by Statistics Canada

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

EDITORIAL POLICY

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

Submission of Manuscripts

Survey Methodology is published twice a year. Authors are invited to submit their articles in English or French in electronic form, preferably in Word to the Editor, Dr. M.P. Singh, singhmp@statcan.ca (Household Survey Methods Division, Statistics Canada, Tunney's Pasture, Ottawa, Ontario, Canada, K1A 0T6). For formatting instructions, please see the guidelines provided in the Journal.

Subscription Rates

The price of *Survey Methodology* (Catalogue no. 12-001-XPB) is CDN $58 per year. The price excludes Canadian sales taxes. Additional shipping charges apply for delivery outside Canada: United States, CDN $12 ($6 × 2 issues); Other Countries, CDN $30 ($15 × 2 issues). Subscription order should be sent to Statistics Canada, Dissemination Division, Circulation Management, 120 Parkdale Avenue, Ottawa, Ontario, Canada, K1A 0T6 or by dialling 1 800 700-1033, by fax 1 800 889-9734 or by E-mail: order@statcan.ca. A reduced price is available to members of the American Statistical Association, the International Association of Survey Statisticians, the American Association for Public Opinion Research, the Statistical Society of Canada and l'Association des statisticiennes et statisticiens du Québec.

SURVEY METHODOLOGY

A journal Published by Statistics Canada

Volume 30, Number 2, December 2004

CONTENTS

Survey Methodology, December 2004
Vol. 30, No. 2, pp. 125-126
Statistics Canada

In This Issue

This issue of Survey Methodology opens with a discussed paper by Paul Biemer. He provides evidence of reduced accuracy due to the redesign of employment questions in the Current Population Survey (CPS). This is an extension of the previous study by Biemer and Bushery (2000). In the current paper, the author attempts to trace the source of the error through extended analysis of the CPS data before and after the redesign. A new approach, using Markov Latent Class Analysis, is presented. This work aims at providing guidance for further investigation into the root causes of the errors in the collection of labour force data in the CPS. Discussions of this paper are provided by Jeroen Vermunt, Stephen Miller and Anne Polivka, and Clyde Tucker.

In their paper, Gunning and Horgan propose a new algorithm for the construction of stratum boundaries in skewed populations. Their algorithm uses an auxiliary variable and achieves equal coefficients of variation for this auxiliary variable in each stratum. The method is based on the assumption that the auxiliary variable is uniformly distributed. One advantage of the method is that it is very easy to apply in practice. In an empirical study, the authors show that the proposed algorithm compares favourably with the cumulative root frequency method of Dalenius and Hodges (1957) and to the Lavallée and Hidiroglou (1988) algorithm.

Hedlin and Wang consider the problem of bias coming from feeding back information from sample surveys to frames. They investigate the bias incurred by updating deaths on a frame that is used for future occasions of the same survey. They quantify this bias and develop an unbiased estimator for this situation. The theoretical results presented in the paper are illustrated through a simulation study.

In their paper, Mudryk and Xie present the Quality Assurance (QA) and Quality Control (QC) aspects of the Intelligent Character Recognition operation of the 2001 Canadian Census of Agriculture. They show how an effective QA and QC plan was developed to ensure the highest quality data from the data capture operation of the Census. Results from an analysis of the Average Outgoing Quality of the data indicate the importance of a QA/QC plan.

In Park and Lee, the design effects for the weighted mean and total estimators are investigated for complex surveys. In particular, they decompose the design effect for the weighted mean and total estimators under a two-stage design. Given this decomposition, they illustrate several common misconceptions about the design effects for the weighted mean and total estimators through several examples using commonly used designs.

In their paper, Beaumont and Alavi investigate a robust generalized regression estimator. They look at alternatives to the optimal Best Linear Unbiased (BLU) estimator that are robust to design ignorability and/or model misspecification. In the situation where the design ignorability assumption may not hold, they propose a least squares estimator that is obtained by shrinking the design weights to their mean. To deal with model misspecification, they propose a weighted generalized M-estimator to reduce the influence of units with large weighted population residuals. Their theoretical results are illustrated with a simulation study.

Zheng and Little propose a non-parametric model-based alternative to Horvitz-Thompson estimation of a total in the case of two-stage sampling with pps sampling at the first stage. This is an extension of their earlier work in which an outcome variable y_i is modeled as a smooth function of the inclusion probability π_i. They show how to fit the model and estimate the total using a penalized spline, and also develop alternative variance estimation procedures. Simulations are used to compare the proposed method to the Horvitz-Thompson estimator and to a model-assisted estimator.

Liang and Kuk consider an alternative to the standard approach for regression estimation in a finite population. Instead of the usual linear model they use an arbitrary smooth function to allow for a non-linear regression, and then they apply Bayesian neural networks to the problem. The advantage of the neural network approach is that the problem of model misspecification is avoided. Liang and Kuk place a prior on each network connection instead of on the number of hidden units as is usually done. This permits a unified approach to the selection of the network structure and the selection of the auxiliary variables. Finally, they handle outliers by introducing a heavy tail distribution to model the disturbances of the data.

In the last paper of this issue, Reiter uses multiple imputation to handle simultaneously both missing data and disclosure limitation. The basic idea is to fill in the missing data first to generate m completed datasets and then replace sensitive or identifying values in each completed dataset with r imputed values. Then, the author develops new combining rules for obtaining valid inferences from such multiply-imputed datasets. These rules take into account both sources of variability in the point estimators.

Finally, the Editorial Board met this past summer at the Joint Statistical Meetings in Toronto. A suggestion was made at that meeting to have a Short Communications section in the journal. These would be shorter papers, typically around four Survey Methodology pages. Possible topics of short communications would include presentation of new ideas without the full development of a regular paper, brief reports of empirical work, and discussions or supplements to other papers published in the journal. All short communications would be refereed, although the reviewing process may be streamlined. I hope that this new format will be attractive to many authors, and look forward to receiving your submissions.

M.P. Singh

Survey Methodology, December 2004
Vol. 30, No. 2, pp. 127-140
Statistics Canada

An Analysis of Classification Error for the Revised Current Population Survey Employment Questions

PAUL P. BIEMER [1]

ABSTRACT

The reduced accuracy of the revised classification of unemployed persons in the Current Population Survey (CPS) was documented in Biemer and Bushery (2000). In this paper, we provide additional evidence of this anomaly and attempt to trace the source of the error through extended analysis of the CPS data before and after the redesign. The paper presents an novel approach decomposing the error in a complex classification process, such as the CPS labor force status classification, using Markov Latent Class Analysis (MLCA). To identify the cause of the apparent reduction in unemployed classification accuracy, we identify the key question components that determine the classifications and estimate the contribution of each of these question components to the total error in the classification process. This work provides guidance for further investigation into the root causes of the errors in the collection of labor force data in the CPS possibly through cognitive laboratory and/or field experiments.

KEY WORDS: Survey redesign; Measurement error; Latent class analysis; Unemployment rate; Specification error.

1. INTRODUCTION

The Current Population Survey (CPS) is a monthly survey of approximately 60,000 households conducted by the U.S. Bureau of the Census for the Bureau of Labor Statistics (BLS). The primary purpose of the survey is to provide estimates of employment, unemployment, and other characteristics of the general U.S. labor force population. Estimates of the size, composition, and dynamic characteristics of the labor force are published each month by BLS and comprise one of the Nation's key economic indicators.

In January 1994, a revised questionnaire was introduced in the CPS to address the recommendations by the Levitan Commission in the late 1970s to convert the mode of interview for the CPS from paper and pencil questionnaire to computer-assisted interviewing methods, to clarify some of the questions on employment, as well as for a number of other reasons described in Rothgeb (1994). The overall objective of the redesign was to improve the quality of the data collected in the CPS. The CPS questionnaire had remained essentially unchanged since the last major revision in 1967.

The revised CPS questionnaire was introduced after considerable research and testing that began in the mid-1980s. The purpose of the testing was to evaluate the quality and operational feasibility of various redesign options including moving the CPS from a paper and pencil questionnaire format to computer assisted interviewing. During these years of testing, more than 100,000 persons were interviewed in the various studies that were conducted (Rothgeb 1994). The CPS redesign research program culminated in a large national study (referred to in the literature as the CATI/CAPI Overlap or CCO Field Test) that was conducted in 1993. The key component of this test consisted of a computer assisted survey of approximately 12,000 households implementing revised CPS interviewing procedures and the revised questionnaire. This survey, referred to in this report as the Parallel Survey, was conducted from July 1992 to December 1993 concurrently with the ongoing CPS survey which used the original questionnaire. This type of split panel design makes it possible to estimate the effect of the redesign changes on the CPS labor force estimates.

A number of papers and reports were published documenting the findings from the CCO Field Test (Cohany, Polivka and Rothgeb 1994; Rothgeb 1994; Polivka 1994; Kostanich and Cahoon 1994; Miller 1994; Thompson 1994; Dippo, Polivka, Creighton, Kostanich and Rothgeb 1994). One key finding from this research was that the Parallel Survey unemployment rate and the labor force participation rate were higher than in the CPS. The higher unemployment and labor force participation rates associated with the revised questionnaire were explained primarily by changes in the definition of employment. The revised questionnaire has a broader approach to both work and job search activities, which would tend to classify more persons as "in the labor force" and, thus, more persons who are not working as unemployed rather than out of the labor force (see, for example, Polivka 1994 and Rothgeb 1994).

The increase in the unemployment rate due to the new design was originally estimated at about one-half percentage point. However, further analysis of the Parallel Survey data

[1] Paul P. Biemer, 3040 Cornwallis Road, PO Box 12194 Research Triangle Park, NC 27709-2194, U.S.A.

called that estimate into question and subsequently a report was release estimating the increase to be less than one-tenth percentage point (Polivka and Miller 1994). The concerns raised in the subsequent reports regarding the utility of the Parallel Survey data for assessing the effect of the redesign are discussed further below and will be considered in our analysis of these data.

An independent analysis conducted by Biemer and Bushery (2000) revealed an anomaly in the revised CPS labor force data that had not been detected by any of the previous research on the CPS redesign. Using a Markov latent class analysis (MLCA) approach, Biemer and Bushery compared the accuracy of labor force classifycations under the original and revised designs by estimating and comparing the error rates using the 1993 CPS data and the 1995 and 1996 CPS data. They defined labor force classification accuracy as the probability that a person who is truly in some labor force category, say category a, is classified as being in a by the CPS; *i.e.*, Pr(classified in a | truly in a). For example, the classification accuracy for unemployment is the probability a person who is truly unemployed, according to the CPS definition, is correctly classified as unemployed by the CPS classification rules.

In Table 2 of their paper, Biemer and Bushery report that the classification accuracy for unemployment dropped by 5.7 percentage points, from approximately 81.8 percent (s.e. = 0.90) in 1993 to 76.1 (s.e. = 1.2) in 1995 and 74.4 percent (s.e. = 1.2) in 1996. These results suggest that the redesigned CPS misclassifies the true unemployed at a higher rate than the old CPS design. The authors first considered that this result could be an artifact of the MLCA methodology. As shown below, MLCA does not require a true or "gold standard" measurement of employment to estimate classification error. Rather the method relies a model describing the true month to month changes in employment status and as well as for the process of classifying individuals into labor force categories. It is possible that labor force transitions that deviate from the model specification could be regarded as misclassifications in the estimation process.

To check the validity of the MLCA results, the authors conducted a series of analyses using traditional estimation approaches, analysis of the error by population groups, comparisons of the error estimates to other published estimates, and simulations to assess the effect of model failure on the results. As an example, there is evidence that the test-retest reliability of the unemployment category decreased after the redesign. Prior to the redesign, the index of inconsistency (The index of inconsistency is a measure of unreliability traditionally used at the Census Bureau. It is equal to $1 - \kappa$ where κ is Cohen's kappa coefficient (Cohen 1960) for the unemployed labor category averaged 30 percent for the period 1992–1993. Following the redesign, the index of inconsistency increased to almost 40 percent for the period 1995–1996. These analyses support their claim that the accuracy of the CPS methodology for classifying unemployed persons declined after the redesign.

In their discussion of the results, the authors speculated that the drop in classification accuracy could indicate a problem with the revised unemployment questions. That is, the revised unemployment questions may be subject to greater classification error and, thus, less classification accuracy. Another possibility they considered is change in the characteristics of the unemployed populations from 1993 to 1995 and 1996. Since the unemployment rate dropped from 1993 to 1996, it is possible that persons who would be more accurately classified by the CPS system left the ranks of the unemployed, leaving persons who would be less accurately classified in the category. This hypothesis could be tested by estimating the accuracy rates for the two methodologies for the same time period. The Parallel Survey offers a means to conduct such an analysis.

The current paper continues the investigation of the reduction in MLCA unemployment classification accuracy rates observed by Biemer and Bushery. The current analysis uses MLCA models very similar to those used by Biemer and Bushery for estimating the classification accuracy for the original and revised versions of the CPS questionnaire. However, the time period considered here is expanded to include the15 months prior to and following the introduction of the revised questionnaire: a total of 30 contiguous months. In addition, data from the Parallel Survey from the period January 1993 through December 1993 is used to compare the employment accuracy for original and revised questionnaire for the same time period.

Our analysis focuses on a labor force classification variable that is derived from a number of questions on the employment section of the CPS questionnaire. This variable is often referred to as a "recoded" labor force variable since it is determined by mapping a pattern of CPS responses to questions about employment onto particular labor force categories such as employed – at work, employed – not at work, unemployed – looking for work, and so on. Biemer and Bushery used a three-category employment classification variable: employed (EMP), unemployed (UEM), and not in the labor force (NLF). For the present analysis, a four-category variable is used that subdivides the UEM category into unemployed-on layoff (UEM-LAYOFF) and unemployed-looking for work (UEM-LOOKING). This is done as a first step toward isolating the source of the apparent inaccuracy in unemployment classification. However, further decomposition of these categories will be necessary to arrive at the root source of the error as will be shown subsequently.

In section 2 we describe the CPS labor force concepts that are most relevant to our study and the structure of the data sets in the analysis. In section 3 we review the MLCA estimation methodology and models used by Biemer and Bushery in their analysis and describe the application of their methodology for the present purposes. In section 4 we present the results of our analysis and what they suggest regarding the source of the classification error in the new questionnaire. Finally, section 5 provides a summary of the key findings and our conclusions from the study.

2. DATA AND CONCEPTS

2.1 The Data Sets for Our Study

Except for the Parallel Survey, the CPS data in our analysis were downloaded from the National Bureau of Economic Research (NBER) website (www.nber.org). This website contains microdata for the CPS for every month from January 1976 through December 2004. The MLCA approach was applied directly to these microdata without the need for supplementary data or data external to the CPS.

In the preliminary analysis, we investigated the CPS classification accuracy for a six-year period: January 1992 through December 1997. That analysis was aimed at determining whether the anomaly first noted in Biemer and Bushery (2000) is a transient phenomenon affecting only the months immediately following the introduction of the new questionnaire or whether it persisted for some years after the new questionnaire was introduced. If temporary or transient, the anomaly might be related to problems during the phase-in of the new design; for example, interviewer training or issues related to the startup of data collection. However, evidence of a persistent, continuing effect could suggest problems with the survey design; for example, the questionnaire, interviewing procedures, or the recoding algorithm.

By applying MLCA across all months from 1992 through 1997 we determined that, although the magnitude of the reduction in accuracy varies somewhat from month to month, it does indeed persist for all months following the introduction of the revised questionnaire. The results confirmed Biemer and Bushery's conjecture of a systemic effect possibly linked to the new unemployment questions introduced in January 1994.

Due to space considerations, in this paper we present results from a somewhat shorter time frame than considered in the preliminary analysis, *viz.*, the years 1992, 1993, 1994, and 1995. This time period covers two years of the CPS using the original questionnaire and two years using the revised questionnaire. In addition, we will also present some results from an MLCA of the 1993 Parallel Survey data that can be compared with results from the main CPS.

The data sets in our study are quite large. Each estimate of classification error we obtain is based upon all households that were interviewed in the CPS for three consecutive months. Across the four years in our analysis, the total number of households responding for all three months in any three-month period varies from about 37,000 to more than 40,000. For the 1993 Parallel Survey, the number of households satisfying this criterion is approximately 10,000. The estimates we produce are appropriately weighted for probabilities of selection and other post-survey adjustments and, therefore, reflect the response probabilities of the published CPS estimates. Weights were constructed by taking an average weight across the three consecutive months that were combined to form a longitudinal record for the analysis (unweighted analyses were also conducted and the results were very similar to the weighted analysis. This suggests the choice of weights has little effect on the study outcomes).

Because of a problem in the identification variables required for linking households for the months June 1995 though December 1995, it was not possible to include these months in our analysis. Further, since our conclusions would not change by including data from the 1996 or later years of the CPS, we confine our analysis to 15 months prior and 15 months following the introduction of the revised questionnaire. Thus, for most of the analysis to follow, we will provide averages of estimates from August 1992 through December 1993 for the original questionnaire and from January 1994 through May 1995 for the revised questionnaire (note that since our estimates are based upon a moving average of three consecutive months, seasonal variations in the labor rates and transitions probabilities are accounted for in the estimates of classification error).

2.2 Labor Force Concepts

The revised CPS questionnaire was introduced in 1994 to improve the overall quality of labor market information through extensive question changes and through the use of computer technology in the data collection. In the following, we describe a few concepts that were affected by the questionnaire redesign and that are relevant for the current analysis.

Employed. The labor force questions in the original questionnaire began with the question "What were you doing most of LAST WEEK (working, keeping house, going to school, or something else)?" Interviewers were allowed to modify the parenthetical part of this question according to the age of the respondent. In some cases, the word "work" or "working" was not part of the question. As an example, if the respondent looked of student-age, the interviewer was allowed to leave out the word "working." The revised questionnaire replaced this question with two

questions: "Does anyone in this household have a business or a farm?" and "LAST WEEK, did you do ANY work for (either) pay (or profit)?" where the parenthetical parts of the question are read if anyone in the response to the first question is "yes." Further, additional questions were added to clarify whether earnings or profits were received from the family business or farm. Thus, the revised questionnaire concept of employment appears to be somewhat broader and better defined than the original questionnaire concept.

Unemployed. The definition of unemployment was slightly modified in the revised questionnaire. In the original questionnaire, persons waiting for a new job to start were classified as unemployed. Under the revised questionnaire definition, a person is unemployed only if all of the following are true: (1) without a job, (2) actively seeking work or on layoff from a job and expecting recall within the next six months, and (3) currently available to take a job (except for a possible temporary illness).

On Layoff. Persons on layoff are defined as persons separated from a job and who are awaiting a recall to return to that job. The original questionnaire did not consider or collect information on the expectation of recall. This was problematic because to most people, the term "layoff" could mean permanent termination from the job rather than the temporary loss of work economists are trying to measure.

Job Search Methods. To be counted as unemployed and looking for work, a person must have engaged in an active job search during the four weeks prior to the survey. The revised questionnaire includes a somewhat broader question about job search methods with expanded and restructured response categories to allow interviewers to more easily record and distinguish between active and passive job search activities. In addition, it provides additional followup questions for those who respond "nothing" or "don't know."

Reference Week. While the original questionnaire referred to LAST WEEK, the reference period was never explicitly defined. The revised questionnaire provides specific dates of the reference week.

We will refer to these changes later in the report when we discuss the differences in the classification error and specification error between the revised and original questionnaires.

As previously noted, Biemer and Bushery focused on a three-category labor force recoded variable with categories: employed (EMP), unemployed (UEM), and not in the labor force (NLF). For the present analysis, we used an expanded recoded variable also available on the CPS public use data files. This variable divides the UEM category into two categories corresponding to persons on layoff (LAYOFF) and persons looking for work (LOOKING). The seven-category variable also divides the EMP and NLF categories into subcategories; however, this level of detail in the EMP and NLF categories is not needed in our analysis. Thus, the seven-category variable will be collapsed to a four-category variable corresponding to EMP, UEM-LOOKING, UEM-LAYOFF, and NLF. The correspondence between the three- and four- category variables is shown in Figure 1.

3. LATENT CLASS MODELS FOR CPS CLASSIFICATION ERROR

Markov latent class models were first proposed by Wiggins (1973) and refined by Poulsen (1982). Van de Pol and de Leeuw (1986) established conditions under which the model is identifiable and gave other conditions of estimability of the model parameters. In this section we describe the basic model proposed by Biemer and Bushery (2000) and its extensions for application in the current analysis.

Let the CPS target population be divided into L groups (such as age, race, or sex groups) and let the variable G be the label for group membership. For example, $G_i = 1$ if the

Original Seven-Variable Category		Four-Category	Three-Category
Old Questionnaire	New Questionnaire	Analysis Variable	Analysis Variable
1. Working—at work	1. Employed—at work	1. EMP	1. EMP
2. With job—not at work	2. Employed—absent		
3. Unemployed—on layoff[1]	3. Unemployed—on layoff	2. UEM—LAYOFF	2. EM
4. Unemployed—looking for work[1]	4. Unemployed—looking	3. UEM—LOOKING	
5. Working without pay (less than 15 hours in a family farm or business) or temporarily absent from a without pay job	5. Retired—not in labor force	4. NLF	3. NLF
6. Unavailable to take a job if one had been offered	6. Disabled—not in labor force		
7. Not in the labor force	7. Other—not in labor force		

[1] Note: In the original questionnaire, categories 3 and 4 are reversed compared to corresponding categories in the revised questionnaire.

Figure 1. Association of the Seven-Category Employment Recode Variable with the Three- and Four-Category Variables Used in the Analysis

i^{th} population member is in group 1, $G_i = 2$ for group 2 and so on. Let X_{gi}, Y_{gi}, and Z_{gi} denote the true labor force classifications for the i^{th} person in group $G = g$ (for $g = 1,\ldots,L$ and $i = 1,\ldots,n_g$) where X_{gi} is defined as

$$X_{gi} = \begin{cases} 1 & \text{if person } (g,i) \text{ is employed in time period 1} \\ 2 & \text{if person } (g,i) \text{ is unemployed – on layoff in time period 1} \\ 3 & \text{if person } (g,i) \text{ is unemployed – looking in time period 1} \\ 4 & \text{if person } (g,i) \text{ is not in the labor force in time period 1} \end{cases}$$

with analogous definitions for Y_{gi} and Z_{gi} for periods 2 and 3 respectively. Consistent with the conventions of the LCA literature, we will drop the subscripts from the variables to simplify the notation.

Let $\pi_{x,y,z|g}$ denote $\Pr(X = x, Y = y, Z = z \mid G = g)$, let $\pi_{y|g,x}$ denote $\Pr(Y = y \mid X = x, G = g)$ and let $\pi_{z|g,y,x}$ denote $\Pr(Z = z \mid Y = y, X = x, G = g)$. Then, the probability that an individual in group g has labor status x in period 1, y in period 2, and z in period 3 is $\pi_{xyz|g}$ which may be written as

$$\pi_{xyz|g} = \pi_{x|g}\ \pi_{y|gx}\ \pi_{z|gxy}. \qquad (1)$$

Finally, under the first order Markov assumption, which is a necessary condition for model identifiability (see Van de Pol and de Leeuw 1986), we assume

$$\pi_{z|gxy} = \pi_{z|gy} \qquad (2)$$

i.e., at period 3, the true status of an individual does not depend on the period 1 status, once the period 2 status is known. An alternate interpretation is that the current status, given the prior period's status, does not depend upon the prior period's transition.

Now, consider the observed labor force classifications from the CPS denoted by A_{gi}, B_{gi}, and C_{gi} for periods 1, 2, and 3, respectively, where

$$A_{gi} = \begin{cases} 1 & \text{if person } (g,i) \text{ is classified as EMP in time period 1} \\ 2 & \text{if person } (g,i) \text{ is classified as UEM – LAYOFF in time period 1} \\ 3 & \text{if person } (g,i) \text{ is classified as UEM – LOOKING in time period 1} \\ 4 & \text{if person } (g,i) \text{ is classified as NLF in time period 1} \end{cases}$$

with analogous definitions for the response indicators, B_{gi}, and C_{gi} for periods 2 and 3, respectively. Using an extension of the notation established above, we denote the response probabilities in each of these classifications as $\pi_{a|gx} = \Pr(A = a \mid X = x)$, with analogous definitions for $\pi_{b|gy}$ and $\pi_{c|gz}$. Thus, $\pi_{a=1|g,x=2}$ is the probability that the CPS classifies a person in group g as employed $(A = 1)$ when the true status is unemployed – on layoff $(X = 2)$. Likewise, $\pi_{a=2|g,x=2}$ is the probability that the CPS correctly classifies a person in group g as unemployed – on layoff.

Finally, we assume

$$\pi_{a,b,c|g,x,y,z} = \pi_{a|gx}\ \pi_{b|gy}\ \pi_{c|gz} \qquad (3)$$

or that classification error in the observed labor force status is independent across the three months.

The CPS labor force classifications for each month of a three consecutive month interval are the outcome variables in our analysis. Let A, B, and C denote the observed classifications and let X, Y, and Z denote the (unobserved) true classifications for Month 1, Month 2, and Month 3, respectively. Let G denote some grouping (or stratification) variable to be defined later in the analysis. Under these assumptions, we can write the probability for classifying a CPS sample member in cell (g, a, b, c) of the $GABC$ table as follows:

$$\pi_{gabc} = \sum_{x,y,z} \pi_g\ \pi_{x|g}\ \pi_{y|gx}\ \pi_{z|gy}\ \pi_{a|gx}\ \pi_{b|gy}\ \pi_{c|gz}. \qquad (4)$$

Extensions to more than one grouping variable are straightforward.

Under multinomial sampling, the likelihood function for the $GABC$ table is

$$\Pr(GABC) = C \prod_{g,a,b,c} \pi_{g,a,b,c}^{n_{gabc}} \qquad (5)$$

where C is the multinomial constant and Π denotes the product of the terms over the subscripts g, a, b, and c. Under the assumptions made previously, the model parameters are estimable using maximum likelihood estimation methods. Van de Pol and de Leeuw (1986) provide the formula for applying the E-M algorithm to estimate the parameters of this model and describe the conditions for their estimability. The ℓEM software (Vermunt 1997) was used to fit the MLCA models.

In their investigations of the validity of MLCA estimates for analyzing CPS labor force classification error, Biemer and Bushery analyzed CPS data collect during the first quarter of each of three years – 1993, 1995, and 1996. They also conducted several types of analysis using the CPS unreconciled reinterview data for the same time period. The reinterview analysis provided another approach for

estimating CPS classification error as well as evidence of the validity of the MLCA approach. Their evaluation of MLCA validity considered five criteria: (1) model diagnostics, (2) model goodness of fit across years of CPS, (3) agreement between the model and test-retest estimates of response probabilities, (4) agreement between the model and test-retest estimates of inconsistency, and (5) plausibility of the patterns of classification error. The MLCA method performed well in all five test. For example, the same model provided the best fit of the data for each year analyzed, there was good agreement between the latent class estimates of reliability and those derived from traditional test-retest methodology; and the estimated error rates were consistent with those of previous studies – for *e.g.*, Chua and Fuller 1987; Abowd and Zellner 1985; Porterba and Summers 1995; and Sinclair and Gastwirth 1998.

Ostensibly, the Markov assumption seems very unlikely to hold for labor force data. As an example, persons who are unemployed in months 1 and 2 of a consecutive three-month period may not have the same probability of being unemployed in a month 3 as persons who just became unemployed in month 2. The former group could contain more chronically unemployed persons than the group entering unemployment in month 2. Further, the group just entering unemployment in month 2 could contain a higher proportion of people temporarily out of work while changing jobs. Biemer and Bushery considered the consequences for the MLCA estimates of misclassification when the Markov assumption is violated.

Using simulation, Biemer and Bushery found that the bias in MLCA estimates of classification probabilities depends upon the severity of the departures of the CPS data from the Markov assumption. They defined two parameters, λ_1 and λ_2, which are ratios of conditional probabilities. λ_1 is the ratio of the probability of being employed in period 3 for a person with an (EMP, UEM) pattern for periods 1 and 2, respectively, divided by the probability of being employed in period 3 for a person with a (EMP, EMP) pattern. Similarly, λ_2 is the ratio of the probability being employed in period 3 for a person with an (UEM, UEM) pattern to the probability of being employed in period 3 for a person with a (EMP, UEM) pattern. Note that when $\lambda_1 = \lambda_2 = 1$, the Markov assumption holds exactly and greater departures of λ_1 and λ_2 from 1 correspond to greater departures of the data from the Markov assumption. Biemer and Bushery found that over a fairly wide range of values for λ_1 and λ_2, the absolute bias in the MLCA estimates of unemployment classification accuracy never exceeded 3 percentage points. For example, in the extreme case of a Markov assumption violation, the expected value of an MLCA estimate of unemployment accuracy would be 77 percent when the true parameter value is 80 percent. Their results suggest that, for the CPS application, MLCA is fairly robust to failures of the Markov assumption to hold.

Although it is virtually impossible to prove their validity, MLCA error estimates can be quite useful for identifying survey questions that are prone to classification error; *i.e.*, flawed questions. For example, Biemer (2004) and Biemer and Wiesen (2002) demonstrate the utility of MLCA methodology for identifying question problems and classification process deficiencies in large scale surveys. Notwithstanding that the MLCA assumptions may be violated to an unknown extent, its usefulness as a tool for exploring a number of important questionnaire design issues has been well-documented. For the present application, MLCA will be used to develop and test hypotheses regarding the sources of the anomaly reported by Biemer and Bushery for 1994 CPS redesign.

The MLCA model use in the present analysis is essentially the same model selected by Biemer and Bushery for their analysis. To account for population heterogeneity, they considered a number of demographic and other explanatory variables that might be highly correlated with classification error. The best performing variable a proxy or self-response indicator variable denoted by P where

$$P = \begin{cases} 1 & \text{if all three interviews are conducted by self response } (SELF) \\ 2 & \text{if two of the interviews are conducted by self response } (MOSTLY\ SELF) \\ 3 & \text{if two of the interviews are conducted by proxy response } (MOSTLY\ PROXY) \\ 4 & \text{if all three interviews are conducted by proxy response } (PROXY). \end{cases}$$

Their empirical findings showed this variable to be strongly related not only to reporting accuracy, but also current employment status and month to month employment transitions. For example, responses for the PROXY group were considerably less accurate than for the SELF group and, further, the PROXY group had somewhat higher unemployment than the SELF group.

The MLCA model also allows transition probabilities to vary by P (referred to as group heterogeneity) as well as by time periods (referred to as non-stationary transitions). In addition, the model assumes that response probabilities $\pi_{a|px}$, $\pi_{b|px}$, and $\pi_{c|px}$ are group-heterogeneous but are equal for all three months in the time interval. This leads to the following model for describing the cell probabilities in the $PABC$ table:

$$\pi_{p,a,b,c} = \sum_{x,y,z} \pi_p \, \pi_{x|p} \, \pi_{y|px} \pi_{z|py} \pi_{a|px}^{A|PX} \pi_{b|py}^{A|PX} \pi_{c|pz}^{A|PX} \qquad (6)$$

where $\pi^{A|PX}_{b|py} = \Pr(A = b \mid P = p, X = y)$ with similar definitions for $\pi^{A|PX}_{a|px}$ and $\pi^{A|PX}_{a|pz}$. That is, the three sets of response probabilities are equal to $\pi^{A|PX}$.

Note that for the present analysis, interest is focused on the overall response probabilities associated with the revised and original questionnaires and not the variation in error rates across proxy groups. Therefore, our analysis focuses on the overall accuracy of response, *i.e.*, $\pi^{A|X}_{a|x}$ or the mean response probability for the four levels of P combined.

4. COMPARISON OF REVISED AND ORIGINAL QUESTIONNAIRE CLASSIFICATION ERROR PROBABILITIES

4.1 Reduction in UEM Classification Accuracy for the Revised Questionnaire

As mentioned in section 2, the CPS data sets for this analysis are monthly samples from August 1992 through May 1995. Figure 2 shows how this the time interval was divided into 30 overlapping three-month intervals: 15 for the original questionnaire and 15 for the revised questionnaire. The intervals are numbered in the table for later reference. For example, time interval 1 covers the period from August 1992 through October 1992 in which the original questionnaire was in use. Therefore, this time interval can provide one estimate of the response probabilities, $\pi^{A|X}$, for the model in (6). Since there are 30 time intervals across the entire 34–month period in our analysis, 30 estimates of $\pi^{A|X}$ can be formed from these consecutive overlapping time intervals: 15 estimates for the original questionnaire and 15 estimates for the revised questionnaire.

To obtain a more stable estimate of $\pi^{A|X}$ for each questionnaire, the 15 estimates corresponding to the 15 time periods per questionnaire in Figure 2 were averaged. These estimates are shown in Tables 1 and 2. Since they are based on simple random sampling assumptions, the standard errors in the tables do not account for the unequal weighting and clustering effects of the CPS. Since the average CPS design effect is about 1.5 for estimates of unemployment, the standard errors in the tables are probably understated by 20 percent or less. This level of bias in the standard errors is inconsequential for the purposes of this paper due to the extremely large sample sizes in the analysis.

Table 1 compares the MLCA estimates of the classification error probabilities for the original and revised questionnaire versions for the three-category labor force classification scheme used by Biemer and Bushery. The first column of the table is the true (or latent) category, the second column is the observed (or CPS) category, and the cell entries are the response probabilities estimated from the MLCA using model (6). For each true class (EMP, UEM, or NLF), the accuracy rate is the cell corresponding to the observed category with the same label. For example, the accuracy of classifying persons who are truly employed is 98.68 percent (for the original questionnaire) and 98.84 percent for the revised questionnaire. Note that this entry corresponds to the cell where both the true category and the observed category are EMP. The other cells for EMP in column 1 are the error rates for EMP. For example, the MLCA estimate of the probability CPS classifies a person as UEM who is truly EMP is 0.42 for the original questionnaire and 0.39 for the revised questionnaire. The other cell entries are interpreted analogously.

Consistent with Biemer and Bushery's findings, the accuracy of the classification of unemployed persons is substantially and highly significantly lower for the revised questionnaire: 79.06 percent versus 73.50 percent, a difference of 5.6 percentage points. Further, the increase in classification error for unemployed persons is due to misclassifications in both the EMP and NLF force categories with slightly more misclassification in the latter category. Our estimates differ slightly from theirs since, as noted earlier, we are analyzing more months of data and using weighted estimates rather than unweighted as in their analysis.

Months Using Questionnaire	Old	Aug. 1992	Sept. 1992	Oct. 1992	Nov. 1992	Dec. 1992	Jan. 1993	...†	Aug. 1993	Sept. 1993	Oct. 1993	Nov. 1993	Dec. 1993
Month Using Questionnaire	New	Jan. 1994	Feb. 1994	March 1994	Apr. 1994	May 1994	June 1994	...†	Jan. 1995	Feb. 1995	March 1995	Apr. 1995	May 1995
Interval													
1 (Old), 16 (New)		X	X	X									
2 (Old), 17 (New)			X	X	X								
3 (Old), 18 (New)				X	X	X							
4 (Old), 19 (New)					X	X	X						
...†								...†					
13 (Old), 28 (New)									X	X	X		
14 (Old), 29 (New)										X	X	X	
15 (Old), 30 (New)											X	X	X

† The "..." symbol is used in this table to indicate that the pattern established for the preceding months continues for the remaining months.

Figure 2. The 30 Three-Month Time Intervals Analyzed for the Revised and Original Questionnaires

Table 1
Comparison of CPS Labor Force Response Probabilities for the Original and Revised Questionnaires

True Class	Observed Class	Original (1992–1993)	Revised (1994–1995)	Original – Revised Diff	S.E.
EMP	EMP	98.68	98.84	−0.15	0.40
	UEM	0.42	0.39	0.03	0.40
	NLF	0.90	0.78	0.13	0.16
UEM	EMP	8.23	10.57	−2.34*	0.45
	UEM	79.06	73.50	5.56*	0.54
	NLF	12.71	15.93	−3.32*	0.26
NLF	EMP	2.14	1.99	0.15	0.36
	UEM	1.43	1.56	−0.13	0.33
	NLF	96.43	96.45	−0.02	0.18

* Significant at $\alpha = 0.001$.

Table 2
Comparison of Two Unemployed Subcategories for the Original and Revised Questionnaires

True Class	Observed Class	Original (1992–1993)	Revised (1994–1995)	Original – Revised Diff	S.E.
UEM–LAYOFF	EMP	16.32	26.67	−10.35*	0.91
	UEM – Layoff	61.30	55.63	5.66*	1.03
	UEM – Looking	17.61	8.41	9.20*	0.45
	NLF	4.77	9.29	−4.52*	0.28
UEM–LOOKING	EMP	7.03	7.51	−0.48	0.29
	UEM – Layoff	1.03	0.65	0.38	0.26
	UEM – Lookling	78.00	74.61	3.39*	0.21
	NLF	13.94	17.23	−3.29*	0.18

* Significant at $\alpha = 0.001$.

Table 2 shows the same set of estimates for the truly employed population only in somewhat greater detail. In this table, we considered the two primary subclassifications of unemployed: UEM-LAYOFF and UEM-LOOKING. This table provides information regarding the source difference in accuracy rates between the two questionnaire versions. We first consider the misclassification of true LAYOFF persons (top half of the table) and then consider the LOOKING persons (bottom half of the table).

For persons on layoff, classification accuracy appears to have dropped an average of 5.66 percentage points with the introduction of the revised questionnaire: from 61.30 percent to 55.63 percent. However, the patterns of classification error also changed. For the original questionnaire, the probability that a person on layoff is misclassified as looking for work is estimated at about 18 percent. The corresponding estimate for the revised questionnaire is less than half that: 8.5 percent. In addition, the data suggests that misclassification of unemployed persons on layoff as either employed or not in the labor force increased by 10.35 and 4.52 percentage points, respectively.

Now consider persons who are truly looking for work in the bottom half of Table 2. According to the MLCA model, classification accuracy for the redesigned CPS decreased significantly from 78.00 to 74.61 percent. Most of the misclassification is attributed to misclassifying persons looking for work as NLF. This result would arise, for example, if the questions regarding active and passive job search activities are prone to error. To further investigate this finding, we conducted an analysis of each of the questions used to determined the LOOKING recode. In the next section, we first consider the sources of error in the LAYOFF classification and then investigate the sources of error for the LOOKING classification.

4.2 Specific Questions Responsible for the Reduction in LAYOFF Accuracy

4.2.1 Decomposition of the LAYOFF Recode

Individuals in the CPS are classified as LAYOFF on the basis of their responses to five questions in the original questionnaire and eight questions in the revised questionnaire. These questions are listed in Figure 3. Initially, we consider which questions or combinations of questions contribute most to the error rate observed in Table 2 for the LAYOFF recoded variable and then show how MLCA models can be applied to estimate the contributions to classification error of individual questions that are used to classify an individual as LAYOFF. The methodology employed for this is similar to the MLCA approach used previously for estimating the aggregate classification error. We will describe this technique in terms of the LAYOFF classification, but it will be applied subsequently to

decompose the error in both the LAYOFF and LOOKING classification processes.

First, we combine the questions in Figure 3 using the logical operators such as "and," "or," "if-then-else," *etc.* to form a set of dichotomous "compound" questions with the property that each compound question must be answered positively in order for an individual to be classified as LAYOFF by the CPS classification process. Let $Q_k, k = 1,\ldots,K$ denote the outcomes to the K compound questions that were formed for the LAYOFF classification, where $Q_k = 1$ denotes a positive outcome and $Q_k = 2$ denotes a negative outcome. Then an individual in the CPS is classified as LAYOFF if and only if $Q_k = 1$ for $k = 1,\ldots,K$. In Figure 4, we define a set of four compound questions for original questionnaire, labeled O1–O4, and five compound questions for the revised questionnaire, labeled N1–N5.

For each classification, Q_k there is a corresponding true, unobservable (latent) classification, T_k defined in analogy to Q_k; *i.e.*, an individual is truly on layoff by the CPS definition if and only if $T_k = 1, k = 1,\ldots,K$. Next, we will use MLCA to estimate the misclassification error rates for each compound question Q_k by treating these as indicators for the unknown true latent characteristics, T_k.

The probability of an error in the classification of LAYOFF can be written as

$$\Pr(Q_k = 2 \text{ for some } k, k = 1,\ldots,K \mid T_k = 1, k = 1,\ldots,K) \quad (7)$$

which is the probability that an individual who is truly on layoff answers at least one the K compound questions negatively.

Next, we define the latent variable, W, as the number of compound questions for which the true response is positive, *i.e.*,

$$W = \begin{cases} 0 & \text{if } T_1 = 2, T_2 = 2, \ldots T_K = 2 \\ 1 & \\ \ldots etc \ldots & \text{if } T_1 = 1, T_2 = 2, \ldots T_K = 2 \\ K & \text{if } T_1 = 1, T_2 = 1, \ldots, T_K = 1. \end{cases} \quad (8)$$

For example, $W = 0$ if a person's true response pattern to the questions O1–O4 is (2,2,2,2), $W = 1$ if the true response pattern is (1,2,2,2), and so on. Note that $W = K$ corresponds to a true layoff. Thus, for the original questionnaire, $W = 0, \ldots, 4$ and for the revised questionnaire, $W = 0, \ldots, 5$.

To decomposing the probability in (7) into individual components for the compound question, Q_k, we rewrite (7) in terms of the error probabilities associated with each compound question. Thus, it can be shown that (7) can be rewritten as

$$\sum_{k=1}^{K} \Pr(Q_1 = 1, \ldots, Q_{k-1} = 1, Q_k = 2 \mid W = K). \quad (9)$$

The k^{th} term in the sum may be interpreted as the contribution of question Q_k to probability of being misclassified given a true LAYOFF.

To estimate the components of (9) using MLCA, we define a classification variable, R, which is defined in analogy to W for the observed values of Q_k,; *i.e.*,

$$R = \begin{cases} 0 & \text{if } Q_1 = 2, Q_2 = 2, \ldots, Q_K = 2 \\ 1 & \\ \ldots etc \ldots & \text{if } Q_1 = 1, Q_2 = 2, \ldots, Q_K = 2 \\ K & \text{if } Q_1 = 1, Q_2 = 1, \ldots, Q_K = 1. \end{cases} \quad (10)$$

Original Questionnaire	Question Wording
Q19	What were you doing most of LAST WEEK?
Q20	Did you do any work at all LAST WEEK not counting work around the house?
Q21	Did you have a job or business from which you were temporarily absent or on layoff LAST WEEK?
Q21A	Why were you absent from work LAST WEEK?
Q22E	Could you have taken a job LAST WEEK if one had been offered?
Revised Questionnaire	
Q20	LAST WEEK, did you do ANY work (either) for pay (or profit)?
Q20B-a	LAST WEEK, (in addition to the business,) did you have a job, either full or part time? Include any job from which you were temporarily absent.
Q20B-b	LAST WEEK, were you on layoff from a job?
Q20B-1	What was the main reason you were absent from work LAST WEEK?
Q21	Has you employer given you a date to return to work?
Q21A	Have you been given any indication that you will be recalled to work within the next 6 months?
Q21A-1	Could you have returned to work LAST WEEK if you had been recalled?
Q21A-2	Why is that?

Figure 3. Primary Components of UEM for the Original and Revised Questionnaires

Compound Question Number	Source Question(s) from the CPS Questionnaire	Compound Question Response is Positive if Source Question Response is....
Original Questionnaire		
O1	Q19: What were you doing most of LAST WEEK? or Q20: Did you do any work at all LAST WEEK not counting work aournd the house?	Q19: Any response except working and Q20: No
O2	Q21: Did you have a job or business from which you were temporarily absent or on layoff LAST WEEK?	Yes
O3	Q21A: Why were you absent from work LAST WEEK?	Temporary layoff (Under 30 days) or Indefinite layoff (30 days or more or no definite recall date)
O4	Q22E: Could you have taken a job LAST WEEK if one had been offered?	Yes
Revised Questionnaire		
N1	Q20: LAST WEEK, did you do ANY work (either) for pay (opr profit)?	No
N2	Q20B-a: LAST WEEK, (in addition to the business,) did you have a job, either full or part time? Include any job from which you were temporarily absent.	Any response except "retired," "disabled", or "unable to work"
N3	Q20B-a: LAST WEEK, were you on layoff from a job? or Q20B-1: What was the main reason you were absent from work LAST WEEK?	Q20B-b: Yes Or Q20B-1: "On layoff" or "slack work/business conditions"
N4	Q21: Has your employer given you a date to return to work? or Q21A: Have you been given any indication that you will be recalled to work within the next 6 months?	Q21: Yes or Q21: No and Q21A: Yes
N5	Q21A-1: Could you have returned to work LAST WEEK if you had been recalled? or Q21A-2: Why is that?	Q21A-1: Yes or Q21A-1: No and Q21A-2: Own temporary illness

Figure 4. Compound Questions Used in the LAYOFF Recode for Original and Revised Questionnaire Versions

Let $\pi_{k|K}^{R|W}$ denote $\Pr(R = k \mid W = K)$. Then for $k > 0$ we may write

$$\pi_{k|K}^{R|W} = \Pr(Q_1 = 1, \ldots, Q_{k-1} = 1, Q_k = 2 \mid W = K). \qquad (11)$$

Thus, the contributions to error of each LAYOFF question can be obtained from the probabilities in (11).

To estimate the probabilities $\pi_{k|K}^{R|W}$ we fit MLCA models to the same data from the 1993 and 1994 CPS as used in the previous analysis and replicated the analysis on the 1993 parallel survey data. Data from the 1992 and 1995 CPS were not part of this analysis. The MLCA models used were similar to those described in the analysis for Tables 1 and 2. That is, we used three consecutive months of data and estimated the components in (10) for 10 consecutive, overlapping intervals for each year (*i.e.*, January–March, February–April, and so on to October–December). For the original questionnaire, the model specified three latent variables corresponding to the three months within a time period, each with $K + 1 = 5$ latent classes. For the revised questionnaire, we use an identical model except each latent variable had $K + 1 = 6$ latent classes.

As before, the best MLCA model for this analysis incorporated the proxy-self grouping variable, P, and specified non-stationary transitions, equal response probabilities within time period, group heterogeneous transition probabilities, and heterogeneous response probabilities. The model provides an adequate fit to the data for all months in the analysis (*i.e.*, $p > 0.05$).

Table 3 provides a summary of the results from this analysis. In the column labeled "percent of total" we report $p_k \times 100$ percent where

$$p_k = \frac{\hat{\pi}_{k|K}^{R|W}}{\sum_{k=1}^{K} \hat{\pi}_{k|K}^{R|W}} \qquad (12)$$

is the proportion of the classification error due to compound question k in Figure 4 and where $\hat{\pi}_{k|K}^{R|W}$ are the MLCA estimates of $\pi_{k|K}^{R|W}$.

The contribution to total error presented in Table 3 (Percent of Total column) is estimated by $p_k \times \Pr(A \neq 2 \mid X = 2)$ where p_k is given by (12) and $\Pr(A \neq 2 \mid X = 2)$ is estimated from Table 2 as 1 minus the accuracy rate for LAYOFF. For the original questionnaire, the components that contribute most to LAYOFF classification error are question O2 (64.2 percent) and question O1 (27.2 percent). These two questions taken together explain more than 90 percent of the error in the LAYOFF classification.

For the revised questionnaire, estimates from the 1994 CPS indicate that more than 90 percent of the error in the LAYOFF classification arises from two components: N1 and N4.

The analysis for the revised questionnaire was repeated on the Parallel Survey with very similar results. The same two components emerge as contributing more than 90 percent of the error. As mentioned in section 2, the utility of the 1993 Parallel Survey as an indicator of data quality for the revised questionnaire is in doubt. Nevertheless, the agreement of the results from the Parallel Survey and the 1994 CPS adds strength to the findings from the 1994 CPS analysis

Thus, reduction in LAYOFF classification accuracy for the revised questionnaire appears to be due primarily to error in the responses to two compound questions: N1, the revised global question "LAST WEEK, did you do ANY work (either) for pay (or profit)?" and N4, which determines whether an individual reporting some type of layoff has a date or indication of a date to return to work. The MLCA estimates indicate that almost 60 percent of the error in the revised LAYOFF classification maybe attributed to N1 while about 34 percent may be attributed to N4.

4.2.2 Decomposition of the LOOKING Recode

The estimation process described for LAYOFF was also applied to the LOOKING recode. Note that compound question O1, O2, N1, and N2 defined in Figure 5 for LOOKING are the same questions as defined in Figure 4 for LAYOFF. Since O1, O2, and N1 appeared to be problematic for LAYOFF, we might expect that they might also be problematic for LOOKING.

Following the approach used for LAYOFF, for each survey year, we defined a latent variable, W in (8) and an indicator variable, R in (9). As we did in the LAYOFF analysis, we fit MLCA models to the data and determined that the best MLCA model for the analysis is the model incorporating the proxy-self grouping variable, P, and specifying non-stationary transitions, equal response probabilities within time period, group heterogeneous transition probabilities, and heterogeneous response probabilities. This model provides an adequate fit to the data for all months in the analysis (*i.e.*, $p > 0.05$). As before, we include the results from the Parallel Survey for comparison with the 1994 CPS results; however, the latter results will be emphasized.

Table 4 displays the values of p_k defined in (11) for the LOOKING classification. For the original questionnaire, the major contributors to classification error appear to be questions O1 and O3, which contribute 31.5 and 56.3 percent of total classification error, respectively. Question O2, which was quite problematic for the LAYOFF population, appears less so for the LOOKING population. While it contributes 64.2 percent of the LAYOFF error estimate (or 24.8 percentage points to the error rate), O2 only contributes 11.3 percent of the LOOKING error estimate (or 2.5 percentage points to the error rate).

For the revised questionnaire, the results from the analysis of the Parallel Survey and the 1994 CPS are again quite similar. The component N1 appears to be an important source of error for LOOKING as it was for the LAYOFF analysis. However, its contribution to LOOKING is smaller: 10 percentage points compared with 25 percentage points for LAYOFF. The biggest contributor to LOOKING error seems to be question N3 which contributes 64.5 percent of the error based on the CCO analysis and 51.1 percent based on the 1994 CPS analysis.

Thus, the initial labor force question appears to be problematic for both questionnaire versions. The MLCA suggests that persons who are looking for work as well as persons who are on layoff experience some difficulty responding to the question "LAST WEEK, did you do ANY work (either) for pay (or profit)?". The changes made to this question in 1994 do not appear to have improved the accuracy of this question for the either population.

Table 3
Percent Contributions to Error in LAYOFF Classifications for Compound Questions for the 1993 CPS, Parallel Survey, and the 1994 CPS

Question	1993 CPS (Original Version)		Parallel Survey (Revised Version)		1994 CPS (Revised Version)	
Old Questionnaire	Error Rate	Percent of Total	Error Rate	Percent of Total	Error Rate	Percent of Total
O1	10.53	27.20	–	–	–	–
O2	24.84	64.19	–	–	–	–
O3	2.35	6.08	–	–	–	–
O4	0.67	1.74	–	–	–	–
New Questionnaire						
N1	–	–	23.19	52.26	25.34	57.12
N2	–	–	0.00	0.00	0.00	0.00
N3	–	–	2.76	6.22	3.06	6.90
N4	–	–	18.42	41.52	15.07	33.98
N5	–	–	0.00	0.00	0.89	2.00
Total	38.39	100.00	44.37	100.00	44.37	100.00

Table 4
Percent Contributions to Error in LOOKING Classifications by Compound Questions for the 1993 CPS, Parallel Survey, and the 1994 CPS

Question	1993 CPS (Original Version)		Parallel Survey (Revised Version)		1994 CPS (Revised Version)	
Old Questionnaire	Error Rate	Percent of Total	Error Rate	Percent of Total	Error Rate	Percent of Total
O1	6.93	31.51	–	–	–	–
O2	2.49	11.34	–	–	–	–
O3	12.39	56.33	–	–	–	–
O4	0.18	0.83	–	–	–	–
New Questionnaire						
N1	–	–	8.38	33.00	10.00	39.40
N2	–	–	0.00	0.00	0.00	0.00
N3	–	–	16.38	64.5	12.97	51.08
N4	–	–	0.46	1.81	2.27	8.96
N5			0.18	0.71	0.14	0.56
Total	22.00	100.00	25.39	100.00	25.39	100.00

Compound Question Number	**Source Question(s) from the CPS Questionnaire**	**Compound Question Response is Positive if Source Question Response is....**
Old Questionnaire		
O1	Q19: What were you doing most of LAST WEEK? or Q20: Did you do any work at all LAST WEEK not counting work aournd the house?	Q19: Any response except working and Q20: No
O2	Q21: Did you have a job or business from which you were temporarily absent or on layoff LAST WEEK?	No
O3	Q22: Has ... been looking for work during the past 4 weeks? and Q22A: What has ... been doing in the last 4 weeks to find work?	Q22: Yes or response to Q19 was LK (LOOKING) and Q22A: Response other than "nothing"
O4	Q22E: Could ...have taken a job LAST WEEK if one had been offered?	Yes or No, and reason is "Already has job" or "Own temporary illness"
New Questionnaire		
N1	Q20: LAST WEEK, did you do ANY work (either) for pay (opr profit)?	Q20: No
N2	Q20B-a: LAST WEEK, (in addition to the business,) did you have a job, either full or part time? Include any job from which you were temporarily absent.	Q20B-a:No[1]
N3	Q22: Have you been doing anything to find work during the last 4 weeks?	Yes
N4	Q22A: What are all the things you have done to find work during the last 4 weeks? Or Q22A-DK: You said you have been trying to find work. How did you go about looking? And Q22A-DK1: Can you tell me more about what you did to search for work?	Mention of at least 1 active activity.
N5	LAST WEEK, could you have started a job if one had been offered?	Yes

[1] Note: In a few cases, N2 was positive if response to Q20B-a was "Disabled" or "Unable" and response to Q20A-1: "Does you disability prevent you from accepting any kind of work during the next six months?" was "No".

Figure 5. Compound Questions Used in the LOOKING Recode for Original and Revised Questionnaire Versions

The key difficulty for the LOOKING category appears to be determining whether persons who are truly looking for work have made efforts of any type (either passive or active) in the past four weeks to find work. If a respondent is classified correctly as having made some effort, the next step in the process – *viz.*, determining whether the efforts satisfy the definition of active looking – is not problematic according to the estimates in Table 4.

5. CONCLUSIONS

Biemer and Bushery (2000) provides some evidence that unemployment classification accuracy rates in the 1994 CPS redesign survey were smaller than for the original survey design used prior to 1994. This paper provides additional evidence of their findings based upon a more extensive analysis of CPS data from 1992 through 1994. Our results

indicate that the probability of correctly classifying unemployed persons decreased from 79.1 percent to 73.5 percent – a difference of 5.6 percentage points. We estimate that roughly 60 percent of the reduction (3.4 percentage points) is due to an increase in the classification error for persons on layoff while the remainder (2.2 percentage points) is due to an increase in the classification error for persons looking for work.

For the revised questionnaire, both LAYOFF and the LOOKING classifications are each based upon five compound questions. For LAYOFF, two compound questions emerged as being problematic. One is the initial labor force question, which asks "LAST WEEK, did you do ANY work (either) for pay (or profit)?" The contribution of this component to LAYOFF misclassification is estimated to be approximately 57 percent which is more than double the corresponding rate for this question in the original questionnaire. In addition, a large error rate is estimated for the compound question formed by two questions: "Has your employer given you a date to return to work?" and "Have you been given any indication that you will be recalled to work within the next 6 months?" Approximately 34 percent of the estimated LAYOFF error rate is due to this combination. Since there are no corresponding questions in the original questionnaire, most of the error in classifying persons on layoff in the revised questionnaire may be linked to these two questions.

For classifying persons who are looking for work in the redesigned survey, two questionnaire components appear to contribute most to classification error: "LAST WEEK, did you do ANY work (either) for pay (or profit)?" and "Have you been doing anything to find work during the last 4 weeks?/What has...been doing in the last 4 weeks to find work?" The error rates for both questions are slightly larger for the revised questionnaire than for the original questionnaire. These increases, therefore, explain the slight increase in LOOKING classification error observed for the revised questionnaire.

The error in CPS unemployment classification is well-documented; for example, see Chua and Fuller 1987; Abowd and Zellner 1985; Porterba and Summers 1995; and Sinclair and Gastwirth 1998. A widely accepted measure of reliability for the CPS – *viz.*, index of inconsistency computed CPS reinterview – shows the reliability of the CPS unemployment classification decreased after the redesign. Results provided in this paper are consistent with these prior studies and help determine the source of the error in the CPS classification of the unemployed. At a minimum, our results provide a basis for further investigation into the root causes of the errors in the collection of labor force data in the CPS. Through cognitive laboratory experiments and field experiments, we may identify causes of the error in the unemployment questions that would suggest ways to improve the questions. Such improvements could be implemented in a future redesign of the CPS.

ACKNOWLEDGEMENT

The author would like to acknowledge the assistance of Pamela McGovern at the U.S. Census Bureau who commented on early drafts of the paper. Appreciation is also expressed to the Associate Editor and an anonymous referee, both of whom were very helpful in preparing the article. Financial support for this research was provided by the U.S. Census Bureau.

REFERENCES

ABOWD, J, and ZELLNER, A. (1985). Estimating gross Labor-Force flows. *Journal of Business and Economic Statistics*, 3, 3, 254-283.

BIEMER, P. (2004). Modeling measurement error to identify flawed questions. In *Methods for Testing and Evaluating Survey Questionnaires*, (Eds. S. Presser, J. Rothgeb, M. Couper, J. Lessler, E. Martin, J. Marting and E. Singer), Hoboken, New Jersey: John Wiley & Sons, Inc., 225-246.

BIEMER, P., and BUSHERY, J. (2000). On the validity of markov latent class analysis for estimating classification error in labor force data. *Survey Methodology*, 26, 139-152.

BIEMER, P., and WIESEN, C. (2002). Latent class analysis of embedded repeated measurements: An application to the National Household Survey on Drug Abuse. *Journal of the Royal Statistical Society, Series A*, 165, 1.

CHUA, T.C., and FULLER, W.A. (1987). A model for multinomial response error applied to labor flows. *Journal of the American Statistical Association*, 82, 397, 46-51.

COHANY, S., POLIVKA, A. and ROTHGEB, J. (1994). Revisions Current Population Survey. Employment and Earnings BLS Report.

COHEN, J.A. (1960). Coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 37-46, 1960.

DIPPO, C., POLIVKA, A., CREIGHTON, K., KOSTANICH, D. and ROTHGEB, J. (1994). Redesigning a Questionnaire for Computer-Assisted Data Collection: The Current Population Survey Experience.

KOSTANICH, D., and CAHOON, L. (1994). CPS Bridge Team Technical Report 3: Effect of Design Differences Between the Parallel Survey and the New CPS. Unpublished Report.

MILLER, S. (1994). What Would the Unemployment Rate Have Been Had the Redesigned Current Population Survey Been in Place from September 1992 to December 1993?: A Measurement Error Analysis. CPS Bridge Team Technical Report 1.

POLIVKA, A. (1994). Comparisons of Labor Force Estimates from the Parallel Survey and the CPS During 1993: Major Labor Force Estimates. CPS Overlap Analysis Team Technical Report 1.

POTERBA, J., and SUMMERS, L. (1995). Unemployment benefits and labor market transitions: A multinomial logit model with errors in classification. *The Review of Economics and Statistics,* 77, 207-216.

POULSEN, C.S. (1982). *Latent Structure Analysis with Choice Modeling Applications.* Doctoral dissertation, Wharton School, University of Pennsylvania.

ROTHGEB, J. (1994). Revisions to the CPS Questionnaire: Effects on Data Quality, U.S. Bureau of the Census. CPS Overlap Analysis Team Technical Report 2, April 6.

SINCLAIR, M., and GASTWIRTH, J. (1998). Estimates of the Errors in classification in the Labour Force Survey and their effects on the reported unemployment rate. *Survey Methodology,* 24, 157-169.

THOMPSON, J. (1994). Mode Effects Analysis of Labor Force Estimates. CPS Overlap Analysis Teach Technical Report 3.

VAN DE POL, F., and DE LEEUW, J. (1986). A Latent Markov Model to Correct for Measurement Error. *Sociological Methods and Research,* 15, 1-2, 118-141.

VERMUNT, J. (1997). ℓ EM: A General Program for the Analysis of Categorical Data, Tilburg, University.

WIGGINS, L.M. (1973). Panel Analysis, Latent Probability Models for Attitude and Behavior Processing, Elsevier S.P.C., Amsterdam.

Comment

JEROEN K. VERMUNT [1]

1. INTRODUCTION

I enjoyed very much reading this very well written paper. The topic addressed by Paul Biemer – classification errors in the measurement of employment status – is a very important one. Employment statistics belong to the most important macro-economic indicators and, actually, we would wish they would be free of error. It, however, turns out to be impossible to measure a person's employment without error. The best that can be done is design the data collection in such a manner that the classification errors at the individual level are minimized as much as possible. The current paper contributes to this objective.

An earlier study by Biemer and Bushery (2000) indicated that the 1993 changes in the measurement procedure that intended to reduce classification errors actually increased measurement error. In the current paper, Paul Biemer replicates these former analyses with a longer time series and with an extra employment category obtained by splitting the unemployed group into "on layoff" and "looking for work". The reported results confirm the earlier conclusions that the new procedure is worse than the old procedure. In a second step, Biemer tries to disentangle the sources of measurement error for the two unemployed categories by modeling the separate questions that are used to determine whether a person is "on layoff" and "looking for work", respectively. Sources of error are identified that point at possible improvements in the questionnaire.

Because of my background, my commentary will mainly concern methodological and statistical issues. More precisely, I will discuss some methodological problem related to application of the LC Markov model, as well as indicate how the statistical analysis could be somewhat refined. It is, however, not clear whether such a more elegant modeling will yield very different conclusions. I want to stress ones more that this is a great paper. My critical remarks are only meant to stimulate the discussion.

2. LATENT CLASS MARKOV: METHODOLOGY

The main engine of the study performed by Paul Biemer is the LC of hidden Markov model. Several assumptions that may affect the encountered results have to be made when – as in this study – the model is applied with a single indicator per occasion. The assumption that is discussed in detail by Biemer is the first-order Markov process assumption. Simulation studies by Biemer and Bushery showed that, fortunately, estimates of classification error are not very sensitive to this assumption. Another assumption that is needed here for model identification is that the measurement error is constant over time. This assumption does not seem to be very problematic in the current study since we are looking for a single time-constant measure for classification error. Moreover, there is no good reason to assume that the quality of the measurement procedure changed over time while the procedure itself did not change (of course, apart from the questionnaire redesign). I am much more concerned about the third assumption; that is, the assumption of independent classification errors (ICE) over time (Bassi, Hagenaars, Croon and Vermunt 2000). Is it realistic to assume that the occurrence of a certain type of classification error at time point t does not affect the probability of making the same mistake at time point $t + 1$? In my opinion, this assumption is not realistic in the current application. For example, a respondent who makes a mistake because (s)he did not understand one of the questions will most probably (or at least be more likely than others) make the same error again at the next occasion. In my opinion, it is necessary to conduct a simulation study to determine the sensitivity of the estimated classification errors for violations of the ICE assumption.

I have another critical remark concerning the use of the LC Markov model for quantifying measurement error in a person's employment state. According to the model, there is a probabilistic relationship between an individual's true and observed states. What is, however, the true state? Is it the true employment state occupied at a particular time point, or the state that would have been recorded with an error-free or gold-standard instrument? Or is it the state a person would have occupied under "normal conditions"? That is, if also randomness in his/her behavior is filtered out.

I will illustrate my point with a small example. Suppose that there is two types (two latent segments) of coffee consumers: consumers who prefer brand A and consumers who prefer brand B, and that I belong to the brand B segment, which means that under normal circumstances I buy brand B coffee. In an interview, I am asked which

[1] Jeroen K. Vermunt, Department of Methodology and Statistics, Tilburg University, The Netherlands.

brand I bought last week. Suppose I report that I bought a brand A package of coffee, and that am neither lying nor making a mistake. In other words, there is no classification error in the sense of making a mistake: I really bought brand A this week (the researcher doesn't know that of course). On the other hand, my behavior from this week is inconsistent with my preference, which means that in terms of measurement of my preference there is a classification error. This example illustrates that there are two types of "errors" that can be made: an error in the reporting and an "error" in the behavior. The "error" in my behavior of this week may have many causes, such as "brand B was sold out", "brand A was offered at a lower price this week", "I could not find the brand B package because of changes in the arrangement of the supermarket", *etc.* The LC Markov model is not able to distinguish such randomness in the behavior that is uncorrelated across time points from real classification errors.

What does this imply for the employment application? It implies that an individual's true state may be "on layoff", but for some reason (by chance) this particular month (s)he has worked. If this "some reason" is uncorrelated with other "some reasons" for being in the "wrong" observed state at other occasions, it will be labeled classification error by the LC Markov model. While in the case of the measurement of preferences based on revealed (or stated) preferences correcting for randomness in behavior seems to be exactly what we wish to accomplish, this is clearly not the case in the measurement of employment status. I, therefore, have the strong feeling that the error rates reported by Biemer might be somewhat overestimated because of randomness in employment behavior, for instance, caused by randomness in the functioning of the labor market.

A well-known consequence of modeling individual change by means of a LC Markov model is that the estimated number of latent transitions is much smaller that the corresponding observed numbers. The reason for this is that both independent classification errors and independent random behavior is filtered out; that is, part of the observed change is attributed to these phenomena.

3. LATENT CLASS MARKOV: MODEL SPECIFICATION

Paul Biemer estimated a separate three-occasion LC Markov model for each of the 30 three–month data sets. Interview mode was used as a grouping variable in order to take into account some of the heterogeneity in the true employment distributions and classification errors. The reported error rates in the tables are averages over interview modes and rotation groups.

I would have set up the model in a somewhat more elegant and less ad hoc manner. Instead of running a separate analysis for each of the rotation groups, I would have tried to build a simultaneous model for all rotation groups. The main problem of doing a series of separate analyses is that parameters that should actually be equated across rotation groups are now estimated without constraints. For example, the employment distribution in March 1994 should be the same in the rotation groups that were interviewed between January and March, February and April, and March and May, respectively. Moreover, the transition probabilities between March and April should be the same in the February–April and March–May rotation groups. This has also implications for the Parallel Survey groups: their time-specific latent distributions and transitions should be assumed to be equal to the ones of the standard CPS. That would have been a much better manner to test whether measurement error differ between the two questionnaires. Especially for the period in which the questionnaire forms overlap, it is crucial to assume equal latent distributions in order to be able to prevent that differences in measurement error appear partially as differences in true states.

A similar problem of the separate analyses applies for the estimation of the classification errors. These are assumed to be time-constant within the 3–month period that a rotation group is interviewed, but are allowed to differ across rotation groups, even if they are interviewed in the same month. It would, of course, be much better to impose equality constraints across rotation groups. A consistent application of the time-homogeneity assumption would imply that – both for the old and the new questionnaire form – the measurement errors are constant within the full investigation period.

What we, actually, need is a LC Markov model covering all 30 months; that is, a model for 30 instead of 3 time points. Such a simultaneous model for all rotation groups is as easily specified as a model for 3 time points. Of course, for each rotation group, only 3 of the 30 months are observed, which means that the other time points have to be treated as missing values. This is not a problem in the maximum likelihood estimation of the model parameters since we can simply assume that the data are missing at random (Vermunt 1997). Questionnaire type (old/new) serves as grouping variable (in addition to interview mode) and affects the time-homogenous classification error probabilities. In other words, we estimate only two sets of classifications errors, one for the old and one for the new questionnaire. Transition probabilities may change over time, but will be equal across rotation groups interviewed at the same occasions. Moreover, the initial state probabilities of a rotation group are not estimated as separate parameters

since they are defined by the current state of the latent Markov chain.

A practical problem of the simultaneous modeling is that with so many time points it no longer possible to estimate the model parameters with the standard EM algorithm. With a variant of EM called the Baum-Welch algorithm, however, the model can also be applied with many time points (Vermunt 2003; Paas, Bijmolt and Vermunt 2003). This algorithm is implemented in an experimental version of the Latent GOLD program (Vermunt and Magidson 2000, 2003) and will be available in a next version of this program.

An alternative way to implement a simultaneous model is as a LC Markov model for 3 occasions in which rotation group serves as grouping variable and in which the relevant across rotation group equality restrictions are imposed on the classification errors, transition probabilities, and initial state probabilities. The most complicated part of this approach is that it requires the use of restrictions on marginal probabilities (Vermunt, Rodrigo and Ato-Garcia 2001). More precisely, the initial state probabilities should be in agreement with the marginal class sizes in the rotation groups that are interviewed at the same occasion.

Other aspects of the modeling that could be refined are the treatment of missing values and the coding of the interview mode. It is not necessary to eliminate cases with missing values from the analysis as is done by Paul Biemer because ML estimation with missing values is straightforward. As far as the interview mode is concerned, it would be much more elegant to work with only two categories – proxy and self – instead of four categories and let the interview mode vary across occasions within cases. In other words, interview mode could be used as a time-varying covariate. Vermunt, Langeheine and Böckenholt (1999) proposed such a latent class Markov model with time-varying covariates.

4. MODEL FOR RESPONSE PROCESS

It is a very nice idea to try to disentangle which questions in the questionnaire are causing the classification errors by modeling the response process itself. This may yield lots of valuable information for redesigning the questionnaire. I, however, think that the extended models for the employment statuses "on layoff" and "looking for work" are formulated in an overly complicated manner.

The form of the created variable R is the same as of the outcome variable in a sequential choice analysis or in a discrete-time survival analysis. Answering the next question is fully determined by whether the current one is answered positively or not. The information we have is how many steps a person takes, which is conceptually equivalent to a discrete survival time. A person "surviving" till the end is classified as being "on layoff" ("looking for work").

In my opinion, it is not very helpful to treat this variable as being generated by K latent variables (Ts). This only makes sense if theoretically there should be a response hierarchy at the latent level, which, however, because of measurement error, is not encountered at the manifest level. That is, if at the manifest level there are 2^K instead of K possible responses. Even if is the case, it often suffices to conceptualize the model as a model with a latent variable with $K + 1$ classes and K indicators, a structure that is sometimes referred to as a probabilistic Guttman model.

Paul Biemer recognizes the complexity of the K latent and K manifest variables formulation and decides to simplify the model. However, I assume because of his starting point, he decided to keep $K + 1$ latent classes. I do not see why so many latent classes are needed. There are not even so many employment states. More logical would be to have only two classes – "on layoff" and "not on layoff" ("looking for work" and "not looking for work") – since the questions are only intended to make this particular distinction. It can, of course, happen that the questions turn out to be informative about the type of "not on layoff" ("not looking for work") status, in which case an extra latent class might be needed. What is clear to me is that $K + 1$ classes are far too many.

I was wondering how many persons were classified as "on layoff" ("looking for work") at the various time points in the analysis with composite variable R as indicator. Are these numbers, as well as the number of transitions into and out off this state similar to the ones obtained with the standard four-state LC Markov model. In my opinion, this is a requisite for the validity of the calculation performed to obtain the figures presented in Tables 3 and 4.

A final thing that occurred to me is the following. Why not building a LC Markov model using the full questionnaire information as is done in the second part of the analysis. In other words, an alternative to using the observed constructed classification consisting of 4 employment categories would be to use the full set of CPS employment questions answered by the respondents. Such an analysis with multiple indicators would not only be much more informative, it would also make it possible to test and relax some of the assumptions that were made in the current analysis. For example, the ICE assumption could be relaxed for some of the questionnaire items.

REFERENCES

BASSI, F., HAGENAARS, J.A., CROON, M. and VERMUNT, J.K. (2000). Estimating true changes when categorical panel data are affected by uncorrelated and correlated classification errors. *Sociological Methods and Research*, 29, 230-268.

PAAS, L.J., BIJMOLT, T.H. and VERMUNT, J.K. (2003). Extending dynamic Segmentation with Lead Generation: A Latent Class Markov Approach. Center Paper, Tilburg University (submitted for publication).

VERMUNT, J.K. (1997). *Log-linear models for event histories.* Techniques in the Social Sciences Series, Thousand Oakes: Sage Publications. 8.

VERMUNT, J.K. (2003). Multilevel latent class models. *Sociological Methodology,* 33. In press.

VERMUNT, J.K., LANGEHEINE, R. and BÖCKENHOLT, U. (1999). Latent Markov models with time-constant and time-varying covariates. *Journal of Educational and Behavioral Statistics,* 24, 178-205.

VERMUNT, J.K., and MAGIDSON, J. (2000). *Latent GOLD User's Manual.* Boston: Statistical Innovations Inc.

VERMUNT, J.K., and MAGIDSON, J. (2003). *Addendum to Latent GOLD User's Guide: Upgrade for Version 3.0.* Boston: Statistical Innovations Inc.

VERMUNT, J.K., RODRIGO, M.F. and ATO-GARCIA, M. (2001) Modeling joint and marginal distributions in the analysis of categorical panel data. *Sociological Methods and Research,* 30, 170-196.

Comment

STEPHEN M. MILLER and ANNE E. POLIVKA [1]

1. INTRODUCTION

We are grateful for the opportunity to comment on this interesting paper. We will focus most of our comments on the empirical findings about the 1994 Current Population Survey (CPS) redesign, rather than a technical discussion of the Markov Latent Class Analysis (MLCA) methodology itself.

In his article, "An Analysis of Classification Error for the Revised Current Population Survey Employment Questions," the author applies MLCA models in an effort to trace the source of what he believes to be the "reduced accuracy of the revised classification of unemployed persons" after the redesign. In the CPS individuals are considered to be unemployed either because they are classified as being on layoff or because they are classified as looking for work. The author reports a particularly large reduction in the accuracy of the measurement of persons on layoff. Consequently, we will focus our attention on the classification of individuals on layoff, although similar comments can be made about the change in the measurement of those looking for work. In examining the accuracy of the measurement of those on layoff, the author assumes that those classified as on layoff were conceptually the same before and after the 1994 redesign, and that these individuals should exhibit the same labor force flows month-to-month. There are, however, many reasons why the improved measurement embodied in the redesign should conceptually change who is classified as on layoff. In addition, there are several factors unrelated to changes in question wording that could affect the composition of those classified as on layoff. Therefore, what the author describes as a reduction in accuracy due to the redesign more appropriately could be attributed to conceptual changes in those classified as on layoff, and the fact that what was being measured by the CPS before the redesign is not the same as what is being measured by the CPS after the redesign.

2. IMPROVED MEASUREMENT

One of the main reasons for the CPS redesign was to more accurately measure official definitions and concepts. Layoff was found to be an especially problematic concept, in that its meaning in general usage in the 1990's – a permanent job separation – was very different from the official CPS definition – a temporary job separation with the expectation of recall. When the questions were originally written in the 1940's, the term layoff was commonly used to refer to temporary spells of unemployment due to retooling or slowing of business conditions. Consequently, recall expectations were not asked about in the pre-redesign questionnaire. Research conducted in the 1980s and early 1990s in preparation for the redesign indicated that respondents' interpretation of layoff had become considerably broader than the official definition. Focus group interviews and large scale respondent debriefings found that between 30 and 50 percent of those who said they were on layoff did not expect to return to their former employers (Rothgeb 1982; Palmisano 1989; Polivka and Rothgeb 1993) Also, in 1993, 5.4 percent of those classified as on layoff had last worked 1 to 5 years ago, and another 0.6 percent had not worked in the last 5 years. This lack of recent work experience further supports the notion that many of those classified as on layoff prior to the redesign had no expectation of recall.

To better measure the official CPS definition of layoff, two questions were added in the revised questionnaire asking about individuals' recall expectations – "Has your employer given you a date to return to work?" and "Have you been given any indication that you will be recalled to work within the next 6 months?" Individuals for whom the answer is "yes" to either of these questions are classified as on layoff if they are available for work; all others are excluded from being classified as on layoff (these individuals can be classified as unemployed later in the questionnaire if they meet the active job search and availability criteria).

As a result of the addition of these direct questions, a somewhat different group of people would be expected to be classified as on layoff. Prior to the redesign, a substantial proportion, if not the majority, of individuals classified as on layoff were in fact permanently separated from their employers. After the redesign, those classified as on layoff had to expect to be recalled to their former employers; thus the vast majority of these individuals should be only temporarily separated from their employers. It is not

[1] Stephen M. Miller and Anne E. Polivka, U.S. Bureau of Labor Statistics, 2 Massachusetts Ave. NE, Room 4985, Washington, D.C. 20012.

surprising that these two groups of individuals would exhibit different month-to-month flows between labor force groups. It is reasonable to expect that individuals who expect to be recalled to their job would be more likely than those who are permanently separated to go from being temporarily on layoff to employed in consecutive months. Further, compared with permanently separated workers, those in industries in which temporary layoffs are prevalent would be more likely to be on layoff one month, employed the next month, and then laid off again.

Month-to-month gross flows of individuals between labor force states indicate that there was an increase in the proportion of the unemployed who went to employment after the 1994 redesign. Specifically, in 1994, 26.6 percent of those who were unemployed in the first month were employed in the second month, compared with 23.7 percent in 1993.

The author's MLCA estimates of a supposed decrease in the accuracy of those classified as on layoff after the redesign because more individuals are classified as employed subsequent to being on layoff, in reality is exactly in accord with what would be expected with a tightening of the definition of on layoff, and is consistent with the increase in the month-to-month gross flows between unemployment and employment (although the increased flow also is in accord with a declining unemployment rate that was observed during the time period covered by the author's study). The MLCA's smaller, but still significant, estimated decrease in accuracy due to more individuals on layoff being classified as not in the labor force after the redesign also is consistent with the tightening of the definition of on layoff through the requirement that individuals expect to be recalled in the next six months, given that individuals may adapt or change their recall expectations over time. For instance, when first interviewed, individuals may expect to be recalled in the next six months. However, in subsequent months, as the time from the initial separation increases, these individuals may no longer say that they expect to be recalled. If, at the same time, these individuals have not started searching for alternative employment, perhaps because they are still eligible to receive unemployment insurance payments, these individuals would transition to being not in the labor force. Alternatively, individuals may initially expect to be recalled; however; in subsequent months due either to poor weather conditions or a deteriorating economic situation for their former employers these individuals may become more uncertain about the probability of being recalled and thus they may not say that they expect to be recalled. If in later months, economic conditions for their former employers improve or the weather becomes less inclement, these individuals again may correctly feel that they will be recalled. The existence of changing expectations could generate a three month pattern where individuals truly were on layoff in the first month, not in the labor force the second month, and on layoff again in the third month. Those who were permanently separated from a job and were incorrectly classified as on layoff in the unredesigned survey would be unaffected by changing recall expectations. Consequently, individuals who were permanently separated from their jobs probably would be more likely to report themselves as on layoff in consecutive months with the unredesigned survey. The MLCA model would interpret this greater stability as indicating that those on layoff were more accurately measured prior to the redesign. However this greater "accuracy" would only be amongst those who were incorrectly classified because they used too broad a definition.

The author concludes that 60 percent of the misclassification of those on layoff in the redesigned survey is due to the question "LAST WEEK, did you do ANY work for pay?" This actually is consistent with more people being on temporary layoff and being recalled by their former employers in the redesigned survey (although if individuals on layoff engage in temporary employment while waiting to be recalled to their former employers, an increase in transitions to employment after 1994 may also be at least partially attributable to the broader employment question used in the redesigned survey). Similarly, the author concludes that 40 percent of the misclassification of those on layoff in the redesigned survey is due to the expectation of recall questions ("Has your employer given you a date to return to work?" and "Have you been given any indication that you will be recalled to work within the next 6 months?"). This is consistent with changing recall expectations and a slight increase in the flow between on layoff and not in the labor force. The author is obtaining different MLCA estimates of those classified as on layoff before and after the redesign because the composition of those groups has been changed, and the composition of the groups have changed in a manner that was desired and intended by those who redesigned the questionnaire.

Further evidence of the different composition of those classified as on layoff can be found in a comparison of data that were collected to determine the effect of the redesign on labor force estimates generated from the CPS. Prior to January 1994, the redesigned questionnaire was administered to 12,000 households monthly from late 1992 to December 1993. After the new questionnaire was implemented in 1994, the old questionnaire was administered monthly from January 1994 to May 1994 to 12,000 households drawn from the same sample. The experimental administration of the old and redesigned questionnaires has been referred to as the "Parallel Survey". Parallel Survey estimates from before 1994 using the new methodology and

after 1994 using the old methodology were generated to compare to official CPS estimates using the unredesigned CPS procedures prior to 1994 and the redesigned procedures after 1994. Polivka and Miller (1998) illustrate the importance of using both parts of the Parallel Survey to obtain a complete picture of the effects of the redesigned survey. For instance, if just the first part of the Parallel Survey were used, it would have been estimated that the redesign increased the unemployment rate by 0.5 percentage point. In fact, when both parts of the Parallel Survey were used, the redesign was estimated to have no statistically significant effect on the unemployment rate.

Using both parts of the Parallel Survey and the official CPS estimates, Polivka and Miller estimate that the redesigned CPS decreased the proportion of unemployed men who were on layoff by a little less than 7 percent, while it increased the proportion of unemployed women classified as on layoff by almost 7 percent (although the latter estimate was not statistically significant at a 5 percent level). These estimates imply that the redesign would decrease the proportion of those on layoff who were male and increase the proportion who were female compared to the proportions that were obtained prior to the redesign, if all else were equal. Comparison of annual averages for those over the age of 20 support this notion, since they indicate that, in 1993, 67.2 percent of those on layoff were male, compared to 63.6 percent of those on layoff in 1994 (although in addition to questionnaire changes these proportions could be affected by changes in economic conditions).

The industry distribution of those classified as on layoff, using data from both parts of the Parallel Survey and the official CPS, reveals other compositional changes in those classified as on layoff before and after the redesign. Examination of estimates from the redesigned survey to the official CPS estimates for January to May 1993 and from the unredesigned survey to official CPS estimates for January to May 1994 reveals particularly dramatic differences for those in the durable manufacturing industry. The proportion of those on layoff who were formerly employed in durable manufacturing when the unredesigned questions were used was almost half the proportion obtained when the redesigned questions were used (for January to May 1993 the proportion of those on layoff who were formerly employed in durable manufacturing averaged 16.8 percent among those who received the unredesigned questions and 9.8 percent among those who received the redesigned questions. For January to May 1994 the proportions were 8.7 percent among those who received the unredesigned questions and 15.5 percent for those who received the redesigned questions). At the same time the proportion of those on layoff who were in construction was 10 to 15 percent larger when the redesigned questions were used compared to when the unredesigned questions were used (for January to May 1993 the proportion of those on layoff who were formerly employed in the construction industry averaged 33.3 percent for those who received the redesigned questions and 27.4 percent for those who received the unredesigned questions. For January to May 1994 the proportions were 33.3 percent and 25.9 percent respectively).

Averaging the average difference between the first part of the Parallel Survey and the CPS for January 1993 to May 1993 (which is equal to the new method effect plus the Parallel Survey effect) with the average difference between the CPS and the second part of the Parallel Survey for January 1994 to May 1994 (which is equal to the new method effect minus the Parallel Survey effect) indicates that the redesign decreased the proportion of those classified as on layoff who were formerly employed in the durable manufacturing industry by 7.3 percentage points and increased the proportion classified as formerly employed in the construction industry by 3.7 percentage points (averaging the average difference between the first part of the Parallel Survey and the CPS with the average difference between the CPS and the second part of the parallel survey is in the spirit, albeit a simplified version, of the main-effects linear models estimates using generalized least squares that were presented in Polivka and Miller).

Individuals in different industries could have very different true labor force transition patterns which in turn could be influencing the MLCA estimates. For instance, given that a substantial proportion of employment in the construction industry is sensitive to weather conditions and may be more project-oriented than other types of employment, it is not unreasonable to expect that workers in construction might truly be more likely to be temporarily laid off in the first of three consecutive months, employed on a short term basis in the second month (either because the weather improved in the second month or because a short term construction project was undertaken), and then temporarily laid off again in the third month (either because weather conditions deteriorated or the project for which they were hired was completed). On the other hand, employment in the durable manufacturing industry has been steadily declining since the 1970's (for example, comparing non-recession years, it was estimated that in 1971 14.9 percent of U.S. workers as measured by BLS's establishment survey were employed in the durable manufacturing industry, compared to 9.2 percent in 1993 and 8.5 percent in 2000). This long term decline in employment makes it likely that a large proportion of workers in the manufacturing industry classified as "on layoff" prior to the redesign were permanently separated from their employers (the change in the industry distribution when the expectation of being recalled was imposed is consistent with this notion). Being

permanently separated from a job in combination with the relatively high wages workers in durable manufacturing received may increase the likelihood of these individuals being unemployed in three consecutive months, because it takes time to find employment in another industry at a similar wage.

Comparison of MLCA model estimates before and after the redesign without accounting for differences in industry composition of those classified as on layoff could cause analysts to mistakenly conclude that the redesign decreased the accuracy of labor force classifications. In reality, the increase in transitions that were measured after the redesign represented a true increase in transitions to employment after layoff was properly asked about in the CPS questionnaire. Failure to account for the fact that the redesigned CPS questionnaire intentionally classified a somewhat different group of individuals on layoff than did the unredesigned questionnaire could lead to incorrect conclusions being drawn from the MLCA models. Workers permanently separated from their employers who were classified as on layoff using the unredesigned questions are appearing to be more accurately classified in MLCA models, but they are more stable in a classification that was incorrect in the first place. Further, a proportion of individuals who are correctly classified as on layoff according to the official definition inherently could have less stable employment histories due either to their personal tastes or the industries with which they are associated.

In addition to compositional changes related to differences in question wording, the author also may have inadvertently captured in his estimates several other compositional changes unrelated to wording differences. These include differences in the time periods the author used for his estimates, as well as technological changes in the data collection process and economic conditions.

3. SEASONALITY

The first inadvertent compositional difference the author may have introduced is related to seasonality and the different time frames the author used for estimation. The number of individuals classified as on layoff in the CPS has a great deal of seasonal variability, with typically a larger number of individuals being on layoff early in the year. For instance, there were 358 individuals who were classified as on layoff in January 1995 who matched to February and March, while there were 294 individuals classified as on layoff in March 1995 who matched to April and May, and only 188 people classified as on layoff in June 1995 who matched to July and August. This means that there were 18 percent more people initially classified as on layoff in January 1995 than in March 1995 and 47 percent more individuals classified as initially on layoff in January 1995 than in June 1995. Using three month moving averages generated with the same calendar months probably would help to mitigate the effects of seasonality. However, the author did not use the same monthly time spans to generate his three-month moving averages to estimate the MLCA models before and after the redesign. The majority of the author's pre-redesign estimates were generated using data from August 1992 through December 1993, while the majority of his post-redesign estimates were generated using data from January 1994 to May 1995. Using these time spans means that the author only has, for instance, one January to March matched set of data for the pre-redesign estimates, while he has two January to March matched sets of data for the post-redesign estimates.

4. TECHNOLOGICAL CHANGES IN DATA COLLECTION

A second reason that the composition of the groups in various labor force states may be different for data collected with the unredesigned and the redesigned methodology is related to the ability to match individuals' data from month to month and the quality of these matches. The vast majority of data collected using the unredesigned methodology either in the official CPS prior to January 1993 or in the Parallel Survey from January 1994 to May 1994 were recorded using a paper form, and interviewers were required to transcribe by hand household and person identification numbers from master files to the paper survey forms. All of the data collected using the redesign methodology, either in the official CPS after January 1994 or in the Parallel Survey in 1993, were collected using an automated instrument that was loaded onto either a laptop computer or on a centralized computer. As part of the computerized data collection process, household and person identification numbers were automatically and consistently carried forward month to month. Using paper forms and transcribing data by hand has the potential to introduce errors and cause researchers to eliminate as non-matches individuals who actually are the same individuals and thus true matches.

Using the same public-use data that the author used, in combination with additional information about whether an individual had moved (that is periodically collected in the CPS), Madrian and Lefgren (1999) estimated that, depending on the stringency of the match criterion used, between 64 and 87 percent of those who were eliminated as an invalid match probably legitimately did match. Further, Madrian and Lefgren noted that there was a substantial decline between 1993 and 1996 in the fraction of invalid matches that probably should have been retained in the data set based on the criterion of whether an individual had

moved (since Madrian and Lefgren were using publicly released data, they were not able to investigate the validity of matches for 1994 to 1995 and 1995 to 1996 because the ability to match this data was suppressed to protect individuals' confidentiality). Madrian and Lefgren suggest that the increased number of valid matches for 1996 onward was due to improvements attributable to the redesign (it should be noted that, although a better match can be obtained using data internal to BLS and the Census Bureau in which information has not been suppressed, the quality of a match using internal data still will be affected by the data collection methodology. Thus the quality of the match will be better after the redesign than before the redesign). In their research, Madrian and Lefgren also found that individuals who were incorrectly excluded from the matched data sets were much more likely to be young and have their information provided by another member of the household (a proxy responder). These individuals are also the ones that Biemer argues are more likely to have classification errors in their labor force status. Consequently, by potentially including more of these individuals in his study due to the improved quality of the match, the author could be obtaining a decrease in the accuracy of his measures that he incorrectly is attributing to the questionnaire.

5. ECONOMIC CONDITIONS

Economic conditions may also contribute to differences in the composition of the groups classified as on layoff before and after the redesign. From 1992 to 1995, the period which the author uses for the majority of his MLCA modeling, the unemployment rate was steadily declining. Specifically, in 1992 the annual average unemployment rate was 7.5 percent while in 1995 it was 5.6 percent.

At a higher unemployment rate, it is likely that the proportion of individuals who remain unemployed month to month is larger than at lower unemployment rates. As the economy improves and the unemployment rate declines, it is not unreasonable to expect an increase in the proportion of individuals who transition from being on layoff to employment. With the increase in these transitions to employment, the proportion of individuals who transition to temporary jobs might also increase. Indeed, although undoubtedly related to many factors, the number of individuals employed in the temporary help supply industry (as defined under the NAICS coding system) increased 44 percent between 1992 and 1995 – from 1.1 percent to 1.5 percent of the U.S. establishments' payrolls (as measured by the BLS's establishment survey).

In addition, as the unemployment rate declines, the type of individual classified as unemployed may change. Specifically, those who remain unemployed when the unemployment rate is low tend to find it more difficult to become steadily employed and are more likely to transition quickly between labor force states. This is the logic behind studies that analyze the effects of different types of employment separations on subsequent labor force outcomes. For instance, in a study comparing individuals who were separated from their employers due to slack business conditions as opposed to complete plant shut downs, Gibbons and Katz (1991) found that, with regard to both duration of joblessness and earnings, workers who were separated from their employers due to slack business conditions did significantly worse than did those who were separated due to a plant closing. Gibbons and Katz argue that these differences were due to employers being able to dismiss their least productive workers, while retaining their more productive workers, when business conditions were slack, as opposed to employers having to dismiss both their least productive and most productive workers when a plant was completely shut down. Similarly, Darby, Haltiwanger and Plant (1985) argue that as economic conditions worsen, the duration of unemployment increases as a result of a change in the composition of those who are unemployed. This is because in more adverse economic conditions, the proportion of the unemployed who are high-skill workers (who also are less used to being unemployed and more likely to be able and willing to hold out for a more satisfactory job) will increase and the proportion of the unemployed who are less skilled and who frequently transition between labor force states will decrease.

It is important to note that the majority of the author's pre-redesign estimates were generated using 1992 and 1993 data, when the unemployment rate averaged 7.0 percent, while the majority of the redesigned estimates were generated using data from 1994 and 1995, when the unemployment rate averaged 6.0 percent. Changes in general economic conditions, and corresponding changes in the composition of the unemployed, may be affecting the supposed accuracy of the author's estimates in a way that is unrelated to the questionnaire. For instance, between 1992 and 1995, the proportion of the unemployed who were teenagers steadily increased from 14.8 percent to 18.2 percent, while the overall unemployment rate steadily declined from 7.5 percent to 5.6 percent. Similarly, the proportion of the unemployed who were Hispanic steadily increased from 13.6 percent to 15.4 percent between 1992 and 1995, though some of this may be due to the increasing proportion of Hispanics in the population (which rose from 8.8 percent to 9.4 percent). Both teenagers and Hispanics tend to be lower skilled workers who historically have been more likely to become unemployed or withdraw from the labor market. It should be noted that, regardless of the

source, an increase in the proportion of the unemployed drawn from groups with less stable labor force histories will influence the MLCA model estimates of accuracy if the change is not accounted for in the modeling.

6. DIFFERENTIAL VALIDITY OF THE MARKOV ASSUMPTIONS

In addition to differences in the composition of those classified as on layoff affecting the estimates generated by the MLCA models, differences in the composition of the various labor force groups before and after the redesign could affect the validity of the underlying assumptions of the MLCA models. As the author notes, a key assumption when implementing MLCA models is that an individual's transition from the second to third month is independent and thus uninfluenced by how the individual was classified in the first month. When estimating MLCA models for individuals' labor force states this obviously is untrue, and the validity of the assumption will likely differ amongst the various labor force categories. For instance, an individual who is employed in the first month is much more likely to be employed in the third month than is an individual who has never worked. More importantly, an individual cannot be classified as on layoff in either the redesigned or unredesigned questionnaire if he or she has not previously worked. Addition, under the official definition of layoff that was implemented in the redesign, individuals also have to expect to be recalled. This leads to a much tighter relationship between employers and workers across months using the redesigned questionnaire. Given that individuals on layoff under the redesign are much more likely to be recalled and thus employed than under the unredesigned questionnaire, the likelihood of an individual's labor force status in the third month depending on their initial labor force status in the first month is much higher. Consequently, not only is it likely that the Markov assumptions are often violated in labor force studies; it is much more likely that the Markov assumptions are violated after the redesign. This differential violation of the model's assumptions could be fundamentally influencing the author's results.

7. CONCLUSION

In summary, although the author believes that he identified a problem that was introduced into the CPS with the 1994 redesign, the supposed increase in misclassification of those on layoff in reality reflects the greater precision of the survey questions. Rather than identify a true error, we believe the author may have failed to recognize that the composition of the groups identified as on layoff before and after the redesign were different due to both intentional changes (such as the definition of on layoff being built into the questionnaire or improved quality of matches obtained because of computerization of the survey) and to uncontrolled changes such as developments in the overall economy. Finally, we would like to see further work in this area which combines the MLCA modeling approach along with a careful consideration of the economic concepts being measured, the time periods being examined and the assumptions being made. We believe this could lead to a more accurate understanding of the effects of the 1994 CPS redesign, and more useful application of the MLCA modeling approach in general.

ACKNOWLEDGEMENTS

Any opinions expressed in this paper are those of the authors and do not constitute policy of the Bureau of Labor Statistics. The authors would like to thank Sharon Cohany, U.S. Bureau of Labor Statistics, for helpful commentary on this discussion.

REFERENCES

DARBY, M.R., HALTIWANGER, J. and PLANT, M. (1985). Unemployment rate dynamics and persistent unemployment under rational expectations. *American Economic Review*, 75, 614-637.

GIBBONS, R., and KATZ, L.F. (1991). Layoffs and Lemons. *Journal of Labor Economics*, 9, 351-380.

MADRIAN, B.C., and LEFGREN, L.J. (1999). A Note on Longitudinally Matching Current Population Survey (CPS) Respondents. Technical Working Paper 247, *National Bureau of Economic Research Technical Working Paper Series.*

PALMISANO, M. (1989). Respondents' Understanding of Key Labor Force Concepts Used in the CPS. *Proceedings of the Section on Survey Research Methods,* American Statistical Association, Alexandria VA.

POLIVKA, A.E., and MILLER, S.M. (1998). The CPS After the Redesign: Refocusing the Economic Lens. In *Labor Statistics Measurement Issues,* (Eds., J. Haltiwanger, M.E. Manser and R. Topel). National Bureau of Economic Research Studies in Income and Wealth, Chicago: University of Chicago Press, 60, 249-286.

POLIVKA, A.E., and ROTHGEB, J. (1993). Overhauling the Current Population Survey: Redesigning the Questionnaire. *Monthly Labor Review,* 116, 10-28.

ROTHGEB, J. (1982). Summary Report of July Follow-up of the Unemployed. U.S. Bureau of the Census Memorandum, Washington D.C.

Comment

CLYDE TUCKER [1]

1. INTRODUCTION

I first would like to congratulate Paul Biemer for offering an innovative approach to the study of measurement error in surveys. Although he chose to illustrate his approach with the employment series in the Current Population Survey (CPS), the method can be applied to many surveys. My comments largely will be conceptual in nature, but I will supplement these comments with examples from the same data that Biemer analyzed.

Using Markov Latent Class Analysis (MLCA), the Biemer paper relies on an evaluation of the consistency over time of respondents' answers to the questions in the employment series. The increase in inconsistency found in the new series as compared to the old one, after controlling for self versus proxy reports, may serve as an indicator of one type of measurement error in the assignment of labor force status. Presumably, this error is the result of the failure of the new questions (at least, compared to the old ones) to collect the correct information for classifying an individual into the right labor force category. Thus, the error can be attributed to poor question design. Because the analysis indicates that the errors tend to be in one direction more than in the other – the misclassification of truly unemployed individuals into a different category – some might interpret the result to be a bias in the unemployment rate.

I will argue that not only has bias not been introduced but also that the new series, while certainly not perfect, reduces error, providing a more accurate picture of the employment situation. It does this by taking into account the economic realities of today in a way that the old series did not. This is accomplished by not only better question wording but also by the inclusion of follow-up questions and probes that capture more detailed information for determining a respondent's true employment status. The use of follow-up questions and probes is facilitated by the introduction of a computerized survey instrument. As a result of these innovations, I believe that the new employment series reduces the amount of specification error that existed with the old series. By specification error, I mean the error arising from using questions that do not measure what they are intended to measure. I also will explain why I do not believe that Biemer's method is appropriate for use in this particular case.

2. RECOGNITION OF THE NEED FOR A NEW EMPLOYMENT SERIES

The last major revision of the CPS prior to 1994 took place in 1967. In the ensuing years, the labor market underwent a great transformation. The number of women in the labor force dramatically increased. The number of part-time jobs and multiple job holdings escalated. The relationship between the worker and the employer became more tenuous. Startling technological developments changed the way Americans did work and resulted in the creation of new types of jobs requiring new kinds of skills. Perhaps most importantly, the economy gradually became more service oriented and less manufacturing oriented.

Just one result of these developments that needed to be taken into account in the CPS was the change in the accepted meaning of "layoff" as so ably described by Miller and Polivka (2004), but there were others, as enumerated by Bregger and Dippo (1993). Better information was needed about discouraged workers (those who have given up looking for work), multiple jobholders, marginal workers (*e.g.*, unpaid workers in a family business), and job-changing patterns. In addition, during the 1970s and 1980s, concern mounted about the various types of nonsampling errors that could be affecting CPS estimates as well as about respondent burden and its detrimental effect on data quality.

Until the 1980s, the technology to tackle these problems was not available. However, as Bregger and Dippo (pages 4–5) note, things began to change:

> "...in the early 1980s, the introduction of two new survey methodologies provided the means for understanding and reducing measurement error. These included the application of behavioral science methods and theory – more commonly referred to as the cognitive aspects of survey methodology – and computer-assisted interviewing. It is through the blending of these two methodologies that a new collection procedure, which focuses on reducing measurement error, was made possible."

Cognitive methods (including focus groups and in-depth interviewing) made it possible to develop questions that

[1] Clyde Tucker, U.S. Bureau of Labor Statistics, 2 Massachusetts Ave., NE, Room 1950, Washington, D.C. 20212.

could accurately measure the more complex economic behaviors that the times required. Furthermore, these techniques were able to uncover problems in the existing labor force series (See Polivka and Rothgeb 1993). The accurate measurement of the more complex behaviors also required a more complicated survey instrument. One so complicated that interviewers, left to their own devices, would have difficulty navigating. This is where computer-assisted interviewing played an important role. With a computerized survey instrument, interviewers could easily navigate through the complex skip patterns necessary to obtain answers to questions for measuring the wide variety of economic behaviors of interest.

3. CONSIDERATION OF NONSAMPLING ERRORS IN BOTH THE OLD AND NEW CPS EMPLOYMENT SERIES

Let me begin this section by detailing my reasons why MLCA is not an effective tool for evaluating the new CPS design relative to the old one. MLCA can be a good method for detecting measurement error within a constant series of questions by looking for inconsistencies in response over several administrations to the same respondent. In the case of the CPS, the method might be appropriate, given a careful examination of a well-chosen set of diagnostics, for examining problems in the old employment series and the new employment series independently of one another. However, let me add a caveat here about examining inconsistencies even within the same employment series. Labor force status, in itself, is inherently inconsistent over time. While the employed and not-in-the-labor-force (NILF) categories are relatively stable, the unemployed category is not. Those in that category are trying to get out. Controlling for seasonal effects by looking at March-May of either 1993 or 1994, it turns out that, on average, almost 90% of those in the employed and NILF categories did not move from one month to the next. On the other hand, over half of those in the unemployed category did. Thus, the unemployed are a particularly difficult group for MLCA to handle.

As for comparing the two series, the use of MLCA is problematic because the two series were designed to measure different things. There were some significant changes made in the employment series in the hopes of reducing specification error. Although I do not want to dwell on the measurement of layoff (Miller and Polivka have covered this topic well.), I do want to use it as a case in point for explaining why the comparison of the old and new instrument is a difficult one to make. Apart from what Miller and Polivka have said, I have my own reasons for doubting Biemer's conclusions.

The changes in the layoff questions were designed to reduce the specification error discovered in qualitative research on the meaning of "layoff," as alluded to by Miller and Polivka. In the attempt to eliminate specification error, two additional questions were added. One asked whether a date for recall had been given, and the other inquired about the possibility of returning to the job within the next 6 months. Only those who were given a recall date or expected to return to work within the 6-month period were classified as truly "on layoff."

Clearly, this altered the characteristics of the group classified as unemployed as a result of layoff as well as those asked the remaining questions in the employment series, but I believe there also were more subtle reasons why inconsistencies in respondents' answers could have increased and still not have contributed to measurement error to the extent argued by Biemer. In the first place, respondents had to answer more questions, which would have increased the probability that at least one false inconsistency would arise from one month to another. This might add to measurement error compared to the old series, but specification error, considered to be the greater problem, still would be reduced. Furthermore, false inconsistencies arising from these questions should be minimized for two reasons. These questions are much more specific than the single layoff question in the old series, and they had been well tested (Esposito, Campanelli, Rothgeb and Polivka 1991). Moreover, given that more specific questions were asked, there would be an increased chance that true change had taken place in the state of at least one of them in the intervening month. Finally, and of greatest interest to me, is the fact that these questions attempt to capture information on relatively nuanced changes. For instance, a respondent may have changed his or her mind about the possibility of being recalled in the next 6 months based on little concrete information. With the uncertainties in today's job market, it would be difficult to say that the respondent had given the wrong answer.

I now want to address Biemer's concerns about the initial question in the new employment series asking about whether any work was done last week for "either pay or profit." His results indicate that this question may be contributing to the amount of error he finds in both the "layoff" and "looking" series. The change in this question (as well as the addition of a question on the existence of a family-owned business or farm) was prompted by the concern that the old questions were not stated broadly enough, so that marginal workers, especially those working for profit at home, were not being classified as working. For example, the Parallel Survey showed the percentage of part-time workers in the new CPS was 1.098 times larger than in the old CPS, and, coincidentally, the employment to

population ratio for women 65 and older also increased by about the same amount (Polivka and Miller 1998). The same is true when comparing 1993 to 1994. It stands to reason that the increased precision in the identification of these marginal workers, who are more likely to be inconsistent in their answers from month to month than other workers, might be mistaken for measurement error. The fact is the more narrow "what were you doing last week" question could lead these respondents to consistently, but inaccurately, report they were unemployed.

Finally, let me turn to the other section of the employment series in which Biemer found a problem – the "looking for work" questions. One important change in this series involved clarifying the differences in "active" and "passive" job search in order to reduce misclassification rates in these categories. Studies conducted in the 1980s found that interviewers were confused about what constituted an active (versus a passive) job search (Polivka and Rothgeb 1993). In the redesigned questionnaire, interviewers were given an explicit list of both active and passive job search methods.

Comparisons of the results of the old and new questions are complicated by the fact that different subpopulations were asked these questions in the two series. Those finally defined as looking (and, thus, considered unemployed) in the two different employment series could have arrived there in quite different ways. Half of those considered looking in 1993 received that designation by volunteering they were looking in the first question ("What were you doing most of last week?"); none of those who were looking in 1994 followed that path. Those retired and 50 or older in 1994 never got the chance to say they were looking. In 1993, none of those who said they were on layoff were asked the looking question, so they had no chance to be classified as NILF in a given month. Then there were the two different levels of information given to the interviewers for coding active and passive methods. One difference uncovered in an analysis of the two groups from 1993 and 1994 was that a higher proportion of those looking in 1994 were women compared to 1993 (45.4% *vs.* 41.2%). Referring to the above discussion on the first employment question, increases in the inconsistency in reports to the looking questions could be the result of capturing more marginal workers using the revised employment series. Sometimes these individuals would be looking and sometimes not.

4. CONCLUSIONS

Paul Biemer has made a bold attempt to investigate the error structure in the CPS employment series; however, his findings do not take into account the reasons for the revised questions. Taking these into account would help explain the month-to-month inconsistencies that he found. Not only might these inconsistencies be real, but they could provide evidence of a reduction in specification error. For instance, controls other than for self/proxy could be included in the model to take into account some of the changes in methodology, and measurement error within more limited subpopulations. More exploration of the utility of MLCA with inherently inconsistent classifications also should be undertaken.

ACKNOWLEDGEMENTS

Any opinions expressed in this paper are those of the author and do not constitute policy of the Bureau of Labor Statistics. The author would like to thank Steve Miller, Anne Polivka, and John Dixon for their assistance on this discussion.

REFERENCES

BREGGER, J.E., and DIPPO, C.S. (1993). Overhauling the Current Population Survey: Why is it necessary to change? *Monthly Labor Review,* 116, 3-9.

ESPOSITO, J.L., CAMPANELLI, P.C., ROTHGEB, J.M. and POLIVKA, A.E. (1991). Determining which questions are best: Methodologies for evaluating survey questions. *Proceedings of the Section on Survey Research Methods,* American Statistical Association, 46-55.

MILLER, S.M., and POLIVKA, A.E. (2004). Discussion of the paper An analysis of classification error for the revised Current Population Survey employment questions. *Survey Methodology,* 30, 145-150.

POLIVKA, A.E., and MILLER, S. (1998). The CPS after the redesign: Refocusing the economic lens. In *Labor Statistics Measurement Issues,* (Eds. J. Haltiwanger, M.E. Manser, and R. Topel). National Bureau of Economic Research Studies in Income and Wealth. Chicago: University of Chicago Press, 60, 249-286.

POLIVKA, A.E., and ROTHGEB, J. (1993). Overhauling the Current Population Survey: Redesigning the questionnaire. *Monthly Labor Review,* 116, 10-28.

Response from the Author

PAUL P. BIEMER [1]

1. INTRODUCTION

My sincere thanks to all four discussants for their thoughtful, thorough and constructive comments. They have added considerably to our understanding of the complex issues surrounding Markov Latent Class Analysis (MLCA) and the Current Population Survey (CPS) labor statistics. All four discussants raise a number of important issues that I will try to address to the extent I can. Some issues will require more work and deserve much greater consideration than is possible here. More complete responses to those issues will have to await the results of future research.

Considering all the comments collectively, there seems to be agreement that Markov latent class analysis has considerable potential as a tool for evaluating and exploring the sources of measurement error in the CPS. However, there is some skepticism that it has identified real problems in the CPS questionnaire. Dr. Vermunt, who is also the author of the software I used for this analysis (*viz.*, ℓEM), provides a number of valuable suggestions for improving the models and investigating the validity of the model assumptions. The three other reviewers (Drs. Miller, Polivka, and Tucker) are quite familiar with the CPS since they are employed by the federal agency that sponsors the survey where they played important roles in the 1994 redesign. Their comments remonstrate the various ways in which the MLCA model assumptions could be violated for these data. In addition, they contain valuable information regarding details of the CPS (both pre- and post-redesign) and the construction of the CPS labor force variable. The comments and suggestions of all the discussants should be carefully considered by labor force economists and statisticians who are conducting research in the area of employment measurement error, particularly those using MLCA.

JEROEN VERMUNT'S COMMENTS

I first address the comments of Dr. Vermunt and then the comments of the other three reviewers. I share Dr. Vermunt's concern that the ICE assumption may not hold for these data. As he points out, if respondents misunderstand the labor force questions in the same way from one month to the next, they may make the same errors each month creating correlated errors across the months. As an example, a person who is truly in the UEM category at both Times 1 and 2 may be more likely to be misclassified at Time 2 if they were also misclassified at Time 1. This can be stated probabilistically as

$$\rho = \frac{P(B \neq 2 \mid A \neq 2 \text{ and } X = Y = 2)}{P(B \neq 2 \mid A = 2 \text{ and } X = Y = 2)} - 1 > 0. \qquad (1)$$

The numerator probability of the quantity ρ is the probability that the Time 2 classification (B) is in error given the Time 1 classification (A) is also in error and the true classification at both time points is UEM. The denominator probability is similar except for the condition that no error is made at Time 1 (*i.e.*, $A = 2$). Under the ICE assumption, $\rho = 0$. Therefore, if the $\rho > 0$ (which is the likely direction of the correlated error), the ICE assumption is violated. Dr. Vermunt suggests a simulation study be conducted to study the sensitivity of the estimated classification errors to violations of this assumption. Of course, determining the extent to which the ICE assumption fails for the CPS data is not possible via simulation. Nevertheless, it is still useful for assessing the potential for correlated error to bias the MLCA classification error estimates.

Following his suggestion, I conducted a small simulation study to gain some insight as to the consequences $\rho > 0$ for MLCA using CPS data. A sequence of artificial populations was generated using parameters consistent with those for the CPS (see for example, Table 1 in the main paper) except that ρ was increased in small increments from 0 to its empirical maximum – *i.e.*, the largest value of ρ that is feasible without violating the other model assumptions. Maintaining the other model assumptions in the analysis is necessary so that the consequences of violating just the ICE assumption can be isolated.

The largest feasible value of ρ was determined empirically to be 0.7. At this value of ρ, the MLCA estimate of the probability a correct classification of UEM went from 79% to 85% and the misclassification error rate dropped from 21% to 15%. For mild departures from the ICE assumption,

[1] Paul P. Biemer 3040 Cornwallis Road, PO Box 12194 Research Triangle Park, NC 27709-2194, U.S.A.

say $0 < \rho < 0.3$, the error rates changed by less than 3 percentage points. These results illustrate that if the ICE assumption fails to hold due to positive between interview correlations, the error rates estimated by MLCA will be somewhat underestimated. However, mild departures from the ICE assumption should have little effect on the classification error probabilities for these data. A similar analysis was conducted for the two other labor force categories (*i.e.*, EMP and NLF) but the change in the classification error estimates was negligible. This result was anticipated due to the relatively small error rates for these categories.

The results suggest that mild departures from the ICE assumption should have little or no effect the conclusions of the analysis. Extreme departures might affect the conclusions in the unlikely event that errors are highly correlated for original questionnaire and essentially uncorrelated for the revised questionnaire. Under that scenario, the original questionnaire would appear to have smaller UEM classification error than the revised questionnaire. However, there is no practical reason to expect this condition to hold since both questionnaires present questions that respondents may misunderstand consistently across interviews.

Although these simulation results, as well as those in Biemer and Bushery (2001) for investigating the conesquences of violations of the Markov assumption, are quite useful for studying the sensitivity of the estimates to violations of the MLCA model assumptions, they provide no direct evidence of the validity of the MLCA estimates. Biemer and Bushery (2001) illustrate how the (empirical) validity of latent class estimates can be established using external data and alternative approaches for estimating classification error. A similar analysis based upon test-retest reinterview data will be provided in the sequel.

For the purpose of identifying potential areas where the CPS questionnaire can be improved, it is not essential to establish unequivocally that the MLCA model assumptions hold since model validity is of secondary importance. Instead, the primary issue for questionnaire evaluation work is whether the method of analysis used is successful at identifying questions that have large measurement errors and are in need of revision. In other words, the validity of the model is established by its ability to find important flaws in the questionnaire. Determining whether there truly is error in the UEM classification as suggested by MLCA requires an evaluation using other methods such as cognitive laboratory research. Cognitive interviews could be used to investigate encoding, comprehension, recall, and/or social desirability issues that generate errors in the responses to the UEM questions. If these investigations uncover important problems in questions, then the utility of MLCA for identifying flawed questions will be supported even though the validity of the MLCA modeling assumptions may never be known.

Dr. Vermunt's other suggestions on ways the modeling framework could be improved are quite reasonable and I hope to investigate them further in the future. However, the current software for fitting MLCA models is somewhat limited and the estimation of complex models such as those he suggests may not be feasible. He also notes that problems can arise when fitting large models with the EM algorithm. As an example, initially we attempted to use the proxy/self-response variable as a time-varying covariate in the MLCA models, but encountered problems in the estimation process such as "division by 0" errors and persistent convergence to local maxima. We ultimately had to abandon the approach in favor of the single, time invariant proxy/self grouping variable used in the current analysis. As new and more general software becomes available, the options for MLCA with time varying covariates as well as other model enhancements mentioned by Dr. Vermunt will be feasible.

COMMENTS OF THE BLS DISCUSSANTS

I will address the comments of Drs. Miller and Polivka and those of Dr. Tucker together since the reviewers are from the same agency (BLS) and their comments raise similar concerns about the analysis. The following five points summarized their main concerns:

1. The modifications introduced in the new questionnaire capture more transitions than the old questionnaire. MLCA wrongly interprets these as errors when in fact they are not error.

2. Respondents may change their minds from month to month about whether their employers truly indicated that they might be recalled to work. These changes should not be classified as a response error.

3. The Markov assumption does not hold in labor force studies and it is violated to an even greater extent after the redesign than before the redesign. This differential violation of the model's assumptions could be fundamentally influencing the MLCA results.

4. The differences in the estimates of LAYOFF classification error before and after the redesign are due to the composition of the groups comprising this category. This composition changed after the redesign in a manner that was desired and intended by those who redesigned the questionnaire.

5. The increased inconsistency in reports to the LOOKING questions for the revised questions could be explained by more marginal workers being identified using the revised questions. Sometimes these individuals would truly be looking for work and sometimes not. MLCA misinterprets these ostensibly random changes as response error when they are not.

Point 1 describes an issue that should not pose any difficulties for MLCA. The MLCA model assumes that each individual occupies a true labor force state which may change from month to month. No assumption is made that the transition probabilities are the same for both questionnaires. The true initial labor force probabilities as well as the month-to-month transition probabilities are estimated independently for each questionnaire. In fact, although not discussed in main paper, the model estimates of the true exit probabilities for LOOKING and LAYOFF are in fact greater for the revised questionnaire than for the original questionnaire. Thus, a greater number of flows from one labor category to another for the revised questionnaire does not necessarily bias the estimates of classification error for that category in either direction.

Point 2 suggests that whether an individual is truly on layoff depends upon that individual's opinion about whether he or she was given an indication of possibly being recalled. However, this in not how the revised questionnaire defines the concept. An individual's true layoff status depends upon whether or not the employer truly provided an indication of being recalled. Although the respondent's opinion about what the employer indicated may change from month to month, the true layoff status does not change according to the respondent's opinion. Flows in and out of the LAYOFF category due to the respondent's opinion should be interpreted as error by the model.

Points 3, 4, and 5 could be made for any analysis employing MLCA. They essentially concern the potential bias in the MLCA estimates when month-to-month transitions do not behave according to the MLCA model and consequently real changes are misinterpreted as classification errors. As the reviewers note, there are at least three ways this can occur:

a) the Markov assumption does not hold (point 3),

b) there is unobserved or unexplained heterogeneity in the population (point 4), and

c) employment-related behaviors for two consecutive months are not correlated for some persons; thus, for those persons, past month status does not predict the current month's status (point 5 as well as a point made by Dr. Vermunt).

The implications of (a) were considered in a simulation analysis in Biemer and Bushery (2001). Their results suggest that, for the CPS data, the estimates of classification error are quite robust to violations of the Markov assumption. It is unlikely, then, that non-Markov transitions explain the findings of higher classification error for the revised questionnaire. Still, additional research is needed to more thoroughly understand the implications of non-Markov transitions for our results.

For (b), it is quite possible for MLCA estimates to be biased when the compositions of the unemployed populations are substantially different under the original and revised questionnaires and those differences are not explained by the grouping variables used in the model. Likewise (c) may be regarded as a special case of (b). For (c), the transition probabilities for some population subgroup are uncorrelated with the prior month's employment status; instead it is correlated with other *unobserved* variables. In Jeroen Vermunt's coffee drinker example, the unobserved variable is the availability of a specific brand of coffee at the market. At this stage of the research, we have not conducted simulation studies to quantify the effects of unobserved heterogeneity on the estimates, but this possibility will be examined in future work.

However, this issue as well as the general plausibility of the MLCA estimates can be investigated to some extent by comparing the MLCA estimates with independent estimates from an estimation approach that is not affected by (a) through (c). If the findings from the alternative analysis are consistent with the MLCA findings, the MLCA findings gain credibility. As an example, test-retest reliability for the CPS employment classifications can be estimated both pre- and post-redesign using the CPS reinterview data (see for example Biemer and Forsman 1992 for a description of CPS reinterview program and these data). The validity of the estimates of test-retest reliability does not depend upon the Markov assumption or group homogeneity assumption; the ICE assumption, however, is still relevant for reliability estimation.

Table 1 shows estimates of Cohen's kappa measure of reliability for three time periods: 1992–1993, 1995–1997, and 2002–2003. As shown in the table, the reliability of the CPS classifications of unemployment dropped after the redesign from about 68% to 65%. The most recent estimates of kappa indicate reliability has dropped to below 60%. These results are consistent with the results from the MLCA that classification error in the CPS unemployment statistics has worsened after the redesign. It is possible that the reliability estimates in Table 1 are biased since they also rely on the validity of the ICE assumption. But as discussed previously, in order to the results in the table to be explained by the failure of the ICE assumption, the ICE assumption

would have to hold for the revised questions but not for the original questions. That condition is very unlikely to occur.

Table 1
Estimates of Cohen's Kappa for the CPS Before and After the Redesign

Year	n	Cohen's κ
1992 – 1993[1]	28,063	67.8
1995 – 1997[2]	22,429	64.6
2002 – 2003[3]	19,205	58.8

[1] From Biemer and Bushery 2000.
[2] Bushery and McGovern (1999).
[3] Personal communication with Bac Tran at the U.S. Census Bureau

Given the evidence presented here and in the main paper, it seems reasonable to consider the possibility that CPS unemployment classification error increased after the redesign. The next step is to conduct additional research to evaluate these findings and explore the possible causes for the error. Rather than to focus on the validity of the MLCA or test-retest reinterview models, the focus of the future research should be the revised CPS questions, particularly those used in the LAYOFF classification.

I have already mentioned the possibility of using cognitive interviews to investigating the problems in the response process associated with the revised questions. As an example, one question identified in the MLCA as being potentially flawed is: "Have you been given any indication that you will be recalled to work within the next 6 months?" Some of the issues that could be investigated in the cognitive laboratory for this question include:

- How well do unemployed subjects understand the meanings of terms such as "any indication" and "recalled?"
- Do subjects who were recently separated from employment have difficulty remembering what their employers said about being recalled when they were terminated?
- An employer may say, "If business improves, we may call you." Do respondents answer the question correctly in this situation?
- Do respondents who initially respond that they will be recalled later change their responses to this question as the months pass by and they have not been recalled?

SPECIFICATION ERROR AND MEASUREMENT ERROR

Finally, I will address an important issue raised by Dr. Tucker regarding specification error, measurement error and their net effects. As Dr. Tucker explains, the original questionnaire suffered from specification error bias caused by measuring the wrong concept. The revisions to the labor force questions introduced in 1994 were designed to eliminate the specification error bias by refining the concepts of employment and unemployment and modifying the survey questions to reflect these refinements. These modifications, while reducing specification error, added more complexity to the survey questions which could have increased the measurement error bias in the labor force estimates. Dr. Tucker suggests that while this may be the case, the measurement bias in the new employment series may be less than the combination of specification bias and measurement bias in the old series. To determine whether this could be true, the specification error bias (B_S) and measurement error bias (B_M) were separately estimated using the MLCA estimates provided in the paper as described below.

Let p denote the CPS estimate of UEM and let P denote the expectation of p with respect to sampling and measurement error distributions. Let π denote the true value of the characteristic under the definitions of UEM implied by the specific questionnaire (*i.e.*, without regard to possible specification error). Therefore, $\pi = P - B_M$, *i.e.*, the value of P in the absence of measurement error bias.

As noted above, specification error bias is the bias in P due to a wrong concept or definition of unemployment implied by the questions and/or labor force classification process. For the revised questionnaire design, we assume that the specification error in p is 0 since it will be regarded as the gold standard for estimating the specification error bias in the original questionnaire.

Let π_{old} and π_{new} denote the π-parameter for the original and revised questionnaires, respectively. Then the specification error bias in the pre–1994 estimates of the unemployment rate is

$$B_S = \pi_{old} - \pi_{new}. \quad (2)$$

For each questionnaire, the estimate of P is p, the weighted estimate from the CPS. The estimate of π is obtained by correcting p for classification error bias using the response probabilities from the MLCA. Let $\mathbf{p}' = (p_1, p_2, p_3)$ where p_1, p_2, p_3 denote the estimates of the proportions in EMP, UEM, and NLF, respectively. Let ω_{ij} be the probability that an observation that truly belongs to the i^{th} category is assigned to the j^{th} category and let π_i denote the true proportion in the population in the i^{th} category. Then

$$E(\mathbf{p}) = \Omega'\pi \quad (3)$$

where $\pi = (\pi_1, \pi_2, \pi_3)'$ and $\Omega = [\omega_{ij}]$ is the 3×3 matrix with elements ω_{ij}. It follows that an estimator of π is

$$\hat{\pi} = (\hat{\Omega}')^{-1}\,\mathbf{p} \quad (4)$$

where $\hat{\Omega}$ is a MLCA estimate of Ω. For each questionnaire, $\hat{\Omega}$ was estimated by the average of the 10 MLCA estimates (January–March through October–December) using the 1993 CPS for the original questionnaire and 1993 Parallel Survey for the revised questionnaire.

Table 2 shows the results of this analysis. For UEM, p = 6.38 for the original and 6.98 for the revised questionnaire. If the unemployment rates are corrected for measurement bias using (4), unemployment rate increases to 7.09 percent for the original questionnaire and 8.03 percent for the revised questionnaire. Thus, an estimate of the measurement bias for the original survey is 6.38 – 7.09 = – 0.71 and for the revised survey is 6.98 – 8.03 = –1.05. Note that the measurement biases are negative for both the original and revised questionnaires, indicating that UEM as well is underestimated by both questionnaire versions.

For the revised questionnaire, the specification bias is assumed to be 0. For the original questionnaire, it is estimated by the difference 7.09 – 8.03 = – 0.94 percent. An estimate of the net bias, $B_T = B_M + B_S$, is – 0.71 + (– 0.94) = –1.65 percent for the old series compared with –1.05 + 0 = –1.05 percent for the new series. Thus, while it is subject to greater measurement error bias, the new series has smaller estimated net bias assuming $B_S = 0$.

Several limitations of these results should be mentioned. First, as noted in the main paper, the estimates for revised questionnaire from the Parallel Survey may not be representative of the revised CPS series. Second, the analysis assumes that the revised questionnaire is the gold standard for estimating the specification error bias in the original questionnaire. This assumption could also be challenged. Finally, no standard errors were provided for the estimates in Table 2 and the hypothesis of smaller overall bias in the revised question was not formally tested. Despite these limitations, the results suggest the possibility that the new unemployment series could have substantially lower net bias than the old series.

Table 2
Comparison of Original and Revised Questionnaire Biases for the CPS Unemployment Rate Based Upon Estimates from the 1993 CPS and the Parallel Survey

	p	π	B_M	B_S	B_T
1993 CPS	6.38	7.09	– 0.71	– 0.94	–1.65
Parallel Survey	6.98	8.03	–1.05	0[1]	–1.05

[1]Note: Specification error bias is assumed to be 0 for the revised questions.

REFERENCES

BIEMER, P., and BUSHERY, J. (2001). Application of markov latent class analysis to the CPS. *Survey Methodology*, 26, 2, 136-152.

BIEMER, P.P., and FORSMAN, G. (1992). On the quality of reinterview data with applications to the current population survey. *Journal of the American Statistical Association*, 87, 420, 915-923.

Survey Methodology, December 2004
Vol. 30, No. 2, pp. 159-166
Statistics Canada

A New Algorithm for the Construction of Stratum Boundaries in Skewed Populations

PATRICIA GUNNING and JANE M. HORGAN [1]

ABSTRACT

A simple and practicable algorithm for constructing stratum boundaries in such a way that the coefficients of variation are equal in each stratum is derived for positively skewed populations. The new algorithm is shown to compare favourably with the cumulative root frequency method (Dalenius and Hodges 1957) and the Lavallée and Hidiroglou (1988) approximation method for estimating the optimum stratum boundaries.

KEY WORDS: Efficiency; Geometric progression; Neyman allocation; Stratification.

1. INTRODUCTION

A stratified random sampling design is a sampling plan in which a population is divided into mutually exclusive strata, and simple random samples are drawn from each stratum independently. The essential objective of stratification is to construct strata to allow for efficient estimation. In what follows X represents the known stratification or auxiliary variable while Y represents the unknown study variable. Suppose there are L strata, containing N_h elements from which a sample of size n_h is to be chosen independently from each stratum ($1 \le h \le L$). We write $N = \sum_{h=1}^{L} N_h$ and $n = \sum_{h=1}^{L} n_h$. In the case of the stratified mean estimate,

$$\bar{y}_{st} = \sum_{h=1}^{L} \frac{N_h}{N} \bar{y}_h, \qquad (1)$$

where $\bar{y}_h$ is the mean of the sample elements in the h^{th} stratum, we need to choose the breaks in order to minimise its variance

$$V(\bar{y}_{st}) = \sum_{h=1}^{L} \left(\frac{N_h}{N}\right)^2 \left(1 - \frac{n_h}{N_h}\right) \frac{S_{yh}^2}{n_h}, \qquad (2)$$

where

$$S_{yh} = \sqrt{\sum_{i=1}^{N_h} \left(Y_{hi} - \bar{Y}_h\right)^2 / N_h},$$

is the standard deviation of Y restricted to stratum and h, and

$$\bar{Y}_h = \frac{1}{N_h} \sum_{i=1}^{N_h} Y_{hi},$$

is the mean.

Dalenius (1950) derived equations for determining boundaries when stratifying variables by size, so that (2) is minimised, but these equations proved troublesome to solve because of dependencies among the components. Since then there have been numerous attempts to obtain efficient approximations to this optimum solution. The first such approximation, suggested by Dalenius and Hodges (1957, 1959), constructs the strata by taking equal intervals on the cumulative function of the square root of the frequencies; this method is still often used today. Eckman's rule (1959) of iteratively equalising the product of stratum weights and stratum ranges was found to require arduous calculations, and is less used than the method of Dalenius and Hodges method (Nicolini 2001). Lavallée and Hidiroglou (1988) derived an iterative procedure for stratifying skewed populations into a take-all stratum and a number of take-some strata such that the sample size is minimised for a given level of reliability. Other recent contributions include Hedlin (2000) who revisited Ekman's rule, Dorfman and Valliant (2000) who compared model-based stratified sampling with balanced sampling, and Rivest (2002) who constructed a generalisation of the Lavallée and Hidiroglou algorithm by providing models accounting for the discrepancy between the stratification variable and the survey variable.

In the present paper we propose an algorithm which is much simpler to implement than any of those currently available. It is based on an observation by Cochran (1961), that with near optimum boundaries the coefficients of variation are often found to be approximately the same in all strata. He concluded however that computing and setting equal the standard deviations of the strata would be too complicated to be feasible in practice. In what follows we show that, for skewed distributions, the coefficients of variation can be approximately equalised between strata

[1] Patricia Gunning, School of Computing, Dublin City University, Dublin 9, Ireland; Jane M. Horgan, School of Computing, Dublin City University, Dublin 9, Ireland.

using the geometric progression. This new algorithm is derived in section 2. Section 3 compares the efficiency of the new approximation with the cumulative root frequency and the Lavallée and Hidiroglou approximations. We summarise our findings in section 4.

2. AN ALTERNATIVE METHOD OF STRATUM CONSTRUCTION

To stratify a population by size is to subdivide it into intervals, with endpoints $k_0 < k_1, < \ldots, < k_L$. Ideally, the division should be based on the survey variable Y. Such a construction is of course not possible since Y is unknown; if it were known we would not need to estimate it. In practice therefore we use a known auxiliary variable X, which is correlated with the survey variable.

In order to make the breaks $(k_0, k_1, \ldots, k_L)$ for any given k_0 and k_L, we seek to make the $CV_h = S_{xh} / \overline{X}_h$ the same for $h = 1, 2, \ldots, L$:

$$\frac{S_{x1}}{\overline{X}_1} = \frac{S_{x2}}{\overline{X}_2} = \ldots = \frac{S_{xL}}{\overline{X}_L}. \tag{3}$$

Now S_{xh} is the standard deviation and $\overline{X}_h$ the mean of X in stratum h: If we make the assumption that the distribution within each stratum is approximately uniformly distributed we may write

$$\overline{X}_h \approx \frac{k_h + k_{h-1}}{2}, \tag{4}$$

$$S_{xh} \approx \frac{1}{\sqrt{12}}\left(k_h - k_{h-1}\right). \tag{5}$$

As an approximation to the coefficients of variation, this gives

$$CV_h \approx \frac{(k_h - k_{h-1})/\sqrt{12}}{(k_h + k_{h-1})/2} \tag{6}$$

with equal CV_h therefore we must have

$$\frac{k_{h+1} - k_h}{k_{h+1} + k_h} = \frac{k_h - k_{h-1}}{k_h + k_{h-1}}. \tag{7}$$

This new and exotic recurrence relation reduces however to something familiar:

$$k_h^2 = k_{h+1}\, k_{h-1}; \tag{8}$$

the stratum boundaries are the terms of a geometric progression.

$$k_h = ar^h \ (h = 0, 1, \ldots, L). \tag{9}$$

Thus $a = k_0$, the minimum value of the variable, and $ar^L = k_L$, the maximum value of the variable. It follows that the constant ratio can be calculated as $r = (k_L / k_0)^{1/L}$. For a numerical example take

$$L = 4;\ k_0 = 5;\ k_4 = 50{,}000: \tag{10}$$

thus $k_h = 5.10^h (h = 0,\ 1,\ 2,\ 3,\ 4)$ and the strata form the ranges

$$5 - 50;\ 50 - 500;\ 500 - 5{,}000;\ 5{,}000 - 50{,}000. \tag{11}$$

This is clearly an extremely simple method of obtaining stratum breaks.

The relationship in (8) depends on the assumption that the distributions within strata are uniform. This may be justified by the following heuristic argument. When the parent distribution is positively skewed, then the low values of the variable have a high incidence, which decreases as the variable values increase, which makes it appropriate to take small intervals at the beginning and large intervals at the end. This is what happens with a geometric series of constant ratio greater than one. In the lower range of the variable, the strata are narrow so that an assumption of rectangular distribution in them is not unreasonable. As the value of the variable increases, the stratum width increases geometrically. This coincides with the decreased rate of change of the incidence of the positively skewed variable, so here also the assumption of uniformity is reasonable.

This algorithm will of course not work for normal distributions. Also since the boundaries increase geometrically, it will not work well with variables that have very low starting points: this will lead to too many small strata; the rule breaks down completely when the lower end point is zero. We expect the best results when the distribution is highly positively skewed and the upper part contains a small percentage of the total frequency.

3. THE PERFORMANCE OF THE ALGORITHM

3.1 Some Real Positively Skewed Populations

To test our algorithm, we implement it on four specific populations, which are skewed with positive tail:

Our first population (Population 1) is an accounting population of debtors in an Irish firm, detailed in Horgan (2003). In addition, we use three of the skewed populations that Cochran (1961) invoked to illustrate the efficiency of

the cumulative root frequency method of stratum construction. These are:

- The population in thousands of US cities (Population 2);
- The number of students in four-year US colleges (Population 3);
- The resources in millions of dollars of a large commercial bank in the US (Population 4).

There were five other populations in the Cochran paper, which turned out to be unsuitable for use with our algorithm. In three cases the variable was a proportion: agricultural loans, real estate loans and independent loans expressed as a percentage of the total amount of bank loans. Another, a population of farms in which the variable ranged from 1 to 18, was essentially discrete. Yet another, a population of income tax returns, was not sufficiently skewed: it owed its skewness to the top 0.05% of the population, and when this was removed, or put in a take-all stratum, the skewness disappeared.

These four populations are illustrated and summarised in Figure 1 and Table 1 in decreasing order of skewness.

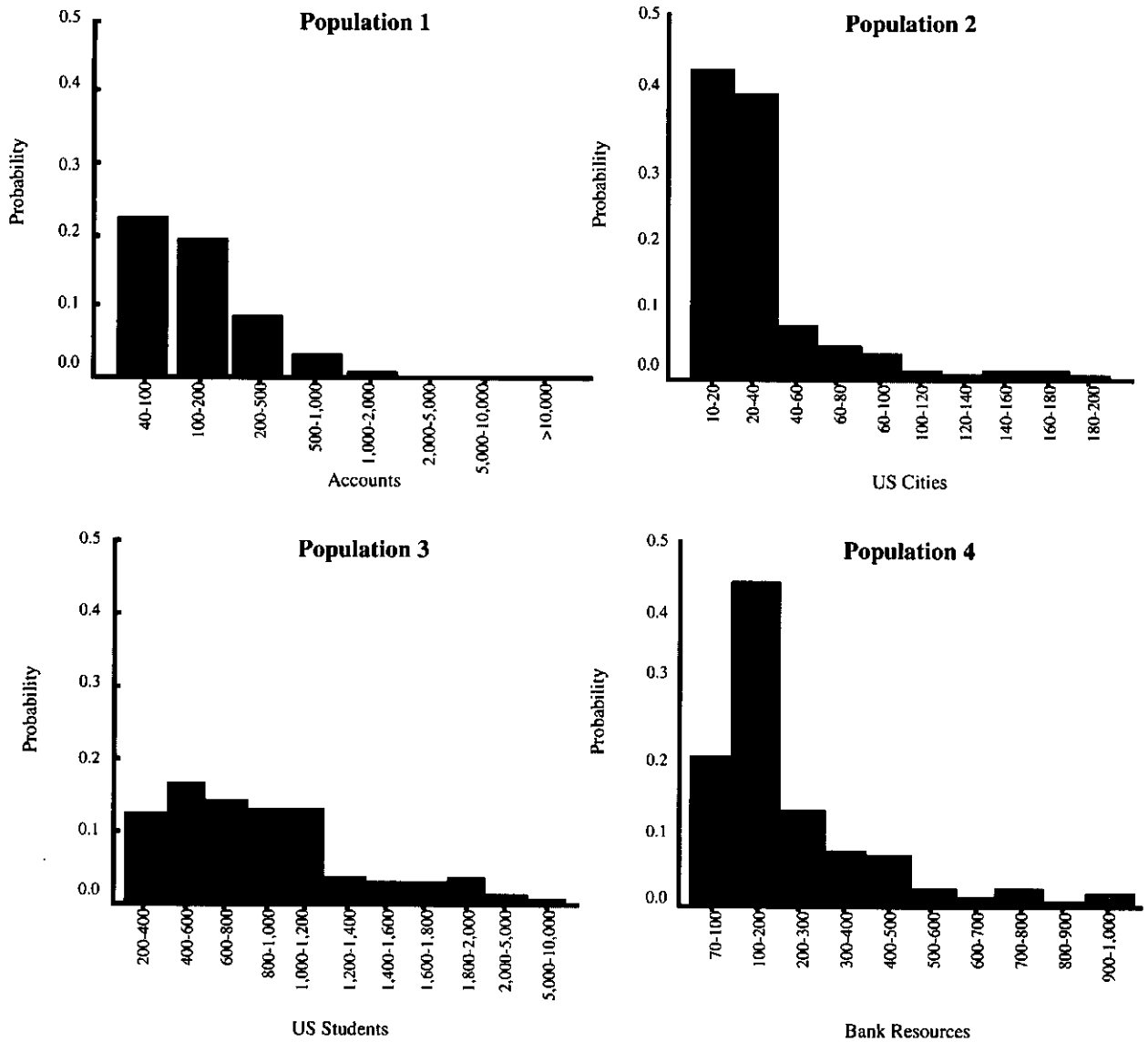


Figure 1. Populations

The new algorithm is implemented on these populations, and compared with the cumulative root frequency (cum $\sqrt{f}$) and the Lavallée-Hidiroglou methods of stratum construction.

3.2 Comparison with the Cumulative Root Frequency Method

We first compare the performance of the new algorithm with cum $\sqrt{f}$ by dividing the populations summarised in Table 1 into $L = 3$, 4 and 5 strata, using both methods to make the breaks. The results are given in Tables 2, 3 and 4.

A cursory examination of the coefficients of variation in Tables 2, 3 and 4 suggests that, in most cases, the geometric method is more successful than cum $\sqrt{f}$ in obtaining near-equal strata CV_h. For example in Population 1, which has the greatest skewness, the CV_h differ substantially from each other when cum $\sqrt{f}$ is used to make the breaks, while the geometric method appears to achieve near-equal CV_h in all cases of 3, 4 and 5 strata: the best results are obtained with $L = 5$. In the other three populations, the CV_h are not as diverse with cum $\sqrt{f}$, but they still appear more variable than those obtained with the geometric method of stratum construction.

The CV_h with the geometric method are more homogeneous when $L = 4$ or 5 than when $L = 3$; this is to be expected since the validity of the assumption of uniformity of the distribution of elements within stratum is strengthened with increased number of strata.

A more detailed analysis of the variability of the CV_h between strata is given in Table 5, where the standard deviation of the CV_h is calculated for each design.

Table 1
Summary Statistics for Real Populations

Population	N	Range	Skewness	Mean	Variance
1	3,369	40 – 28,000	6.44	838.64	3,511,827
2	1,038	10 – 200	2.88	32.57	924
3	677	200 – 10,000	2.46	1,563.00	3,236,602
4	357	70 – 1,000	2.08	225.62	36,274

Table 2
The Geometric *vs* the Cum $\sqrt{f}$: Stratum Breaks with $L = 3$ and $n = 100$

Population	Stratification Method	CV		Stratum 1	2	3
1	Geometric	0.0600	k_h	354	3,152	
			N_h	2,334	1,288	189
			n_h	9	46	45
			CV_h	0.71	0.68	0.64
	Cum $\sqrt{f}$	0.0600	k_h	558	2,236	
			N_h	2,339	735	295
			n_h	19	17	64
			CV_h	0.70	0.42	0.76
2	Geometric	0.0270	k_h	26	72	
			N_h	701	243	94
			n_h	36	29	35
			CV_h	0.28	0.23	0.33
	Cum $\sqrt{f}$	0.0269	k_h	28	66	
			N_h	729	208	101
			n_h	40	22	38
			CV_h	0.29	0.25	0.34
3	Geometric	0.0317	k_h	726	2,645	
			N_h	253	321	103
			n_h	9	38	53
			CV_h	0.32	0.37	0.39
	Cum $\sqrt{f}$	0.0282	k_h	1,179	3,629	
			N_h	456	152	69
			n_h	37	35	28
			CV_h	0.41	0.31	0.27
4	Geometric	0.0184	k_h	168	405	
			N_h	211	93	53
			n_h	27	27	46
			CV_h	0.23	0.24	0.30
	Cum $\sqrt{f}$	0.0198	k_h	162	441	
			N_h	207	107	43
			n_h	25	39	36
			CV_h	0.23	0.30	0.27

Table 3
The Geometric *vs* the Cum $\sqrt{f}$: Stratum Breaks with $L = 4$ and $n = 100$

Population	Stratification Method	CV		Stratum 1	2	3	4
1	Geometric	0.0430	k_h	205	1,057	5,443	
			N_h	1,416	1,382	483	88
			n_h	6	22	40	32
			CV_h	0.45	0.44	0.48	0.50
	Cum $\sqrt{f}$	0.0480	k_h	558	1,117	2,795	
			N_h	2,339	483	325	222
			n_h	23	5	10	62
			CV_h	0.70	0.19	0.27	0.69
2	Geometric	0.0194	k_h	20	43	93	200
			N_h	459	398	130	51
			n_h	22	31	25	22
			CV_h	0.22	0.20	0.22	0.22
	Cum $\sqrt{f}$	0.0213	k_h	19	38	85	
			N_h	393	428	155	62
			n_h	15	26	30	29
			CV_h	0.20	0.17	0.25	0.26
3	Geometric	0.0214	k_h	526	1,386	3,653	
			N_h	138	343	127	69
			n_h	5	27	26	42
			CV_h	0.27	0.26	0.26	0.27
	Cum $\sqrt{f}$	0.0230	k_h	690	2,160	5,100	
			N_h	235	319	75	48
			n_h	13	43	21	23
			CV_h	0.31	0.33	0.29	0.19
4	Geometric	0.0142	k_h	134	261	504	
			N_h	156	109	63	29
			n_h	20	23	29	28
			CV_h	0.18	0.19	0.19	0.20
	Cum $\sqrt{f}$	0.0143	k_h	162	255	488	
			N_h	207	58	57	35
			n_h	33	9	23	35
			CV_h	0.23	0.11	0.18	0.24

Table 4
The Geometric *vs* the Cum $\sqrt{f}$: Stratum Breaks with $L = 5$ and $n = 100$

Population	Stratification Method	CV		Stratum 1	2	3	4	5
1	Geometric	0.0360	k_h	147	549	2,037	7,552	
			N_h	1,054	1,267	732	265	51
			n_h	2	14	27	33	24
			CV_h	0.37	0.38	0.40	0.37	0.41
	Cum $\sqrt{f}$	0.0349	k_h	279	838	1,677	4,193	
			N_h	1,644	1,010	332	249	134
			n_h	9	14	7	15	55
			CV_h	0.52	0.30	0.20	0.25	0.57
2	Geometric	0.0144	k_h	17	32	59	108	
			N_h	364	418	130	87	39
			n_h	18	28	17	20	17
			CV_h	0.18	0.14	0.15	0.16	0.15
	Cum $\sqrt{f}$	0.0186	k_h	28	38	57	104	
			N_h	729	92	89	88	40
			n_h	58	4	7	16	15
			CV_h	0.28	0.08	0.11	0.16	0.16
3	Geometric	0.0184	k_h	433	941	2,043	4,434	
			N_h	100	255	1,989	74	56
			n_h	2	16	27	20	35
			CV_h	0.22	0.21	0.24	0.21	0.21
	Cum $\sqrt{f}$	0.0212	k_h	1,179	1,669	3,139	6,079	
			n_h	50	3	17	15	15
			CV_h	0.40	0.09	0.20	0.19	0.13
4	Geometric	0.0110	k_h	118	200	339	576	
			N_h	114	116	64	39	24
			n_h	12	20	24	18	24
			CV_h	0.14	0.14	0.17	0.12	0.16
	Cum $\sqrt{f}$	0.0119	k_h	162	255	395	627	
			N_h	207	58	37	36	19
			n_h	44	11	10	19	16
			CV_h	0.23	0.11	0.10	0.13	0.11

Table 5
The Variability of the CV_h for the Geometric and the Cum $\sqrt{f}$ Methods

Strata		Population			
		1	2	3	4
3	Geometric	0.035	0.050	0.036	0.038
	Cum $\sqrt{f}$	0.181	0.045	0.072	0.035
4	Geometric	0.027	0.010	0.006	0.008
	Cum $\sqrt{f}$	0.276	0.042	0.062	0.059
5	Geometric	0.018	0.015	0.013	0.020
	Cum $\sqrt{f}$	0.166	0.076	0.119	0.054

We see from Table 5 that, with just two exceptions, the standard deviations of the CV_h are substantially lower with the geometric method of stratum construction than with cum $\sqrt{f}$. In the two cases where the cumulative root has a lower standard deviation than the geometric, the differences between them is not great, and occur with the smallest number of strata, $L = 3$, in Populations 2 and 4. We may conclude therefore that the new algorithm is successful in breaking the strata in such a way that the CV_h are near equal.

What remains is to investigate whether the geometric breaks lead to more efficient estimation than cum $\sqrt{f}$. To do this, the two methods are compared in terms of the relative efficiency or variance ratio obtained with $n = 100$ allocated optimally among the strata using *Neyman allocation* (Neyman 1934):

$$n_h = \left(\frac{N_h S_{xh}}{\sum_{i=1}^{L} N_i S_{xi}} \right) n. \qquad (12)$$

The relative efficiency is defined as

$$eff_{\text{cum, geom}} = \frac{V_{\text{cum}}(\bar{x}_{st})}{V_{\text{geom}}(\bar{x}_{st})}, \qquad (13)$$

where $V_{\text{cum}}(\bar{x}_{st})$ and $V_{\text{geom}}(\bar{x}_{st})$ are the variances of the mean respectively with the cumulative root frequency and the geometric methods, with $n = 100$ and n_h allocated as in (12) for each of the stratification methods. In sample size planning the relative efficiencies may be interpreted as the proportionate increase or decrease in the sample size with cum $\sqrt{f}$ to obtain the same precision as that of the geometric method with $n = 100$.

The variance calculations are based on the auxiliary variable X, and since this is assumed to be highly correlated with the unknown survey variable Y, we can assume the relative efficiency eff, given in (13), will be a reasonable approximation of the relative efficiency of Y.

Table 6 gives the variance ratio when the number of strata $L = 3, 4$ and 5.

From Table 6 we see that, while this new method is not always more efficient than the cumulative root frequency method of stratum construction, when it is, it is substantially so, and when it is not it is only marginally worse. For example, large gains in efficiency are observed when $L = 5$ in Populations 2, 3 and 4: here the relative efficiencies are 1.69, 1.33 and 1.17 respectively indicating that samples of sizes $n = 169$, 133 and 117 are required with cum $\sqrt{f}$ to obtain the sample precision as that of the geometric method with $n = 100$.

Table 6
Efficiencies of the Cum $\sqrt{f}$ Relative to the Geometric Method

Strata	Population			
	1	2	3	4
3	0.97	0.99	0.79	1.16
4	1.23	1.19	1.16	1.04
5	0.94	1.69	1.33	1.17

We also see from Table 6 that while there are four cases where the relative efficiency is less than 1, with one exception, all are greater than 0.9. The exception is Population 3 with $L = 3$, the smallest number of strata; the relative efficiency in this case is 0.79.

3.3 Comparison with the Lavallée and Hidiroglou Algorithm

With the Lavallée-Hidiroglou algorithm, the optimum boundaries $k_1, k_2 \cdots k_{L-1}$ are chosen to minimise the sample size n for a given level of precision. The requirement on precision is usually stated by requiring the coefficient of variation to be equal to some specified level between 1% – 10%. Obtaining the minimum n is an iterative process, and the SAS code used for implementing it was obtained from the web at *http://www.ulval.ca/pages/lpr/*.

To compare the performance of the new method with Lavallée-Hidiroglou, the *CV*s from the geometric algorithm given in Tables 2, 3 and 4 are used as input for the Lavallée-Hidiroglou algorithm, and the sample sizes required to obtain the same precision as that of the geometric method with $n = 100$ are computed. The results are given in Table 7.

The first thing to notice from Table 7 is that the sample size required with the Lavallée-Hidiroglou algorithm to obtain the same precision as the geometric method is greater than 100 in all but four cases. In Population 2 with 5 strata, it is necessary to increase the sample size by 36% to

n = 136, to obtain the same precision as the geometric method with n = 100. With three and four strata, sample sizes of n = 121 and 113 are required in Population 1, and samples sizes of n = 123 and n = 117 are required in Population 2, to obtain the same precision as the geometric method. When the sample size falls below n = 100, the drop is not as large. In Population 4, with four and five strata, n = 93 and n = 99 respectively, and in Population 1 with 5 strata a sample size of n = 90 will suffice with the Lavallée-Hidiroglou algorithm to obtain the same precision as the geometric method.

The results in Table 7 might appear to indicate that the geometric method outperforms the Lavallée-Hidiroglou method in terms of the minimum sample size required for a specified precision. We observe however that the geometric method does not give a take-all stratum. If this is required it is more appropriate to use the Lavallée-Hidiroglou to obtain the strata. Often, in financial applications the top stratum is decided judgementally; for example US state taxing authorities typically decide their take-all stratum based on a total percentage of purchase amounts (Falk, Rotz and Young 2003). If after such a take-all stratum has been removed the skewness remains, the geometric method is probably the easier and more efficient way of obtaining the remaining strata.

Table 7
Boundaries and Sample Size Required with the Lavallée-Hidiroglou Method to Obtain the Same *CV* as the Geometric Method when n = 100

Population	n	CV		3 Strata 1	2	3		
1	121	0.0600	k_h	1,248	8,676			
			N_h	2,867	464	38		
			n_h	42	41	38		
			CV_h	0.87	0.57	0.37		
2	123	0.0270	k_h	35	102			
			N_h	795	202	41		
			n_h	47	35	41		
			CV_h	0.31	0.31	0.17		
3	107	0.0317	k_h	1,398	4,197			
			N_h	481	135	61		
			n_h	28	18	61		
			CV_h	0.41	0.30	0.24		
4	100	0.0184	k_h	172	361			
			N_h	212	85	60		
			n_h	22	18	60		
			CV_h	0.23	0.21	0.32		
				4 Strata 1	2	3	4	
1	113	0.0430	k_h	442	1,828	8,411		
			N_h	2,086	915	327	41	
			n_h	16	21	35	41	
			CV_h	0.64	0.41	0.45	38	
2	117	0.0194	k_h	19	37	95		
			N_h	393	420	176	49	
			n_h	13	21	34	49	
			CV_h	0.19	0.16	0.28	0.21	
3	103	0.0214	k_h	740	1,505	3,819		
			N_h	256	234	118	69	
			n_h	9	10	15	69	
			CV_h	0.32	0.18	0.25	0.27	
4	93	0.0142	k_h	117	188	359		
			N_h	111	112	74	60	
			n_h	7	9	17	60	
			CV_h	0.14	0.12	0.19	0.32	
				5 Strata 1	2	3	4	5
1	90	0.0360	k_h	342	1,153	3,431	10,301	
			N_h	1,846	993	357	147	26
			n_h	12	14	17	21	26
			CV_h	0.58	0.34	0.31	0.31	0.32
2	136	0.0144	k_h	14	21	35	80	
			N_h	189	270	336	164	79
			n_h	4	7	16	30	79
			CV_h	0.12	0.10	0.12	0.24	0.30
3	105	0.0184	k_h	512	869	1,577	3,675	
			N_h	133	180	185	110	69
			n_h	4	5	10	17	69
			CV_h	0.27	0.15	0.16	0.23	0.27
4	99	0.0119	k_h	99	130	189	339	
			N_h	70	68	85	71	63
			n_h	4	4	8	20	63
			CV_h	0.10	0.08	0.10	0.18	0.33

4. SUMMARY

This paper derives a simple algorithm for the construction of stratum boundaries in positively skewed populations, for which it is shown that the stratum breaks may be obtained using the geometric distribution. The proposed method is easier to implement than approximations previously proposed. Comparisons with the commonly used cumulative root frequency method using four positively skewed real populations divided into three, four and five strata, showed substantial gains in the precision of the estimator of the mean; the greatest gains occurring when the number of strata was five. Comparisons with the Lavallée-Hidiroglou method indicated that a greater sample size was required to obtain the same precision as the geometric method is most cases; the greatest increase in the required sample size occurred with the largest number of strata. One limitation of the new algorithm compared to the Lavallée-Hidiroglou method of stratum construction is that it does not determine a take-all top stratum.

ACKNOWLEDGEMENTS

This work was supported by a grant from the Irish Research Council for Science, Engineering and Technology.

We are indebted to the referees for their helpful suggestions which have greatly improved the original paper.

REFERENCES

COCHRAN, W.G. (1961). Comparison of methods for determining stratum boundaries. *Bulletin of the International Statistical Institute*, 32, 2, 345-358.

DALENIUS, T. (1950). The problem of optimum stratification. *Skandinavisk Aktuarietidskrift*, 203-213.

DALENIUS, T., and HODGES, J.L. (1957). The choice of stratification points. *Skandinavisk Aktuarietidskrift*, 198-203.

DALENIUS, T., and HODGES, J.L. (1959). Minimum variance stratification. *Journal of the American Statistical Association*, 88-101.

DORFMAN, A.H., and VALLIANT, R. (2000). Stratification by size revisited. *Journal of Official Statistics*, 16, 139-154.

ECKMAN, G. (1959). An approximation useful in univariate stratification. *The Annals of Mathematical Statistics*, 30, 219-229.

FALK, E., ROTZ, W. and YOUNG, L.L.P. (2003). Stratified sampling for sales and use tax highly skewed data-determination of the certainty stratum cut-off amount. *Proceedings of the Section on Statistical Computing*, American Statistical Association, 66-72.

HEDLIN, D. (2000). A procedure for stratification by an extended ekman rule. *Journal of Official Statistics*, 16, 15-29.

HORGAN, J.M. (2003). A list sequential sampling scheme with applications in financial auditing. *IMA Journal of Management Mathematics*, 14, 1-18.

LAVALLÉE, P., and HIDIROGLOU, M. (1988). On the stratification of skewed populations. *Survey Methodology*, 14, 33-43.

NEYMAN, J. (1934). On the two different aspects of the representative method: The method of stratified sampling and the method of purposive selection. *Journal of the Royal Statistics Society*, 97, 558-606.

NICOLINI, G. (2001). A method to define strata boundaries. Working Paper 01-2001-marzo, Departimento di Economia Politica e Aziendale, Universita degli Studi di Milano.

RIVEST, L.-P. (2002). A generalization of the Lavallée-Hidiroglou algorithm for stratification in business surveys. *Survey Methodology*, 28, 191-198.

Survey Methodology, December 2004
Vol. 30, No. 2, pp. 167-174
Statistics Canada

Feeding Back Information on Ineligibility from Sample Surveys to the Frame

DAN HEDLIN and SUOJIN WANG [1]

ABSTRACT

It is usually discovered in the data collection phase of a survey that some units in the sample are ineligible even if the frame information has indicated otherwise. For example, in many business surveys a nonnegligible proportion of the sampled units will have ceased trading since the latest update of the frame. This information may be fed back to the frame and used in subsequent surveys, thereby making forthcoming samples more efficient by avoiding sampling ineligible units. On the first of two survey occasions, we assume that all ineligible units in the sample (or set of samples) are detected and excluded from the frame. On the second occasion, a subsample of the eligible part is observed again. The subsample may be augmented with a fresh sample that will contain both eligible and ineligible units. We investigate what effect on survey estimation the process of feeding back information on ineligibility may have, and derive an expression for the bias that can occur as a result of feeding back. The focus is on estimation of the total using the common expansion estimator. An estimator that is nearly unbiased in the presence of feed back is obtained. This estimator relies on consistent estimates of the number of eligible and ineligible units in the population being available.

KEY WORDS: Dead unit; Feed back bias; Overcoverage; Permanent random number sampling; Panel survey; Coordinated samples.

1. INTRODUCTION

To facilitate estimation of change, consecutive samples in a repeated survey are usually overlapping. If several surveys draw samples from the same frame, it is often desirable to spread the response burden out by making sure that samples for different surveys are not overlapping to a greater extent than necessary. This is particularly desirable if the frame is moderately large and used for many continuing surveys, which is a situation that many national statistical institutes face when conducting business surveys. Stratified simple random sampling is a very common design for business surveys. The skewed distribution of businesses calls for large sampling fractions in many strata, which aggravates the response burden for medium size and large businesses. Both estimation of change and response burden issues are of paramount importance in official business statistics. Therefore, sampling systems have been constructed that allow the organisation to co-ordinate samples, either positively or negatively (*i.e.* to create overlap or to make sure that there is little overlap).

For example, the Office for National Statistics (ONS) in the United Kingdom uses the Permanent Random Number (PRN) technique, which is a widely used method for drawing samples from lists. A PRN from the uniform distribution on [0,1] is attached to each frame unit independently of each other and independently of the unit labels and any variables associated with the units. Each unit will retain the PRN throughout its existence. The units can be ordered along a line starting at 0 and ending at 1 and we refer to this line as the *PRN line*. To draw a simple random sample without replacement, an *SI*, with a predetermined sample size n, a point is selected (randomly or purposively) on the PRN line and the n units to the right (say) are included in the sample. Two SIs are fully co-ordinated if they are drawn from the same interval. For overviews and further details see Ohlsson (1995) and Ernst, Valliant and Casady (2000).

Samples for repeated surveys can also be selected with a panel technique where a set of rotation groups are selected at the first wave and one, say, of the groups is replaced with a fresh rotation group at the second wave and the other groups are retained in the sample. The difference between PRN sampling and panel sampling is more about the way to control overlaps than having different sampling designs.

There are in principle two main sources of data that are used to maintain a frame: administrative ones and surveys. Various administrative bodies send tapes to the ONS on a regular basis with information on, *e.g.*, births and deaths of businesses. While these tapes are sent to the ONS very frequently, the distribution of the time it takes for a new unit or an alteration of an old unit to be registered on the frame is highly skewed. This is partly due to frame maintenance procedures, *e.g.* to avoid duplicates. There is also very often a considerable difference in time between the actual and formal termination of a business. Therefore, most of the ONS's business surveys share the information on deaths

[1] Dan Hedlin, Statistics Sweden, Box 24 300, SE-104 51 Stockholm, Sweden. E-mail: dan.hedlin@scb.se; Suojin Wang, Texas A&M University, Department of Statistics, College Station, Texas 77843-3143, U.S.A. E-mail: sjwang@stat.tamu.edu.

they obtain through their samples with other business surveys to speed up the information process. We examine the effects of using sample surveys to update a frame that is used for repeated surveys. This is in principle how information on dead units is treated in business surveys at the ONS, Statistics Sweden, and some other national statistical institutes.

It would seem natural that this new information should be made available to other sample surveys, which otherwise may include the dead units in their samples and therefore lose precision. However, as pointed out by Srinath (1987) among others, such a procedure may cause bias. We refer to this as *feed back bias*, which results whenever the sampling mechanism is not independent of the feed back procedure. For example, consider a situation where all dead units are found and deleted at the first wave of a panel survey. If no further deaths have occurred up to the second-wave observation of the panel units, the second-wave sample contains only live units. Without knowledge of the total number of live units in the population at the time of the second wave, an unbiased estimator of the total cannot be constructed. While more information about the population has been gathered when the deaths were recorded at the first wave, there is actually less information in the second wave-sample on the proportion of live units in the population. We show how an estimate of the number of live units in the population can be used to construct an approximately unbiased estimate of the population total.

A safe recommendation would be that no information on deaths from sample surveys, other than from completely enumerated strata, may be used to update the frame when samples are co-ordinated over time (*cf.* Ohlsson 1995, page 168, and Colledge 1989, page 103). However, to prohibit feeding back seems to deny oneself the use of all available information. We obtain an expression for the feed back bias and show that the feed back bias can be estimated and used to adjust conventional estimators. Schiopu-Kratina and Srinath (1991) adjust the sampling weights to counter an expected too low proportion of dead units in the rotating sample of the Survey of Employment, Payroll and Hours conducted by Statistics Canada. Hidiroglou and Laniel (2001) discuss the feed back issue briefly. A general discussion of frame issues is given by Colledge (1995) and overviews of issues associated with continuing business surveys include College (1989), Hidiroglou and Srinath (1993), Srinath and Carpenter (1995), and Hidiroglou and Laniel (2001).

Instead of the terms eligible and ineligible we use the more emotive words dead and live, although our reasoning does cover all kinds of ineligibility. The discussion is confined to the estimation of the total

$$t_y = \sum_U y_k \qquad (1)$$

of some study variable $\mathbf{y}' = (y_1, y_2, ..., y_N)$ on a population U with unit labels $\{1, 2, ..., N\}$.

When the sampled units are observed, we assume that *all* dead units in the sample are classified as dead and the frame is updated with this information. This may be difficult in practice. In some surveys, however, the eligibility of all nonresponding units can be correctly identified.

Section 2 introduces the necessary notation and concepts and gives expressions for the feed back bias when estimating a total. Section 3 discusses three strategies that may be used in the presence of feed back and compares these in a simulation study. The paper concludes with a discussion in section 4.

2. EXPRESSIONS FOR FEED BACK BIAS

2.1 Introduction and Notation

We assume throughout that a dead unit is always out of scope and that the value of the study variable of a dead unit is always zero. (It is conceivable that dead units are eligible in some surveys; for example, a business survey collecting data on production may have defined businesses that were alive at least part of the reference period as eligible.) We adopt the design-based view that the survey population and the study variable are fixed and non-stochastic at any given point in time. The situation we address is as follows. One or more samples are drawn from the frame which comprises the *original survey population*, U_1. Let the set of samples drawn from U_1 be denoted by s_1. For convenience we assume that the frame units and population units are of the same type. We refer to the updated frame, where all dead units that have been included in samples from U_1 have been excluded, as the *current survey population*, U_2. For example, two surveys may simultaneously work with a sample each, and after they have fed back, U_1 has shrunk to U_2. We disregard births of new units and other deaths than those deleted through samples from U_1. We will also disregard undercoverage, nonresponse and measurement errors. In practice, administrative sources will provide information on deaths. They work independently from the sampling procedures employed by the statistical agency and will therefore not contribute to feed back bias. These units are *dead by administrative sources*. We can think of these dead units as being excluded from the population. See Hidiroglou and Laniel (2001) for a discussion of estimation in the presence of units deathed by administrative sources. While the sampling design here is assumed to be SI, it can readily be extended to stratified simple random sampling.

Let $U_{2,d}$ and $U_{2,l}$ be the two subsets of the current survey population, $U_2 = U_{2,d} \cup U_{2,l}$, that consist of dead and live units, respectively. All units in $U_{2,d}$ and $U_{2,l}$ are assumed to be flagged as live. Units that are flagged as dead but for which the independence of detection and the sampling mechanism cannot be assured are called *dead by sample survey sources*. In our set-up, these are the dead units detected in samples taken from U_1. Let the set of these units be denoted by $s_{1,d}$, and we have the relationship $U_1 = U_2 \cup s_{1,d}$. Figure 1 displays the sets and their relationships. Let N and n with a proper subscript be the size of the corresponding population and sample(s), respectively. Then $N_1 = N_2 + n_{1,d}$ and $N_2 = N_{2,l} + N_{2,d}$. At the time when samples are drawn from U_2, N_2 and $n_{1,d}$ are known numbers, whereas $N_{2,l}$ and $N_{2,d}$ are unknown. Moreover, $n_{1,d}$, $N_{2,d}$ and N_2 could be viewed as random depending on feed back results, while $N_{2,l}$ is fixed. Following principles of Durbin (1969) and more recently in Thompson (1997), we would in many situations prefer to condition on $n_{1,d}$. For example, if it is seen that $n_{1,d} = 0$, then it does not seem appropriate to include in the inference the possibility that $n_{1,d}$ could have been large. However, to analyse the development of the feed back bias over a series of waves in a panel survey when planning the survey, unconditional analysis would be preferable. We also provide an expression for the unconditional feed back bias.

Denote by $s_{1,l}$ the live part of s_1, *i.e.*, the part of U_2 that was covered by the previous sample(s) drawn from U_1; see Figure 1. Clearly, $s_{1,l}$ is a random set and we have $s_{1,l} \subset U_{2,l}$. Let the nonsampled part of U_2 be denoted by $U_{2,wd}$ ('wd' for 'with dead units'). It is also a random set and encompasses all of $U_{2,d}$ and part of $U_{2,l}$. We have $U_2 = U_{2,wd} \cup s_{1,l}$.

Let s_2 be an SI taken from U_2. Estimators based on s_2 will suffer from feed back bias unless special information is at hand, such as knowledge about $N_{2,l}$, which is not usually the case. To derive an expression for the feed back bias we shall first obtain the inclusion probabilities. To do this, it is useful to consider the two sample parts of s_2 separately: the sample part $s_{2,a}$ of size $n_{2,a}$ taken from $s_{1,l}$ through PRN sampling or a panel sampling technique, and the remaining part $s_{2,b}$ taken from $U_{2,wd}$. If the sampling is done with a panel technique, the sample parts $s_{2,a}$ and $s_{2,b}$ are the old and new rotation groups, respectively. If the sample is drawn with PRN sampling, $s_{2,a}$ and $s_{2,b}$ consist of units with PRN's that fell in s_1 or did not fall in s_1, respectively. Whether the sample was drawn through PRN sampling or a panel sampling technique, the sample parts can be viewed as two fixed size samples, each drawn with the SI design from their respective subpopulation. We condition on $n_{2,a}$ and $n_{2,b}$ throughout without making it explicit in formulae. With the notation $(k \in s_{2,a})$ we refer to the event that a unit is first included in the first-wave sample(s) from U_1 and then in the second-wave sample taken from what remains of the first-wave sample(s) after dead units have been taken out. The notation $(k \in s_{2,b})$ is analogous. Let $I(k \in s_{2,a}) = 1$ when unit k is included in $s_{2,a}$, otherwise $I(k \in s_{2,a}) = 0$. To derive the overall bias it is convenient to analyse the biases from the sample parts $s_{2,a}$ and $s_{2,b}$. We derive an expression for each of these in section 2.2 and section 2.3, respectively, and in section 2.4 the bias expressions will be amalgamated.

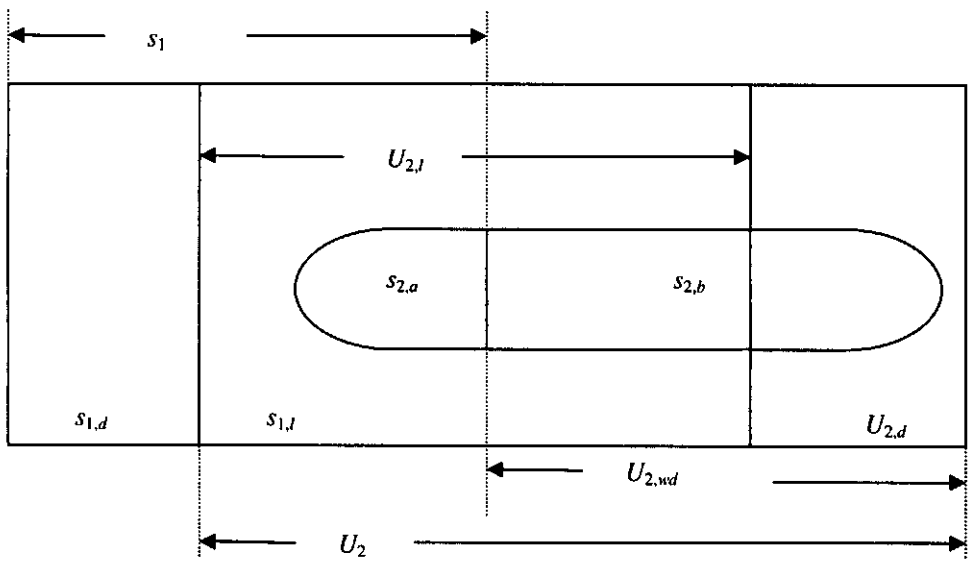


Figure 1. The original survey population, U_1, and its subsets. The grey area represents s_2, the sample from U_2.

2.2 Feed Back Bias from a Sub-sample from the Original Sample

Suppose a sub-sample $s_{2,a}$ is taken from $s_{1,l}$, the live part of the first-wave sample(s). Recall that $y_k = 0$ if k is a dead unit and that $U_2 = U_{2,d} \cup U_{2,l}$. Thus we have $\sum_{s_{2,a}} y_k = \sum_{U_{2,l}} y_k I(k \in s_{2,a}) = \sum_{U_2} y_k I(k \in s_{2,a})$. Assume that $N_{2,l} > 0$. Then we obtain that $\Pr[k \in s_{2,a} \mid n_{1,d}] = n_{2,a} / N_{2,l}$, since a sample of size $n_{2,a}$ is effectively selected from a population of size $N_{2,l}$ with the SI design (through an SI sample from U_1 followed by an SI sample from $U_{2,l}$). Note that a unit k in $s_{2,a}$ must be alive since $U_{2,l}$ consists solely of live units.

Denote the bias of an estimator $\hat{\theta}$ for the parameter θ by $B(\hat{\theta}, \theta)$. Then with respect to the population total $t_y = \sum_{U_2} y_k$, the conditional bias of a general linear estimator $\hat{t}_y^{(s_{2,a})} = \sum_{s_{2,a}} w_k y_k$ based on $s_{2,a}$, with any given w_k's, is

$$\begin{aligned} B\left(\hat{t}_y^{(s_{2,a})}, t_y \middle| n_{1,d}\right) &= \sum_{U_{2,l}} \left\{ w_k \Pr\left[k \in s_{2,a} \middle| n_{1,d}\right] - 1 \right\} y_k \\ &= \sum_{U_{2,l}} \left(\frac{w_k n_{2,a}}{N_{2,l}} - 1 \right) y_k \\ &= \sum_{U_2} \left(\frac{w_k n_{2,a}}{N_{2,l}} - 1 \right) y_k . \end{aligned} \tag{2}$$

For the sample part $s_{2,a}$, the naive expansion estimator that ignores feed back bias would have weights $w_k = N_2/n_{2,a}$. From (2) the bias of the estimator $\hat{t}_{y\pi}^{(s_{2,a})} = N_2 / n_{2,a} \sum_{s_{2,a}} y_k$ is

$$B\left(\hat{t}_{y\pi}^{(s_{2,a})}, t_y \middle| n_{1,d} \right) = \frac{N_{2,d}}{N_{2,l}} t_y . \tag{3}$$

2.3 Feed Back Bias from a Sample Taken Afresh from the Current Survey Population

Next, we derive the bias arising from the sample part $s_{2,b}$ of size $n_{2,b}$ taken from U_2 through U_{2wd}, see Figure 1. First note that

$$\Pr\left[k \in s_{2,b} \middle| k \in U_{2,\mathrm{wd}},\ n_{1,d}\right] = \frac{n_{2,b}}{N_{2,wd}} . \tag{4}$$

From (4) we obtain that the conditional expected value of $\hat{t}_y^{(s_{2,b})} = \sum_{s_{2,b}} w_k y_k$ is

$$\begin{aligned} E\left(\hat{t}_y^{(s_{2,b})} \middle| n_{1,d}\right) &= E\left[\frac{n_{2,b}}{N_{2,wd}} \sum_{U_{2,wd}} w_k y_k \ \middle| n_{1,d} \right] \\ &= \frac{n_{2,b}}{N_{2,wd}} \frac{N_{2,l} - n_{1,l}}{N_{2,l}} \sum_{U_2} w_k y_k . \end{aligned}$$

The second equation above is due to the fact that given $n_{1,d}$, all $N_{2,l}$ live units in U_2 are equally likely to be in $U_{2,wd}$, which has $N_{2,l} - n_{1,l}$ live units. Therefore, the conditional bias of $\hat{t}_y^{(s_{2,b})}$ is

$$B\left(\hat{t}_y^{(s_{2,b})}, t_y \middle| n_{1,d}\right) = \sum_{U_2} \left(\frac{w_k n_{2,b}}{N_{2,wd}} \frac{N_{2,l} - n_{1,l}}{N_{2,l}} - 1 \right) y_k . \tag{5}$$

For the expansion estimator $\hat{t}_{y\pi}^{(s_{2,b})}$ with weights $w_k = N_2 / n_{2,b}$ the bias is

$$B\left(\hat{t}_{y\pi}^{(s_{2,b})}, t_y \middle| n_{1,d}\right) = B t_y , \tag{6}$$

where

$$\begin{aligned} B &= \frac{N_2}{N_{2,wd}} \frac{N_{2,l} - n_{1,l}}{N_{2,l}} - 1 \\ &= \frac{N_2 \left(N_{2,l} - n_{1,l}\right) - N_{2,l} \left(N_2 - n_{1,l}\right)}{N_{2,wd} N_{2,l}} \\ &= -\frac{N_{2,d} n_{1,l}}{N_{2,l} N_{2,wd}} \\ &= -\frac{N_{2,d} \left(n_1 - n_{1,d}\right)}{N_{2,l} \left(N_1 - n_1\right)} . \end{aligned}$$

The bias is always non-positive since $B \le 0$. It is easy to see that B is an increasing function of $n_{1,d}$ since $N_{2,d} = N_{1,d} - n_{1,d}$, where $N_{1,d}$ is the fixed number of all dead units in U_1. It is also readily seen that the maximum of B is attained when $s_{1,d}$ encompasses all dead units in U_1, that is, when $n_{1,d} = N_{1,d}$ and consequently $N_{2,d} = 0$.

2.4 Feed Back Bias from Sample Parts Combined

Combining (6) with (3) we obtain the overall bias of $\hat{t}_{y\pi} = N_2 / n_2 \sum_{s_2} y_k$ to be

$$\begin{aligned} B\left(\hat{t}_{y\pi}, t_y \middle| n_{1,d}\right) &= E(\hat{t}_{y\pi} \mid n_{1,d}) - t_y \\ &= \frac{N_{2,d}}{N_{2,l}} \left(\frac{n_{2,a}}{n_2} - \frac{n_{2,b}}{n_2} \frac{n_{1,l}}{N_{2,wd}} \right) t_y = \tilde{c} t_y . \end{aligned} \tag{7}$$

The bias in the expansion estimator is really down to not knowing the correct population size. In (3) the bias stems from multiplying the sample average over live units with N_2 rather than the unknown $N_{2,l}$. The bias from the sample parts $s_{2,a}$ and $s_{2,b}$ will in absolute terms be less than (3) and (6), respectively, if some of the dead units in the samples from U_1 have not been identified as dead and therefore have not been weeded out. This would happen, for example, if the status of nonresponding units is difficult to determine.

An unconditional analysis in the presence of feed back can be obtained directly by taking expectation of (7) with respect to $n_{1,d}$. Thus, unconditionally, we have

$$E\left(\frac{N_2}{n_2}\sum_{s_2} y_k\right) - t_y$$
$$= \left[\begin{array}{l} \dfrac{N_{1,d} - E(n_{1,d})}{N_{2,l}}\left(\dfrac{n_{2,a}}{n_2} - \dfrac{n_{2,b}}{n_2}\dfrac{n_1 - E(n_{1,d})}{N_{2,wd}}\right) \\ -\dfrac{n_{2,b}}{n_2 N_{2,l} N_{2,wd}} V(n_{1,d}) \end{array}\right] t_y$$
$$= ct_y, \qquad (8)$$

where $E(n_{1,d}) = n_1 N_{1,d} / N_1$ and $V(n_{1,d}) = n_1 N_{1,d} N_{2,l} / N_1^2$.

Lavallée (1996) took an interesting approach to a similar problem with panel survey data. In that paper, the problem of frame update using panels with rotation is addressed among other issues. Our approach is different from the approach of that paper in that we consider the two conditional probabilities $\Pr[k \in s_{2,a} \mid n_{1,d}]$ and $\Pr[k \in s_{2,b} \mid n_{1,d}]$ separately.

3. THREE SIMPLE STRATEGIES AND A SIMULATION STUDY

3.1 Strategies in the Presence of Feed Back

A strategy, which is referred to as Strategy 1 here, is to feed back, delete the set $s_{1,d}$ from the frame and accept the feed back bias. However, the size of the bias is seldom known. The estimator for Strategy 1 under SI is $\hat{t}_{y\pi} = N_2 / n_2 \sum_{s_2} y_k$ where s_2 is a sample taken from U_2. To obtain Strategy 2, note that if consistent estimates of $N_{2,d}$ and $N_{2,l}$ are available these may be plugged into (7) or (8) and an estimator with favourable properties is obtained:

$$\hat{t}'_{y\pi} = \hat{t}_{y\pi}(1+\hat{c})^{-1}, \qquad (9)$$

where

$$\hat{c} = (\hat{N}_{2,d} / \hat{N}_{2,l})[n_{2,a} / n_2 - \{n_{2,b}(n_1 - n_{1,d})\} / \{n_2 (N_1 - n_1)\}]$$

for both the conditional and unconditional cases since the term $n_{2,b} V(n_{1,d})(n_2 N_{2,l} N_{2,wd})^{-1}$ in (8) is almost always negligible. The estimates $\hat{N}_{2,d}$ and $\hat{N}_{2,l}$ of the sizes of the domains $U_{2,d}$ and $U_{2,l}$ can be obtained from a sample from the original or current survey population. If more than one sample is drawn, each can provide an unbiased estimate of $N_{2,d}$ (or $N_{2,l}$), all of which can be combined. The minimum variance combined estimator is the sum of the estimators weighted with the reciprocals of their variances. As the following argument shows, we do not expect the bias of (9) to be large:

$$E(\hat{t}'_{y\pi}) = E\left[\hat{t}_{y\pi}(1+\hat{c})^{-1}\right] \approx E(\hat{t}_{y\pi})(1+c)^{-1}$$
$$= t_y(1+c)(1+c)^{-1} = t_y.$$

Another strategy, here denoted by Strategy 3, is to feed back the information that certain units are dead, but to retain them on the frame and allow them to be sampled. The resulting estimator is unbiased, but the disadvantage of this strategy is that the precision will suffer as part of the sample is lost on ineligible units. The estimator of Strategy 3 is $\hat{t}''_{y\pi} = N_1 / n_2 \sum_r y_k$, where r is a sample from the original survey population U_1.

3.2 A Simulation Study

A simulation study may shed some light on which of the Strategies 1–3 is to be preferred. Natural measures for comparing the strategies are bias and variance. In business surveys, estimates for subpopulations (industries) are often more interesting than the whole population. To simulate a subpopulation, a frame consisting of 1,000 units was created to form the original survey population. A gamma distributed value, Y1, was associated with each unit. We used the same gamma distribution as the one that generated Population 12 in Lee, Rancourt and Särndal (1994, page 236). The coefficient of variation (population standard deviation divided by the mean) was 0.57. Another study variable, Y2, was created by performing independent Bernoulli trials, one for each population unit, which obtained value 1 with probability equal to 0.5 and value 0 otherwise. Unlike in Lee *et al.*, some of the units were dead. Each unit was independently of other units classified as dead with a probability P_{dead}. All dead units were assigned zero values for both Y1 and Y2. A set of Y1 and Y2 were simulated for each of four values of P_{dead}: 0.03, 0.05, 0.2, and 0.5. These sets contained 29, 54, 201 and 494 dead units, respectively.

A PRN was attached to each unit and the units were laid out along a PRN line. The first sample, s_1, was drawn by identifying the 500 units with the smallest PRNs. All dead units in s_1 were flagged as 'dead by sample survey sources'. Hence, s_1 covered approximately the first half of the PRN line. The frame with the units flagged as dead by sample survey sources excluded made up the current survey population. The estimates of $N_{2,d}$ and $N_{2,l}$ used in Strategy 2 were based on s_1. A second sample, denoted by $s_{2\text{current}}$, was drawn by taking 100 units to the right of a starting point, *start* 2, disregarding units dead by sample survey sources. Another sample of 100 units was selected from *start* 2, but units dead by sample survey sources were this time allowed to be included in this sample. Hence, this sample was drawn from U_1, and we denote it by $s_{2\text{orig}}$. The sample $s_{2\text{current}}$ is pertinent to Strategies 1 and 2 while $s_{2\text{orig}}$ will be used for Strategy 3.

The procedure described in the preceding paragraph was repeated 1,000 times. That is, for each of the values of P_{dead} mentioned above and for each of three starting points of s_2, to be defined, 1,000 sets of PRNs were generated and attached to the units. The frame was reordered for each new

set of PRNs, and three samples were drawn for each reordering (s_1, $s_{2\text{current}}$, and $s_{2\text{orig}}$). Two values of *start* 2, 0.0 and 0.7, were chosen so as to make the proportion of $s_{2\text{current}}$ that fell in $s_{1,l}$ 100% and 0%, respectively. That is, $n_{2,a}/n_2$ was set to 100% and 0%. Further, to make $n_{2,a}/n_2$ on average 50% under each of the chosen P_{dead}, appropriate values of *start* 2 were derived. They are 0.448, 0.447, 0.438, and 0.4 for the P_{dead} values 0.03, 0.05, 0.2, and 0.5, respectively.

In summary, the population and samples sizes, the study variables Y1 and Y2, and which of the units that were dead were held fixed in our study. For twelve combinations of P_{dead} and $n_{2,a}/n_2$, the reordering of the units on the PRN line through the simulation of new PRNs made the following factors vary:

- which of the units that were included in s_1, $s_{2\text{current}}$, and $s_{2\text{orig}}$;
- how many and which of the dead units that were dead by sample survey sources;
- which of the units that belonged to $s_{1,l}$ and $U_{2,wd}$.

Thus the quantities $s_{1,d}$, $N_{2,d}$ and N_2 vary in the simulations. It seems practical to let them do so rather than controlling them in an experiment with more factors than P_{dead} and $n_{2,a}/n_2$. Hence the results are unconditional, in accordance with (8).

3.3 Results

Table 1 shows the empirical relative bias of Strategies 1 and 2, computed as the straight average of the 1,000 differences between the estimate and the parameter in terms of the percentage of the total obtained in the simulation. Strategy 3 is unbiased and is therefore not included in Table 1. The empirical bias of Strategy 3 that nevertheless appeared in the simulations reflects the simulation error; it was at most 0.5%. As seen in Table 1, Strategy 2 is virtually unbiased as well. Note that the simulated empirical bias under Strategy 1 is what (8) predicts (with allowance for simulation error). This bias is appreciable in nearly all cases and if the proportion of dead (or ineligible) units is high the bias can be very severe indeed. Figure 2 shows the conditional bias given $n_{1,d}$ for $P_{\text{dead}} = 0.50$ and $n_{2,a}/n_2 = 0\%$. Note that the bias given by (6) is locally well described by the regression line in the figure defined by the OLS fit of the bias conditional on $n_{1,d}$. For example, if $n_{1,d} = 220$, then both $N_{2,d}/N_{2,l}$ and $(n_1 - n_{1,d})/(N_1 - n_1)$ equal 0.56 and $B = -0.31$.

Table 1
Bias, % of Total of Y1. The First Entry in Each Cell is the Bias Under Strategy 1, the Second is the Bias Under Strategy 2.

	Average of n_a/n					
P_{dead}	0%		50%		100%	
0.03	-1.6	-0.1	0.4	0.4	1.5	0.0
0.05	-2.8	0.0	0.4	0.4	2.9	0.0
0.20	-10.2	-0.2	1.5	0.4	12.7	0.1
0.50	-24.6	0.2	12.5	0.3	49.0	0.2

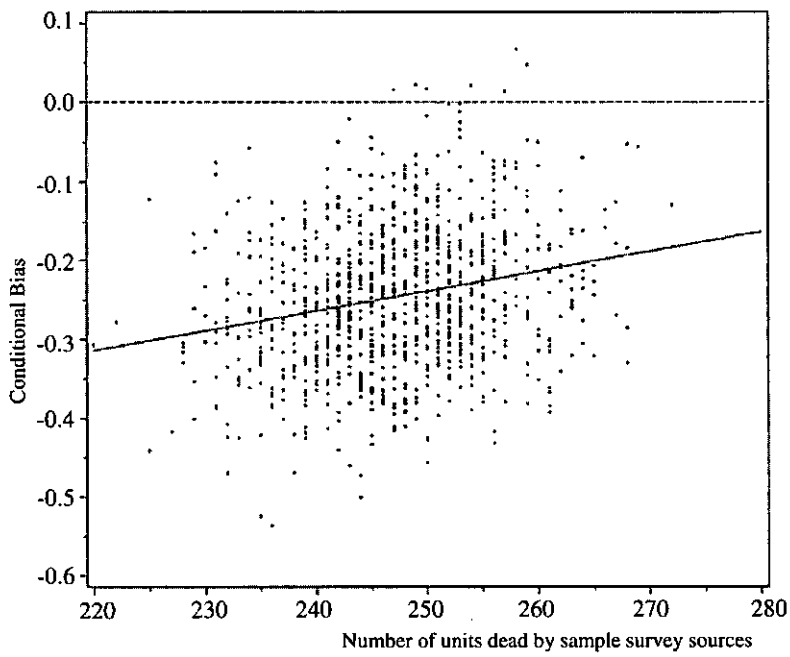


Figure 2. The simulated conditional bias plotted against the number of units dead by sample survey sources, $n_{1,d}$, for $P_{\text{dead}} = 0.50$ and $n_{2,a}/n_2 = 0\%$. An OLS regression line shows the local trend of the conditional bias as a function of $n_{1,d}$.

To assess the bias it helps to look at the coverage probabilities. Table 2 shows the empirical coverage probabilities, based on symmetric 'confidence intervals' with a width of two times the simulated empirical standard deviation of each side of the point estimate. While Strategy 2 gives in all cells coverage probabilities close to the targeted 95%, Strategy 1 achieves that in general only for the population with 3% dead units. The coverage probability under Strategy 1 tends also to be acceptable for populations with a larger proportion of dead units, if half of the sample is taken from the part of the PRN line where dead units have been weeded out, and the other half from the part of the PRN line where the original proportion of dead units has been retained, as the negative bias from the first half of the sample tends to cancel out the positive bias from the second half.

The variance of the simulated estimates was computed. Tables 3 and 4 show the variance comparisons for Y1 and Y2, respectively, under Strategies 2 and 3 relative to that of Strategy 1. As expected, in all cases Strategy 1 gave a smaller variance than did Strategy 3. Strategy 2 performed well in most cases, but considering the extra complexity of this strategy, the feed back Strategy 1 seems preferable for populations with a small proportion of ineligible units, say 3% or less. However, if this proportion is larger than, say, 5%, the bias of Strategy 1 may cause poor coverage probabilities and misleading estimates. The variance of Strategy 2 is no worse than that of Strategy 3; in most cases Strategy 2 is superior. The non-monotone variance ratios in the bottom row of Table 3 is due to the estimation of $N_{2,d}$ and $N_{d,l}$ combined with the specific details of the simulation.

Table 2
The Coverage Probability in Percentage for Estimating Total of Y1. The First Entry in Each Cell is the Coverage Probability Under Strategy 1, the Second is the Coverage Probability Under Strategy 2.

	Average of n_a/n					
P_{dead}	0%		50%		100%	
0.03	94.6	94.3	94.6	94.8	94.3	95.1
0.05	93.3	95.2	94.4	93.9	90.8	95.0
0.20	65.9	94.5	93.8	94.8	46.1	94.6
0.50	21.2	95.1	78.4	94.7	0.0	94.8

Table 3
Variance Ratio of the Estimator of the Total of Y1. The First Entry in Each Cell is the Variance Under Strategy 2 Relative to that of Strategy 1, the Second is the Variance Under Strategy 3 Relative to Strategy 1.

	Average of n_a/n					
P_{dead}	0%		50%		100%	
0.03	1.04	1.04	1.00	1.06	0.98	1.08
0.05	1.08	1.08	0.98	1.14	0.95	1.15
0.20	1.28	1.28	0.85	1.27	0.83	1.46
0.50	1.85	1.85	0.52	1.34	0.58	2.24

Table 4
Variance Ratio of the Estimator of the Total of Y2. The First Entry in Each Cell is the Variance Under Strategy 2 Relative to that of Strategy 1, the Second is the Variance Under Strategy 3 Relative to Strategy 1.

	Average of n_a/n					
P_{dead}	0%		50%		100%	
0.03	1.03	1.03	1.00	1.03	0.97	1.03
0.05	1.06	1.06	0.99	1.04	0.95	1.06
0.20	1.25	1.25	0.92	1.15	0.80	1.19
0.50	1.80	1.81	0.65	1.40	0.50	1.36

4. DISCUSSION

This paper gives conditional and unconditional expressions for the feed back bias when the total is estimated with the common expansion estimator. We have shown that the feed back bias can be large. With as little as 5% ineligible units on the frame, feeding back information of these from sample surveys can result in about 2–3% bias. However, a small-scale simulation study indicates that if the proportion of ineligible units is 3% or less, the feed back strategy does not seem to create problems in terms of bias and variance.

We have also derived a virtually unbiased estimator. The simulation study shows that this estimator compares favourably in terms of variance with the alternative strategy of retaining ineligible units on the frame and letting them be included in further samples. This estimator relies on the availability of consistent estimates of the number of eligible and ineligible units in the population. These estimates may be obtained from an earlier sample through the unbiased strategy of letting units that have been found dead be included in the sample.

In order to facilitate the theoretical development, we have made simplifying assumptions. The most important of these is the assumption that *all* dead units have been found in earlier sample surveys and have been fed back to the frame. We have envisaged a frame with one 'white' area, where all ineligibles have been flagged as such, and one 'black' area, where no ineligibles have been touched. In practice, this is not likely to happen. If the frame is moderately large and used for many continuing surveys, some of which may feed back to varying intensity, the frame will turn 'grey' rather than 'black and white'. The feed back bias will then be less severe than in the 'black and white' situation. It has not, however, been in the scope of this paper to quantify the bias for a 'realistically grey' frame. In this sense, what has been examined in this paper is a worst case scenario.

ACKNOWLEDGEMENTS

The authors thank Mark Pont for very useful initial discussions of this topic. They are also most grateful to an associate editor and two referees for very valuable comments. Both authors' research was partially supported by the UK Office for National Statistics and Wang's research was also supported by the U.S. National Cancer Institute (CA 57030). Hedlin was employed by University of Southampton when he took part in this work.

REFERENCES

COLLEDGE, M.J. (1989). Coverage and classification maintenance issues in economic surveys. In *Panel Surveys*, (Eds., D. Kasprzyk, G.J. Duncan, G. Kalton and M.P. Singh). New York: John Wiley & Sons, Inc., 80-107.

COLLEDGE, M.J. (1995). Frames and business registers: an overview. In *Business Survey Methods*, (Eds., B. Cox, D. Binder, N. Chinnappa, A. Christianson, M. Colledge and P. Kott). New York: John Wiley & Sons, Inc., 21-47.

DURBIN, J. (1969). Inferential aspects of the randomness of sample size in survey sampling. In *New Developments in Survey Sampling*, (Eds., N.L. Johnson and H. Smith). New York: John Wiley & Sons, Inc., 629-651.

ERNST, L.R., VALLIANT, R. and CASADY, R.J. (2000). Permanent and collocated random number sampling and the coverage of births and deaths. *Journal of Official Statistics*, 16, 211-228.

HIDIROGLOU, M.A., and LANIEL, N. (2001). Sampling and estimation issues for annual and sub-annual Canadian business surveys. *International Statistical Review*, 69, 487-504.

HIDIROGLOU, M.A., and SRINATH, K.P. (1993). Problems associated with designing subannual business surveys. *Journal of Business and Economic Statistics*, 11, 397-405.

LEE, H., RANCOURT, E. and SÄRNDAL, C.-E. (1994). Experiments with variance estimation from survey data with imputed values. *Journal of Official Statistics*, 10, 231-243.

LAVALLÉE, P. (1996). Frame update problems with panel surveys. *Proceedings of Statistical Days '96*, Statistical Society of Slovenia, 252-261.

OHLSSON, E. (1995). Coordination of samples using permanent random numbers. In *Business Survey Methods*, (Eds., B. Cox, D. Binder, N. Chinnappa, A. Christianson, M. Colledge and P. Kott). New York: John Wiley & Sons, Inc., 153-169.

SCHIOPU-KRATINA, I., and SRINATH, K.P. (1991). Sample rotation and estimation in the survey of employment, payrolls and hours. *Survey Methodology*, 17, 79-90.

SRINATH, K.P. (1987). Methodological problems in designing continuous business surveys: some Canadian experiences. *Journal of Official Statistics*, 3, 283-288.

SRINATH, K.P. and CARPENTER, R.M. (1995). Sampling methods for repeated business surveys. In *Business Survey Methods*, (Eds., B. Cox, D. Binder, N. Chinnappa, A. Christianson, M. Colledge and P. Kott). New York: John Wiley & Sons, Inc., 171-183.

THOMPSON, M.E. (1997). *Theory of Sample Surveys*. London: Chapman & Hall.

Survey Methodology, December 2004
Vol. 30, No. 2, pp. 175-181
Statistics Canada

Application of Quality Control in ICR Data Capture: 2001 Canadian Census of Agriculture

WALTER MUDRYK and HANSHENG XIE [1]

ABSTRACT

Intelligent Character Recognition (ICR) has been widely used as a new technology in data capture processing. It was used for the first time at Statistics Canada to process the 2001 Canadian Census of Agriculture. This involved many new challenges, both operational and methodological. This paper presents an overview of the methodological tools used to put in place an efficient ICR system. Since the potential for high levels of error existed at various stages of the operation, Quality Assurance (QA) and Quality Control (QC) methods and procedures were built into this operation to ensure a high degree of accuracy in the captured data. This paper describes these QA / QC methods along with their results and shows how quality improvements were achieved in the ICR Data Capture operation. This paper also identifies the positive impacts of these procedures on this operation.

KEY WORDS: Data Capture; Intelligent Character Recognition (ICR); Quality control; Quality improvement; Statistical process control.

1. INTRODUCTION

The data capture of the 2001 Canadian Census of Agriculture was conducted between July and November 2001, using relatively new technology called Intelligent Character Recognition (ICR). This approach to data capture combines Automated Machine Capture which uses optical character, mark and image recognition, with Manual Capture by operators who 'key from image' using a heads-up data capture technique. The heads-up data capture technique is applied only to fields that can not be recognized by the optical system with a sufficiently high degree of confidence (that is pre-specified).

The ICR system offered many benefits to the data capture operation, in terms of resource savings and productivity gains. At the same time, accuracy became an extremely important consideration for processing a large number of documents since the potential for unacceptable levels of error existed at various stages of the process. In the literature, the quality of ICR applications has been studied by a few authors; see, *e.g.*, Kalpic (1994) and Pasley (2000), among others. Kalpic discussed the coding algorithm and the results for the 1991 Census Coding Operation in Croatia and Bosnia-Herzegovina, using intelligent optical readers. Pasley pointed out that the quality of a scanned image usually depends on the quality of the source document, the precision of the scanner, the skill of the scanner operator and the resolution at which the document was scanned. With quality improvement in mind, QA and QC procedures were built into the data capture operation for the 2001 Canadian Census of Agriculture to ensure a high degree of accuracy in this operation.

Quality Control activities for the ICR Data Capture Operation were focused in three main stages of processing, namely: document preparation, scanning calibration, and data capture of the questionnaires. This was done since each of these stages was dependent on one another and each had the potential to contribute significant errors down the line. Therefore, each component should ideally have its own control system.

It is the purpose of this paper to describe the QA/QC methodology and procedures associated with each of the main stages of the ICR Data Capture Operation, summarise the results obtained from their application and show how ongoing quality improvements were achieved in the ICR Data Capture operation.

2. QUALITY PROGRAM OVERVIEW

To better understand the rationale behind the QA/QC procedures, it is worthwhile to give an overview of their objectives and methodologies.

2.1 Objectives

The overall quality objective for this project was to measure, control and improve the quality of the entire ICR Data Capture Operation on a continuous basis. This would be achieved by implementing a series of QA/QC procedures at each critical stage of the operation. The specific objectives for each stage were as follows:

a) Document Preparation: to ensure that only highly readable documents would reach the scanning stage.

[1] Walter Mudryk and Hansheng Xie, Business Survey Methods Division, Statistics Canada, Ottawa, Canada K1A 0T6.

b) Scanning Calibration: to ensure optimal machine set-up and calibration prior to the start of production.
c) Quick Capture (Machine Capture) and Quick Key (Manual Capture): to ensure a high level of quality of data capture during production.

2.2 QA / QC Methodologies

Each major stage of processing was operationally unique and therefore, had different quality requirements. As a result, QA procedures were applied to the Document Preparation operation, and QC procedures to the Scanning Calibration, Quick Capture and Quick Key operations. A flowchart is given in the Appendix, which shows the various stages of the ICR Data Capture Operation and exactly where these procedures were applied.

2.2.1 Document Preparation

The document preparation operation was essentially divided into five sub-processes, specifically: sorting, transcription, batching, cutting and storage. This operation was responsible for preparing the questionnaires and associated batches for scanning by the ICR equipment and was performed manually by clerical staff. It included activities such as separating the contents of the received envelopes by document type (*Sorting*), re-transcribing damaged or illegible questionnaires (*Transcription*), grouping questionnaires into batches for registration (*Batching*), cutting the spine of each booklet questionnaire with an electric cutter (*Cutting*) and filing questionnaires in the archive (*Storage*). One of the most important aspects of this operation was the identification and isolation of problematic questionnaires so that they would not advance undetected to the scanning and data capture stages. These problematic questionnaires were labeled as '*outlier*' questionnaires since they had problems such as questionnaires being *X*'ed out or written over fields, extraneous markings, illegible entries, torn, crumpled or taped documents, *etc.*

The potential for error in this operation could lead to some problems being experienced at the scanning stage. It was felt that QA procedures would be appropriate to ensure quality at this stage since many of the clerical functions were also subject to various automated system cross-checks. The system cross checks ensured that the documents had a valid ID, correct number of pages, and that the pages, once cut, were aligned and in sequential order. The QA procedures consisted of a series of on-going random spot checks for each of the five sub-processes. The results of each spot check were recorded on a control form and summarized for the supervisor to identify if the work was being done correctly. Feedback would then be given to the individual clerk or group on a regular basis, and corrective actions would be taken when necessary. For example, if the work was not being performed well, some re-training would take place and/or an increase in the frequency of spot-checks was done until favorable results were obtained. If extensive problems were identified, the supervisor could also decide on the amount of re-work required, based on the seriousness of the problem observed.

For the sorting, batching, cutting and storage operations, the quality measure selected was '*percent of questionnaires in error*' (*i.e.*, in keeping with the assumptions required for a simple sampling unit). For the transcription operation, the probability of multiple independent errors occurring within a questionnaire was extremely high and therefore the quality measure selected was '*Defects per Hundred Units, DPHU*' (*i.e.*, in keeping with the assumptions required for a complex sampling unit).

2.2.2 Scanning Calibration Check

Experience has shown that if the scanning equipment is not properly configured, the potential for generating poor quality images increases substantially. It is therefore imperative that the scanning equipment be optimally set prior to production and well maintained throughout the scanning operation. To ensure this, a QC procedure called the Scanning Calibration Check was developed to review the machine settings and calibration on an ongoing basis.

Since the equipment settings of the scanning system would tend not to fluctuate too greatly, it was felt that Statistical Process Control (SPC) methods would be appropriate for controlling this portion of the operation. This would essentially be an ongoing spot check of the calibration settings performed on a daily basis prior to the start of production. The calibration check consisted of re-scanning a test batch and comparing the results with the corresponding pre-benchmarked results for the same batch. The differences between the actual and expected results would be compared and error rates computed. These error rates were then plotted on SPC control charts to determine if the process was operating at an acceptable level. If this test batch failed, the scanning process would not be allowed to start production until the machine was re-calibrated and subsequently re-tested successfully.

In the Scanning operation, machine recognition could substitute wrong values when poor quality images are produced. Poor images could be the result of many factors such as dirty read heads, smeared optical windows, mis-alignment, mis-registration of fields, poor contrast / brightness levels, paper feed problems, *etc.* Since a specific quality standard was established for each field type, a separate p control chart was used to evaluate the substitution error rate for each type (specifically, alpha, alphanumeric, numeric, tick boxes and bar codes). The acceptable quality standard for each field type was previously established on a

field type basis by the client area so therefore, the quality measure used was '*percent of fields in error*', *i.e.*, the substitution error rate by field type for each scanner.

Based on SPC control chart theory, a decision for each scanning calibration test was made as follows:

- If each of the sample error rates for the five field types was respectively lower than their corresponding upper control limit (UCL), it was concluded that the scanning system was functioning properly and was ready for scanning production.
- Otherwise, it was concluded that a problem existed with the scanning equipment, and corrective action must be taken before the start of regular production.

The test batches were constructed with minimum sample size requirements in mind for each field type, such that the producer's confidence level would be at least 95%. This was then used as a guide in selecting the actual questionnaires for each of the test batches. The *minimum size* was required for each field type in order to achieve the high efficiency of decisions in the scanning calibration test, while the *Producer's Confidence Level* referred to the likelihood that the scanning system would pass the test for that field type when the system was functioning at the acceptable target level. The Upper Control Limit for each field type was computed assuming a $+2\sigma$ variability. This limit is lower than the customary $+3\sigma$ Upper Control Limits since the scanning calibration check was designed to be more sensitive in detecting smaller shifts at start-up than during normal production.

2.2.3 Quick Capture and Quick Key

Once the questionnaires had been scanned, the system would produce a digital image of each field along with an interpretation of its value and an associated confidence level for its recognition. The actual data capture then consisted of two processes: Quick Capture and Quick Key. Quick Capture was the automatic recognition by the system of all field images whose confidence levels were above a pre-specified threshold value. Quick Key consisted of the heads-up manual capture (by keyers working on terminals) of field images whose confidence levels were below the pre-set threshold value.

Since under ideal circumstances, these two processes were expected to be relatively stable, the QC Procedures were again based on SPC principles and were developed to measure and monitor the quality of each of the processes. This QC approach consisted of a small sample check from the output of a sample of batches taken systematically over time and computing the error rates for each sample. These error rates would then be compared to rejection levels that were calculated by the system based on the expected quality standard and the sample size for that observation. A decision was then made as to the acceptability of each of these sample measurements relative to the expected quality standard for that process.

In the case of the Quick Capture operation, the machine may interpret a different value from the actual value for that field, and therefore, substitution rates were used to evaluate this process. These substitution errors are particularly serious since, if left unchecked, they may affect the recognition rate for many fields for a long period of time. In the case of the Quick Key operation, operators may make keying errors for many reasons such as lack of skill, poor training, fatigue, *etc.*, and therefore, keying error rates were used to evaluate this manual process. For both of these processes, the quality measure was defined as '*percent of fields in error*', across all field types combined.

Within the two capture operations, there were two distinct categories for processing the scanned documents: *Regular* questionnaires and *Outlier* questionnaires. QC procedures were put in place for each category. A separate sample was required for each process, one for Quick Capture and one for Quick Key. The system could distinguish between Quick Capture and Quick Key fields in each sample questionnaire and maintain separate counts of these fields that had been captured under each process. These field counts eventually became the sample size for each sample. Each sample was then compared to its own *threshold rejection rate*, which was a function of the number of fields observed (*i.e.*, the effective *sample size*) and the expected quality standard or target for that process. A decision would then be made to accept or reject the sample. The threshold rejection rate was equivalent to the standard Upper Control Limit (*UCL*) that would be calculated on a standard p control chart. If the sample error rate exceeded this level, the process was rejected and the QC Reviewer proceeded to investigate and implement corrective actions as appropriate; otherwise the process was accepted.

The sampling was done on an individual scanner basis for Quick Capture and an individual operator basis for Quick Key. Some operators required more questionnaires to be sampled from time to time, and others less, based on their actual performance. Since the actual observations were based on samples, a customary $+3\sigma$ variability was permitted above the expected quality standard (*i.e.*, the centerline of a p control chart) for each process. The batch decisions for these sample observations were made by the system during QC verification and these results were then plotted on a p control chart for each scanner and operator, after the fact and updated weekly.

For a detailed description of these QA/QC procedures and their rationale, please refer to Mudryk, Bougie and Xie (2001).

3. QUALITY IMPROVEMENTS

Two essential elements were included in the quality improvement strategy for the ICR Data Capture Operation. These consisted of feedback of QA/QC results and the implementation of corrective and preventive actions when required. These two elements enabled various staff to play an active role in improving the quality of each process through the additional insight into the problems that were identified and through the subsequent corrective or preventive actions that were taken.

Using QC data analysis as the base, all processes were examined to determine if they were operating efficiently. QC meetings were held with operations staff on a weekly basis to review the ongoing progress of the entire operation. Problems that had impacted any of the processes were addressed and recommendations made to treat their root causes and prevent their re-occurrence. The involvement of operational staff in resolving these problems played an important part in facilitating quality improvements on a continuous basis. The following examples illustrate some of the more significant corrective actions that were taken during the operation that led to quality improvements at various stages.

Example 1: Filtering Process for Detecting Outlier Documents

During the first few weeks of production, it was noticed that some documents were causing a high concentration of errors from things like large X's across a page, 0's and dashes in various fields, *etc.* These documents were causing high error rates for both operations but especially for the Quick Capture process. Since these documents were very different from the majority of the regular documents, a procedure was introduced to sort these documents for special treatment and processing after the fact. Some documents in fact had to be re-transcribed at this stage prior to processing them by ICR.

Example 2: Adjusting System Settings for Scanning & Recognition

The highlights of the QC weekly summaries indicated that both scanners made errors frequently on Pages 3 and 14 of the questionnaires during the first few weeks of processing. An investigation was conducted and it was found that there was a template reading problem on Page 3 and the pre-set recognition threshold level for the numeric fields on Page 14 were set too low. After the system settings on both scanners were adjusted, the system showed substantial improvements in the scanning of these two pages.

Example 3: Retraining Operators with High Error Rates

During the keying operation, the QC results showed that certain keyers were experiencing above average difficulties with the 'key from image' process and that their error rates remained high for several weeks. Focusing on continuous improvement, these keyers were offered retraining on an ongoing basis. As a result, many keyers made significant improvements (week by week) in their keying performance.

4. QC EVALUATION AND ANALYSIS

Throughout the operation, many QC reports, charts and estimates, were produced to provide information about the incoming and outgoing quality levels and to evaluate the output of each production process. These reports were used to analyse the quality of each process by week and across weeks.

4.1 Document Preparation

For each of the five sub-processes of the document preparation, individual QA procedures were applied at different frequencies and both corrective and preventive actions were taken on an on-going basis as dictated by the results. The information collected and the feedback that was provided as a result of these QA procedures helped significantly in improving the scanning, imaging, recognition and capture of the questionnaires. In the first few weeks of production, it was discovered from the QC results that problematic documents (*i.e.*, *outliers*) were causing most of the substitution errors (*i.e.*, machine errors) in the *Quick Capture* process. From that point on, a new procedure was introduced into the *Sorting* process of the Document Preparation operation to separate these documents for special treatment from the regular documents (*i.e.*, labeled them for subsequent 100% verification). In general, better quality documents reached the scanning stations while poorer documents were either re-transcribed or processed separately with the addition of post processes such as 100% verification.

4.2 Scanning Calibration Check

In an effort to ensure optimal scanner settings and calibration, a *Scanning Calibration Check* was initially conducted twice a day, and subsequently once a day, prior to production processing. Many test batches were scanned during the operation with a relatively high rejection rate encountered by each scanner. On average, approximately 2-3 tests per day (with corresponding re-calibrations) were required for optimising the set-up of each of the two scanners. This demonstrates the need for re-calibration between processing periods. It should be noted that some rejections occurred due to problems identified with the test batches which were fixed later on. This is definitely an area where some procedural improvement is required in the future.

Both scanners exhibited reasonably high variability during this test. The high number of tests required, high rate of rejection and high variability across processing periods for many of the field types demonstrate the need to calibrate the scanning equipment properly prior to production. Otherwise, the scanners could be inadvertently set up to produce poor images right from the start, which would make good quality capture very difficult. Once a test batch failed, problems were usually identified and subsequent maintenance and corrective actions taken. This included actions such as: re-configuring the scanning equipment, replacing old light bulbs, fixing software problems, cleaning dirty read heads, *etc.* Using this test, the scanners were able to be calibrated and maintained at optimum levels of performance, between production runs.

4.3 Quick Capture and Quick Key

For the ***Quick Capture*** process, over the entire 18 weeks of processing the ***Regular*** **questionnaires**, the overall weekly substitution error rates decreased steadily from 4.3% to 0.8%, resulting in a grand overall substitution error rate of 2.0% (across all field types) for both scanners. The substitution error rates measured during production were maintained very near the Target levels that were established for each field type. These were as follows: Alpha (2.1% relative to a target of 2.0%); Alphanumeric (3.2% *vs.* 3.5%); Bar Code (0.0% *vs.* 0.2%); Numeric (2.8% *vs.* 2.0%) and Tick Boxes (0.8% *vs.* 0.4%). In comparison, processing the ***outlier*** **questionnaires** had a much higher substitution error rate and greater weekly variability than the corresponding ***regular*** questionnaires (*i.e.*, ranged from a high of 22.4% to a low of 1.3%). Although the substitution error rate did tend to reduce substantially over time, it did remain relatively high throughout the process and was measured at 7.0% overall, which was significantly higher than the rate for ***regular*** questionnaires (*i.e.*, 2.0%).

For the *Quick Key* process, the keying error rate for processing the ***regular*** **questionnaires** was relatively high throughout the entire processing period (*i.e.*, mostly over 3%). This was partially due to the fact that this operation was a *heads-up* keying process and these keyers typically processed the most difficult cases. Over the entire 18 weeks however, the weekly keying error rates generally decreased from 5.6% to 1.6%, with an overall average of 3.4%. The keying was also subject to high levels of variability among operators, with individual error rates ranging 1.7% to 7.5%. It is interesting that keying the ***outlier*** **questionnaires** had a similar keying error rate to the corresponding ***regular*** process (*i.e.*, 3.4% *vs.* 3.7%) and ranged from a high of 5.7% to a low of 1.6%.

4.4 Estimates of Average Outgoing Quality

The primary purpose of the QA/QC procedures was to identify problems and to prevent them from occurring again. However, these procedures also had a corrective component in the sense that, errors that were discovered were always rectified. It is therefore possible to estimate the overall Average Outgoing Quality (AOQ) for the data capture component after the application of the QC procedures.

Estimates of AOQ were calculated for each of the two data capture processes. For a sampled outlier batch, all the questionnaires (*i.e.*, sampled and remainder) in that batch would be subjected to subsequent 100% verification, while for a regular batch, only the sampled questionnaires would be verified. This affects the calculation of AOQ since it can be assumed that the outgoing error rate for all verified questionnaires is 0.0%. The overall estimate for each component was based on the information obtained from both the regular and outlier documents, considering estimates of incoming quality and corrections made during verification. In the calculation, any documents reprocessed through either Quick Capture or Quick Key were included in the count.

Table 1 provides estimates of the AOQ for the *Quick Capture* and *Quick Key* processes.

Table 1
Estimates of AOQ for ICR Data Capture

Process	No. Questionnaires in Population	No. Fields in Population	Estimated No. Fields Verified and Corrected	Incoming Error (%)	AOQ (%)
Quick Capture					
Regular	273,818	21,248,277	170,249	2.01	1.99
Outlier	12,702	1,044,358	1,044,358	6.99	0.00
Overall	286,520	22,292,635	1,214,607	2.95	**1.90**
Quick Key					
Regular	281,502	6,376,020	234,253	3.41	3.28
Outlier	25,788	686,734	686,734	3.67	0.00
Overall	307,290	7,062,754	920,987	3.45	**2.97**
Combined					
Regular		27,624,297	404,502	2.82	2.29
Outlier		1,731,092	1,731,092	5.09	0.00
Overall		29,355,389	2,135,594	3.24	**2.16**

It can be seen that the overall AOQ for the *Quick Capture* process was estimated at 1.90% and for the *Quick Key* process at 2.97%. This was down considerably from their corresponding estimates of incoming quality of 2.95% and 3.45% respectively. The overall AOQ for both processes was estimated at 2.16% (relative to an overall incoming error quality of 3.24%). It should be noted that the AOQ for *outlier* documents was assumed to be 0% since all *outlier* documents were subsequently 100% verified.

4.5 QC Summary

The above results clearly indicate the need for the QA/QC procedures at the different stages of processing. It also shows how they collectively contributed to controlling the outgoing quality and generating quality improvements into all phases of the ICR data capture operation.

The QC results clearly showed that the *outlier* documents had a greater negative impact on the *Quick Capture* process (*i.e.*, 7.0% substitution error rate) than the *Quick Key* process (*i.e.*, 3.7% keying error rate). This indicates that the filtering process for special treatment of *outlier* documents was an important step to take. The QC results also showed that if the documents were in good shape for scanning and the machines were well calibrated, the *automated* system was capable of capturing the data faster and with better quality than the manual *key from image* process. This is quite an important observation, since there are obvious savings implied with a corresponding improvement in data capture quality (*i.e.*, 2.0% *vs.* 3.4%). To the defence of the keyers, however, they did process the more difficult cases, thus partially explaining their higher error rates. Overall, it was estimated that about 77% of the fields were captured through the *Quick Capture* process and 23% were captured through the *Quick Key* process.

It should also be noted that the regular feedback of the QC information collected from the various stages of the ICR process was essential in identifying the root causes of many problems and in helping to resolve them. This provided the opportunity for many quality improvements to be generated into the various stages, on an on-going basis.

For a detailed description of these QA/QC results, please refer to Mudryk and Xie (2002).

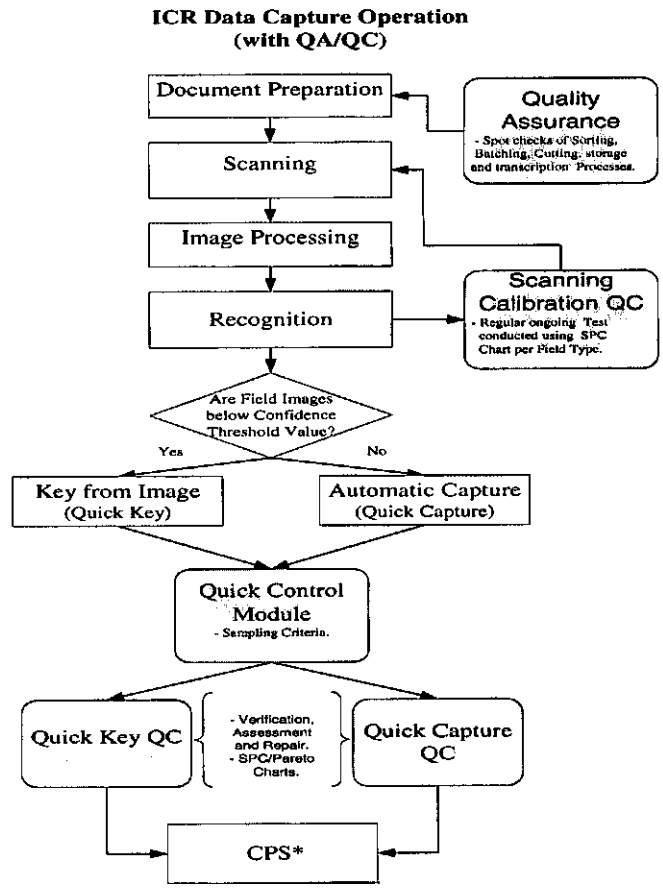
5. CONCLUSIONS

It is clear from the results obtained in this analysis, that the QA/QC procedures were extremely valuable and had a very positive impact on the entire operation. The QA procedures that were applied in the Document Preparation process were effective in preventing many poor documents from reaching the scanning stations and those that did were then labeled for special treatment and subsequent 100% verification.

The QC procedures were then able to optimize the machine set-up by applying the Scanning Calibration Check prior to production. Furthermore during production, QC samples were also able to identify problems with the automatic recognition and key from image processes, so that they could be improved as required.

In all cases, early warning signals were obtained from objective measurements at each stage of processing, and corrective and preventive actions were implemented as needed. Extensive feedback was provided to all stages of the ICR process on an ongoing basis from which continuous quality improvements were generated.

APPENDIX



* CPS = Central Processing System.

ACKNOWLEDGEMENTS

The authors are grateful to the Editor, to an Associate Editor and to an Assistant Editor for their detailed and constructive comments. They also thank Bob Bougie for many helpful comments.

REFERENCES

KALPIC, D. (1994). Miscellanea, Automated Coding of Census Data. *Journal of Official Statistics*, 10, 4, 449-463.

MUDRYK, W., BOUGIE, B. and XIE, H. (2001). Quality Control of ICR Data Capture: 2001 Canadian Census of Agriculture. *International Conference on Quality in Official Statistics in Stockholm*, Sweden.

MUDRYK, W., and XIE, H. (2002). Quality Control Application in ICR Data Capture for the 2001 Canadian Census of Agriculture. *Proceedings of the Joint Statistical Meetings*, American Statistical Association, 2424-2429.

PASLEY, B. (2000). Web Exclusive: The Good and Bad of Scanned Images. Posted on the POB (Point of Beginning) website.

Survey Methodology, December 2004
Vol. 30, No. 2, pp. 183-193
Statistics Canada

Design Effects for the Weighted Mean and Total Estimators Under Complex Survey Sampling

INHO PARK and HYUNSHIK LEE [1]

ABSTRACT

We revisit the relationship between the design effects for the weighted total estimator and the weighted mean estimator under complex survey sampling. Examples are provided under various cases. Furthermore, some of the misconceptions surrounding design effects will be clarified with examples.

KEY WORDS: Simple random sample; pps sampling; Multistage sampling; Self-weighting; Poststratification; Intracluster correlation coefficient.

1. INTRODUCTION

The design effect is widely used in survey sampling for developing a sampling design and for reporting the effect of the sampling design in estimation and analysis. It is defined as the ratio of the variance of an estimator under a complex sampling design to that of the estimator under simple random sampling with the same sample size. An estimated design effect is routinely produced by computer software packages for complex surveys such as WesVar and SUDAAN. It was originally intended and defined for the weighted (ratio) estimator of the population mean (Kish 1995). However, a common practice has been to apply this concept for other statistics such as the weighted total estimator often with success but at times with confusion and misunderstanding. The latter situation occurs particularly when simple but useful results derived under a relatively simple sampling design are applied to more complex problems. In this paper, we examine the relationship between the design effects for the weighted total estimator and the weighted mean estimator under various complex survey sampling designs. In section 2, we briefly review the definition of the design effect and its practical usage while discussing some of the misconceptions surrounding design effects for the weighted total and mean estimators. Subsequently, in section 3, we analyze the difference between the design effect for the weighted total estimator and that for the weighted mean estimator under a two-stage sampling design followed by a discussion regarding the design effects under various two-stage sampling designs and some more general cases in section 4. We try to clarify some of the misconceptions with these examples. Finally, we summarize our discussion in section 5.

2. A BRIEF REVIEW ON DEFINITION AND USE OF DESIGN EFFECT IN PRACTICE

A precursor of the design effect that has been popularized by Kish (1965) was used by Cornfield (1951). He defined the efficiency of a complex sampling design for estimating a population proportion as the ratio of the variance of the proportion estimator under simple random sampling with replacement (srswr) to the corresponding variance under a simple random cluster sampling design with the same sample size. The inverse of the ratio defined by Cornfield (1951) was also used by others. For example, Hansen, Hurwitz and Madow (1953, Vol. I, pages 259 – 270) discussed the increase of the relative variance of a ratio estimator due to the clustering effect of cluster sampling over simple random sampling without replacement (srswor). The name, design effect, or Deff in short, however, was coined and defined formally by Kish (1965, section 8.2, page 258) as "the ratio of the actual variance of a sample to the variance of a simple random sample of the same number of elements" (for more history, see also Kish 1995, page 73 and references cited therein).

Suppose that we are interested in estimating the population mean $(\bar{Y})$ of a variable y from a sample of size m drawn by a complex sampling design denoted by p from a population of size M. Kish's Deff for an estimate $(\bar{y}_p)$ is given by

$$\text{Deff} = \frac{V_p(\bar{y}_p)}{(1-f)S_y^2/m} \tag{2.1}$$

where V_p denotes variance with respect to p, $f = m/M$ is the overall sampling fraction, and $S_y^2 = (M-1)^{-1} \sum_{k=1}^{M}(y_k - \bar{Y})^2$ is the population element variance of the

[1] Inho Park and Hyunshik Lee, Westat, Inc. 1650 Research Blvd., Rockville, MD 20850, U.S.A. E-mail: InhoPark@westat.com.

y-variable. Although the design effect was originally intended and defined for an estimator of the population mean (Kish 1995), it can be defined for any meaningful statistic computed from a sample selected by a complex sampling design.

The Deff is a population quantity that depends on the sampling design and refers to a particular statistic estimating a particular population parameter of interest. Different estimators can estimate the same parameter and their design effects are different even under the same design. Therefore, the design effect includes not only the efficiency of the design but also the efficiency of the estimator. Särndal, Swensson, and Wretman (1992, page 54) made this point clear by defining it as a function of the design (p) and the estimator ($\hat{\theta}$) for the population parameter ($\theta = \theta(y)$). Thus, we may write it as

$$\text{Deff}_p(\hat{\theta}) = \frac{V_p(\hat{\theta})}{V_{\text{srswor}}(\hat{\theta}')}$$

where $\hat{\theta}'$ is the usual form of an estimator for θ under srswor, which is normally different from $\hat{\theta}$. For example, to estimate the population mean, one may use the weighted (ratio) mean $\hat{\theta} = \sum_s w_k y_k / \sum_s w_k$ with sampling weights w_k but $\hat{\theta}'$ would be the simple sample mean $\sum_s y_k / m$, where the summation is over the sample s. We will see the effect of particular estimators $\hat{\theta}$ on the design effect in the later sections.

Kish (1995) later advocated using a somewhat different definition, which is called Deft and uses the srswr variance in the denominator on the ground that without-replacement sampling is a part of the design and should be captured in the definition. He also reasoned that Deft is easier to use for making inferences and that it is better to define the design effect without the finite population correction factor $(1-f)$ because the factor is difficult to compute in some situations. The new definition is given by

$$\text{Deft}_p(\hat{\theta}) = \sqrt{\frac{V_p(\hat{\theta})}{V_{\text{srswr}}(\hat{\theta}')}}$$

or $\text{Deft}_p^2(\hat{\theta}) = V_p(\hat{\theta}) / V_{\text{srswr}}(\hat{\theta}')$. Survey data software such as WesVar and SUDAAN produce Deft^2 instead of Deff. We will use this definition in this paper.

When the population parameter is the total (Y), the unbiased estimator is the weighted sample total, namely, $\hat{Y} = \sum_s w_k y_k$. When the population mean is the parameter of interest, it is usually estimated by the weighted mean, that is, $\hat{\bar{Y}} = \sum_s w_k y_k / \sum_s w_k$. It is a special case of the ratio estimator, $\sum_s w_k y_k / \sum_s w_k x_k$, where $x_k \equiv 1$ for all $k \in s$.

One common misconception about the design effects for $\hat{Y}$ and $\hat{\bar{Y}}$ is that they are similar in values. However, it has been observed that the design effect for $\hat{Y}$, $\text{Deft}_p^2(\hat{Y})$, tends to be much larger than that for $\hat{\bar{Y}}$, $\text{Deft}_p^2(\hat{\bar{Y}})$. This was also noted in, for example, Kish (1987) and Barron and Finch (1978). Some explanation can be found in Hansen *et al.* (1953, Vol. I, pages 336 – 340) who showed that the difference arises from the relative variance of the cluster sizes. More recently Särndal *et al.* (1992, pages 315 – 318) showed that contrary to the case of $\hat{\bar{Y}}$, the design effect for $\hat{Y}$ depends on the (relative) variation of the y-variable. In fact, even the design effect for $\hat{\bar{Y}}$ may depend on the (relative) variation of the y-variable, which we will discuss in section 4. This dependence contradicts what the design effect is intended to measure as Kish (1995) explicitly described:

> "Deft are used to express the effects of sample design beyond the elemental variability (S_y^2/m), removing both the units of measurement and sample size as nuisance parameters. With the removal of S_y, the units, and the sample size m, the design effects on the sampling errors are made generalizable (transferable) to other statistics and to other variables, within the same survey, and even to other surveys."

His statement may be loosely true for the weighted mean $\hat{\bar{Y}}$ as expressed in the frequently used sample approximate formula for $\text{Deft}_p^2(p, \hat{\bar{Y}})$ given by Kish (1987):

$$\text{Deft}_p^2(\hat{\bar{Y}}) = \{1 + \rho(\bar{m} - 1)\}(1 + \text{cv}_w^2) \qquad (2.2)$$

where the sample design p contains complex features such as unequal weighting and cluster sampling, $\rho = \rho_p(y)$ is the intraclass correlation coefficient (often called within cluster homogeneity measure), $\bar{m}$ is the average cluster sample size, and cv_w^2 is the sample relative variance of the weights. Strictly speaking, this formula is not independent of the y-variable because ρ is dependent on the y-variable. Also, the design effect may not be free of the unit of measurement unless $V_p(\hat{\bar{Y}})$ is expressed in a factorial form of S_y^2/m. See Park and Lee (2002). This formula (2.2) is valid only when there is no correlation between the sampling weights and the survey variable y. However, if the correlation is present, the formula may need to be modified as studied by Spencer (2000) and Park and Lee (2001). In the following section, we elaborate this aspect in detail for two-stage sampling and we will also examine this point further in section 4.1.

3. DECOMPOSITION OF THE DESIGN EFFECT UNDER TWO-STAGE SAMPLING

We consider a sampling design conducted in two stages. Suppose that a population $U = \{k : k = 1, \ldots, M\}$ with M elements is grouped into N clusters of size M_i such that

$M = \sum_{i=1}^{N} M_i$. The first stage sample $s_a = \{i : i = 1, \ldots, n\}$ of n clusters (primary sampling units, or PSUs in abbreviation) is selected with replacement from N clusters with probabilities p_i, where $\sum_{i=1}^{N} p_i = 1$. Let $p_a = \Pr(s_a)$ denote the first stage sampling design. The second stage sample $s_{bi} = \{j : j = 1, \ldots, m_i\}$ of m_i elements (secondary sampling units or SSUs in abbreviation) is then selected independently from each PSU i selected at the first stage according to some arbitrary sampling design, say $p_{bi} = \Pr(s_{bi} \mid s_a)$ where $i \in s_a$. Denote the total sample of elements and the overall sampling design by $s = \cup_{i \in s_a} s_{bi}$ and $p = \Pr(s)$, respectively. Associated with the j^{th} element in the i^{th} cluster is a survey characteristic $y_{ij}, j = 1, \ldots, M_i$, $i = 1, \ldots, N$. For a given $i \in s_a$, let $w_{j|i}$ be the second stage sampling weights such that an estimator of the form $\hat{Y}_i = \sum_{j=1}^{m_i} w_{j|i}\, y_{ij}$ is unbiased for the cluster total $Y_i = \sum_{j=1}^{M_i} y_{ij}$, that is, $E_b(\hat{Y}_i) = Y_i$, where E_b represents the expectation with respect to the second stage sampling. Let $w_i = 1/(np_i)$ be the first stage sampling weights and let $Y = \sum_{i=1}^{N} Y_i$ be the population total. It is easy to show that $E_a(Y_i / p_i) = Y$. Assuming that Y_i are known for $i \in s_a$, $\sum_{i=1}^{n} w_i Y_i$ is the average of n unbiased estimators of Y so that $E_a(\sum_{i=1}^{n} w_i Y_i) = Y$, where E_a denotes the expectation with respect to the first stage sampling design. Note that both stages are sampling with replacement. Accordingly, it is possible that the same sampling unit (either cluster or element) is selected more than once but they are treated differently. Define the overall sampling weights by $w_{ij} = w_i\, w_{j|i}$. Clearly, $\hat{Y} = \sum_{i=1}^{n} \sum_{j=1}^{m_i} w_{ij}\, y_{ij}$ is unbiased for Y, that is, $E_p(\hat{Y}) = E_a\, E_b\, (\hat{Y}) = E_a\, (\sum_{i=1}^{n} w_i\, Y_i) = Y$, where E_p represents the expectation with respect to p. The variance of $\hat{Y}$ can be written as

$$V_p(\hat{Y}) = V_a E_b(\hat{Y}) + E_a V_b(\hat{Y}) = \sum_{i=1}^{N} w_i (Y_i - p_i Y)^2 + \sum_{i=1}^{N} w_i V_b(\hat{Y}_i) \quad (3.1)$$

where V_p, V_a and V_b represent variances defined with respect to the overall, the first stage, and the second stage sampling. See Särndal *et al.* (1992, pages 151 – 152).

A commonly used estimator for the population mean $\bar{Y} = Y/M$ is the weighted (ratio) estimator given by $\hat{\bar{Y}} = \hat{Y}/\hat{M}$ where $\hat{M} = \sum_{i=1}^{n} \sum_{j=1}^{m_i} w_{ij}$. Using Taylor linearization, as shown in Särndal *et al.* (1992, pages 176 – 178), $\hat{\bar{Y}}$ can be approximated as

$$\hat{\bar{Y}} \cong \bar{Y} + M^{-1} \hat{D} \quad (3.2)$$

where $\hat{D} = \sum_{i=1}^{n} \sum_{j=1}^{m_i} w_{ij}\, d_{ij}$ is an unbiased estimator of the population total $D = \sum_{i=1}^{N} \sum_{j=1}^{M_i} d_{ij}$ of $d_{ij} = y_{ij} - \bar{Y}$, which represents the deviation of y_{ij} from the population mean $\bar{Y}$. Note that $D = 0$. Denoting $D_i = \sum_{j=1}^{M_i} d_{ij} = Y_i - M_i \bar{Y}$ and $\hat{D}_i = \sum_{j=1}^{m_i} w_{j|i}\, d_{ij}$, we obtain the approximate variance of $\hat{\bar{Y}}$ from expression (3.2) as

$$\mathrm{AV}_p(\hat{\bar{Y}}) = \frac{1}{M^2}\left[\sum_{i=1}^{N} w_i \left(Y_i - \frac{M_i}{M} Y\right)^2 + \sum_{i=1}^{N} w_i V_b(\hat{D}_i)\right]. \quad (3.3)$$

If a simple random sample of size $m = \sum_{i=1}^{n} m_i$ is selected with replacement from the population U, then a sample mean $\bar{y}_{\text{srs}} = \sum_s y_k / m$ and its expansion

$$\hat{Y}_{\text{srs}} = M \bar{y}_{\text{srs}} = \frac{1}{f} \sum_s y_k \quad (3.4)$$

would serve as the estimators of the population mean $\bar{Y}$ and total Y, respectively, under srswr, where $f = m/M$ is the overall sampling fraction. Their variances under this sampling design are given as $V_{\text{srswr}}(\hat{Y}_{\text{srs}}) = M^2\, V_{\text{srswr}}(\bar{y}_{\text{srs}})$, where $V_{\text{srswr}}(\bar{y}_{\text{srs}}) = m^{-1} S_y^2$ and $S_y^2 = (M-1)^{-1} \sum_s (y_k - \bar{Y})^2$. We note that m is the achieved sample size, which is a random quantity in general. From (3.1), (3.3), and above expressions with m replaced by its expected value m' with respect to the overall sampling design p, *i.e.*, $m' = E_p(m)$, the design effects for $\hat{Y}$ and $\hat{\bar{Y}}$ can be written as

$$\mathrm{Deft}_p^2(\hat{Y}) = \frac{m'}{\mathrm{CV}_y^2}\left\{\sum_{i=1}^{N} w_i \left(\frac{Y_i}{Y} - p_i\right)^2 + \sum_{i=1}^{N} w_i V_b\left(\frac{\hat{Y}_i}{Y}\right)\right\} \quad (3.5)$$

and

$$\mathrm{Deft}_p^2(\hat{\bar{Y}}) \cong \frac{m'}{\mathrm{CV}_y^2}\left\{\sum_{i=1}^{N} w_i \left(\frac{Y_i}{Y} - \frac{M_i}{M}\right)^2 + \sum_{i=1}^{N} w_i V_b\left(\frac{\hat{D}_i}{Y}\right)\right\} \quad (3.6)$$

where $\mathrm{CV}_y^2 = S_y^2 / \bar{Y}^2$ represents the population relative variance of the y-variable. From these expressions, the difference in design effects for $\hat{Y}$ and $\hat{\bar{Y}}$ can be written as follows.

$$\mathrm{Deft}_p^2(\hat{Y}) - \mathrm{Deft}_p^2(\hat{\bar{Y}}) \cong \Delta_a + \Delta_b, \quad (3.7)$$

where

$$\Delta_a = \frac{m'}{\mathrm{CV}_y^2}\left\{\sum_{i=1}^{N} w_i \left[\left(\frac{Y_i}{Y} - p_i\right)^2 - \left(\frac{Y_i}{Y} - \frac{M_i}{M}\right)^2\right]\right\}$$

and

$$\Delta_b = \frac{m'}{\mathrm{CV}_y^2}\left\{\sum_{i=1}^{N} w_i \left[V_b\left(\frac{\hat{Y}_i}{Y}\right) - V_b\left(\frac{\hat{D}_i}{Y}\right)\right]\right\}.$$

The two components Δ_a and Δ_b in expression (3.7) reflect the differences arising from the respective sources of variation from the first and second stages of sampling. Of course, the second component disappears if all the elements in selected clusters are observed since it becomes a single-stage design or if a simple random sample is selected in the second stage. This is because both variances $V_b(\hat{Y}_i)$ and $V_b(\hat{D}_i)$ are equivalent under the aforementioned conditions, that is, 1) $V_b(\hat{Y}_i) = V_b(\hat{D}_i) = 0$ if $w_{j|i} = 1$ for all i and j, and

2) $V_b(\hat{Y}_i) = V_b(\hat{D}_i) \geq 0$ if $w_{j|i} = M_i/m_i$ for all i and j. In other words,

$$\Delta_b = 0 \quad \text{if} \quad w_{j|i} = c_i \text{ for all } i \text{ and } j, \tag{3.8}$$

where c_i are nonnegative constants and not necessarily equal for different clusters. Meanwhile, we can show that

$$\Delta_a = \begin{cases} 0 & \text{if} \quad p_i \propto M_i, \\ A_p(y) & \text{if} \quad Y_i \propto M_i, \\ -A_p(y) & \text{if} \quad p_i \propto Y_i, \end{cases} \tag{3.9}$$

for all i, where $A_p(y) = (m'/\mathrm{CV}_y^2)\sum_{i=1}^N w_i(p_i - M_i/M)^2$. Note that $A_p(y)$ is a nonnegative quantity and also that the conditions in expression (3.9) can be restated, respectively, as $p_i = M_i/M, \bar{Y}_i = \bar{Y}$, and $p_i = Y_i/Y$, where $\bar{Y}_i = Y_i/M_i$ for all $i = 1, \ldots, N$. This result reveals the effect of cluster sampling on the precision of the two estimators. For example, if $p_i = M_i/M$, cluster sampling makes no difference in the precision of the two estimators. On the other hand, if $p_i = Y_i/Y$, $\hat{Y}$ becomes more efficient than $\hat{\hat{Y}}$ in precision under cluster sampling, whereas the cluster sampling favors $\hat{\hat{Y}}$ over $\hat{Y}$ in terms of precision if $\bar{Y}_i = \bar{Y}$ for all i.

Now, let us consider some examples of the conditions of (3.8) and (3.9).

Example 3.1 For one or two-stage cluster design with pps cluster sampling using $p_i = M_i/M$ and $w_{j|i} = c_i$ for all $i = 1, \ldots, N$, we have from (3.8) and (3.9) that $\Delta_a = \Delta_b = 0$, that is, there is no difference in the design effects for $\hat{\hat{Y}}$ and $\hat{Y}$.

The same result as given in example 3.1 can be achieved by $\hat{Y} = M\bar{Y}$. This estimator is the ratio estimator, which can be used if M is known. The case that overall sampling weights are a constant for all the elements (*i.e.*, self-weighting sampling design) is a well known special case. We will come back to this in section 4.

Example 3.2 One-stage simple random cluster sampling or two-stage sample design with srs for both stages. Under these designs, we have $w_{j|i} = c_i$ and $p_i = 1/N$ for all i and j and thus, it follows from (3.8) and (3.9) that $\Delta_b = 0$ and

$$\Delta_a = \begin{cases} 0 & \text{if} \quad M_i \equiv M_0 \text{ for all } i, \\ \overline{m}' \dfrac{\mathrm{CV}_M^2}{\mathrm{CV}_y^2} & \text{if} \quad \bar{Y}_i \text{ are all equal}, \\ -\overline{m}' \dfrac{\mathrm{CV}_M^2}{\mathrm{CV}_y^2} & \text{if} \quad Y_i \text{ are all equal}, \end{cases} \tag{3.10}$$

where $\overline{m}' = m'/n, \mathrm{CV}_M^2 = \overline{M}^{-2}\sum_{i=1}^N (M_i - \overline{M})^2/N$ denotes the relative variance of cluster sizes M_i, and $\overline{M} = M/N$ denotes the average size of clusters. The conditions in (3.10) also satisfy the conditions in (3.9) and therefore, (3.10) is a special case of (3.9). Note that the quantity $A_p(y)$ in expression (3.9) approximately reduces to $\overline{m}' \cdot \mathrm{CV}_M^2/\mathrm{CV}_y^2$ when $p_i = 1/N$ for all i.

Example 3.2 shows that when unequal cluster sizes are not reflected in the sampling design, the relative efficiency of $\hat{Y}$ over $\hat{\hat{Y}}$ depends in part on the relative variability of cluster sizes. If the cluster means are all equal, then cluster sampling makes $\hat{\hat{Y}}$ more efficient than $\hat{Y}$, vice versa if all the cluster totals are equal. On the other hand, if all clusters are equal in size, no difference in the design effects arises by simple random sampling of clusters.

In section 4, we utilize the results derived in this section to discuss other examples used in the sampling literature.

4. EXAMPLES ON THE DESIGN EFFECT IN THE SAMPLING LITERATURE

4.1 Unequal Probability Element Sampling

Consider an unequal probability element sampling design without clustering. The discussion in section 3 applies to this example with $M_i \equiv 1$ for all $i = 1, \ldots, N$ and thus, $m = n$. For brevity's sake, we use lower cases y_i to denote the value of the y-variable, and we also assume that N is large so that $N/(N-1) \cong 1$. Due to the absence of the second stage sampling variation, the design effects for $\hat{Y}$ and $\hat{\hat{Y}}$ given in expressions (3.5) and (3.6) reduce to

$$\mathrm{Deft}_p^2(\hat{Y}) \cong \frac{\sum_{i=1}^N p_i^{-1}(y_i - p_i Y)^2}{\sum_{i=1}^N N(y_i - \bar{Y})^2} \tag{4.1}$$

and

$$\mathrm{Deft}_p^2(\hat{\hat{Y}}) \cong \frac{\sum_{i=1}^N p_i^{-1}(y_i - \bar{Y})^2}{\sum_{i=1}^N N(y_i - \bar{Y})^2}. \tag{4.2}$$

Further let us consider an example where the survey variable y is not correlated with the selection probability p_i.

Example 4.1 Unequal probability element sampling with no correlation between y_i and p_i. When y_i and p_i are not correlated, we can approximate $\sum_{i=1}^N p_i^{-1}(y_i - \bar{Y})^2$ by $n\overline{W}\sum_{i=1}^N (y_i - \bar{Y})^2$, where $\overline{W} = N^{-1}\sum_{i=1}^N w_i$. Note that $E_p(n^{-1}\sum_{i=1}^n w_i) = N/n$, $E_p(n^{-1}\sum_{i=1}^n w_i^2) = N\overline{W}/n$ and $E_p(n^{-1}\sum_{i=1}^n w_i^2)/E_p^2(n^{-1}\sum_{i=1}^n w_i) = n\overline{W}/N$. Thus,

$$\mathrm{Deft}_p^2(\hat{\hat{Y}}) \cong n\overline{W}/N = E_p\left(n^{-1}\sum_{i=1}^n w_i^2\right)\Big/E_p^2\left(n^{-1}\sum_{i=1}^n w_i\right). \tag{4.3}$$

It is easy to show that $n\overline{W}/N \geq 1$ using the Cauchy-Schwarz inequality (Apostol 1974, page 14). In addition, routine calculations show from (4.1) and (4.2) that

$$\begin{aligned}&\mathrm{Deft}_p^2(\hat{Y})-\mathrm{Deft}_p^2(\hat{\bar{Y}})\\&\cong \mathrm{CV}_y^{-2}\left\{\sum_{i=1}^N p_i^{-1}(p_i-\bar{p})^2-2Y^{-1}\sum_{i=1}^N p_i(y_i-\bar{Y})(p_i-\bar{p})\right\}\\&=\mathrm{CV}_y^{-2}(n\bar{W}/N-1),\end{aligned}$$

where $\bar{p}=N^{-1}\sum_{i=1}^N p_i=1/N$. The latter expression is obtained from $\sum_{i=1}^N p_i^{-1}(p_i-\bar{p})^2=n\bar{W}/N-1$ and $\sum_{i=1}^N p_i^{-1}(y_i-\bar{Y})(p_i-\bar{p})\cong 0$ because y_i and p_i are uncorrelated. Consequently,

$$\mathrm{Deft}_p^2(\hat{Y})-\mathrm{Deft}_p^2(\hat{\bar{Y}})\cong \mathrm{CV}_y^{-2}\left\{\mathrm{Deft}_p^2(\hat{\bar{Y}})-1\right\}$$

or

$$\mathrm{Deft}_p^2(\hat{Y})\cong(1+\mathrm{CV}_y^{-2})\,\mathrm{Deft}_p^2(\hat{\bar{Y}})-\mathrm{CV}_y^{-2}. \quad (4.4)$$

From (4.4), it is clear that $\mathrm{Deft}_p^2(\hat{Y})\geq \mathrm{Deft}_p^2(\hat{\bar{Y}})$ if $\mathrm{Deft}_p^2(\hat{\bar{Y}})\geq 1$ and the equality holds if $\mathrm{Deft}_p^2(\hat{\bar{Y}})=1$ or $\bar{W}=N/n$. Also, $\mathrm{Deft}_p^2(\hat{Y})<\mathrm{Deft}_p^2(\hat{\bar{Y}})$ if $1/(1+\mathrm{CV}_y^2)<\mathrm{Deft}_p^2(\hat{\bar{Y}})<1$.

Example 4.1 shows that $\hat{Y}$ tends to have a larger design effect than $\hat{\bar{Y}}$ if the correlation between y_i and p_i is weak and $\mathrm{Deft}_p^2(\hat{\bar{Y}})\geq 1$.

The customary quantification of the effect of unequal weights on the design efficiency shown in (2.2) is due to Kish (1965, 11.7). He considered cases where the unequal weights arise from "haphazard" or "random" sources such as frame problems or non-response adjustments. Assuming that (1) a random sample of size n selected with replacement is divided into G weighting classes such that the same weight w_g is assigned to n_g sampling units within class g and $n=\sum_{g=1}^G n_g$, and that (2) all G weighting class variances are equal to the unit variance of y, *i.e.*, $S_{yg}^2=S_y^2$ for all $g=1,\ldots,G$, he proposed a quantity given as

$$\mathrm{Deft}_{\mathrm{Kish}}^2(\hat{\bar{Y}})=n\sum_{g=1}^G n_g\,w_g^2\Big/\left(\sum_{g=1}^G n_g\,w_g\right)^2, \quad (4.5)$$

to measure the increment in the variance of $\hat{\bar{Y}}$ in comparison with the hypothesized variance under srswr of size n. The rationale behind the above derivation is that the loss in precision of $\hat{\bar{Y}}$ due to haphazard unequal weighting can be approximated by the ratio of the variance under disproportionate stratified sampling to that under the proportionate stratified sampling.

In (4.5), letting $n_g=1$ for all g and thus, $n=G$, Kish (1992) later proposed a well-known approximate formula given as

$$\mathrm{Deft}_{\mathrm{Kish}}^2(\hat{\bar{Y}})=n\sum_{i=1}^n w_i^2\Big/\left(\sum_{i=1}^n w_i\right)^2=1+\mathrm{cv}_w^2, \quad (4.6)$$

where $\mathrm{cv}_w^2=n^{-1}\sum_{i=1}^n(w_i-\bar{w})^2/\bar{w}^2$ is the sample relative variance and $\bar{w}$ is the sample mean of w_i. Note that (4.6) is a sample approximate of (4.3). For a sampling design which is inefficient for estimation of Y, the inefficiency diminishes with the ratio estimation. Next, we consider the opposite case where the y-variable is correlated with the selection probability p_i, where the efficiency of $\hat{Y}$ increases.

Example 4.2 Unequal probability element sampling where y_i is correlated with p_i. Suppose that y_i is linearly related with p_i by $y_i=A+Bp_i+e_i$, where A and B are the least-square regression coefficients of the model for the (finite) population and e_i is the corresponding residual. Furthermore, assume that the regression model fits well to the population data and the error variance is roughly homogeneous so that $R_{ew}\cong 0$ and $R_{e^2w}\cong 0$, where R_{ew} and R_{e^2w} denote the population correlations of pairs (e_i,w_i) and (e_i^2,w_i), respectively. For example, $R_{ew}=\sum_{i=1}^N(e_i-\bar{E})(w_i-\bar{W})/\{(N-1)S_eS_w\}$, where $\bar{E}=\sum_{i=1}^N e_i/N$; S_e and S_w are the population standard deviations of e_i and w_i, respectively. Then the design effects given by (4.1) and (4.2) reduce to

$$\begin{aligned}\mathrm{Deft}_p^2(\hat{Y})&\cong(n\bar{W}/N)(1-R_{yp}^2)\\&\quad+(n\bar{W}/N-1)\left(\frac{R_{yp}}{\mathrm{CV}_p}-\frac{1}{\mathrm{CV}_y}\right)^2 \quad (4.7)\end{aligned}$$

and

$$\begin{aligned}\mathrm{Deft}_p^2(\hat{\bar{Y}})&\cong(n\bar{W}/N)(1-R_{yp}^2)\\&\quad+(n\bar{W}/N-1)\left(\frac{R_{yp}}{\mathrm{CV}_p}\right)^2, \quad (4.8)\end{aligned}$$

respectively, where R_{yp} is the population correlation between y_i and p_i and CV_p is the population coefficient of variation of p_i (see Park and Lee (2001) for proof). It follows from (4.7) and (4.8) that $\mathrm{Deft}_p^2(\hat{Y})\geq \mathrm{Deft}_p^2(\hat{\bar{Y}})$ if and only if

$$2R_{yp}\leq \mathrm{CV}_p/\mathrm{CV}_y, \quad (4.9)$$

where the equality holds if and only if $2R_{yp}=\mathrm{CV}_p/\mathrm{CV}_y$. Also, the inequality is reversed when the inequality in (4.9) becomes opposite.

The condition (4.9) indicates that $\hat{Y}$ tends to be less efficient in terms of precision than $\hat{\bar{Y}}$ whenever R_{yp} is small. Thus, we see that R_{yp} plays an important role in determining the design efficiency of unequal probability sampling on $\hat{Y}$ and $\hat{\bar{Y}}$ and their relative efficiency.

In an attempt to develop an approximate expression to the design effect when y_i is correlated with p_i, Spencer (2000) proposed a sample approximate formula for $\hat{Y}$ and compared it with Kish's approximate formula (4.6) for the special case of $R_{yp}=0$. As seen in example 4.2, the two design effects (4.7) and (4.8) are not equal unless $\bar{W}=N/n$ (see Park and Lee (2001) for more discussion

and some numerical examples). In addition, this special case provides the same condition as for example 4.1 and thus, the two approximate design effect formulae (4.7) and (4.8) are equivalent to (4.4) and (4.3), respectively.

4.2 One-Stage Cluster Sampling

Consider a one-stage cluster sampling, where every element in a sampled cluster is included in the sample, *i.e.*, $m_i \equiv M_i$ for all $i \in s_a$. Due to the absence of the second stage sampling variation, the variance of $\hat{Y}$ takes only the first term of expression (3.1) and it can be decomposed as

$$\sum_{i=1}^{N} w_i (Y_i - p_i Y)^2 = \frac{M(N-1)}{n} S_{yB}^2 + \sum_{i=1}^{N} w_i Q_i \bar{Y}_i^2, \quad (4.10)$$

where $S_{yB}^2 = (N-1)^{-1} \sum_{i=1}^{N} M_i (\bar{Y}_i - \bar{Y})^2$ and $Q_i = M_i(M_i - p_i M)$ for $i = 1, \ldots, N$. Note that $Q_i = 0$ if $p_i = M_i / M$, that is, p_i is proportional to the cluster size M_i. Also, note that S_{yB}^2 is the between-cluster mean square deviation in an analysis of variance. Denoting the within-cluster mean square deviation as $S_{yW}^2 = (M-N)^{-1} \sum_{i=1}^{N} \sum_{j=1}^{M_i} (y_{ij} - \bar{Y}_i)^2$, write $S_{yB}^2 = S_y^2 \{1 + \delta(M-N)/(N-1)\}$ with $\delta = 1 - S_{yW}^2 / S_y^2$. Since the expected sample size is $m' = n\bar{M}$, the design effect for $\hat{Y}$ can be written from (4.10) as

$$\mathrm{Deft}_p^2(\hat{Y}) = \left(\frac{N-1}{N}\right)\left(1 + \frac{M-N}{N-1}\delta\right) + \frac{n\bar{M}}{\mathrm{CV}_y^2} \sum_{i=1}^{N} \frac{w_i Q_i}{M_i^2} \left(\frac{Y_i}{Y}\right)^2. \quad (4.11)$$

Similarly, the design effect for $\hat{\bar{Y}}$ can be expressed as

$$\mathrm{Deft}_p^2(\hat{\bar{Y}}) \cong \left(\frac{N-1}{N}\right)\left(1 + \frac{M-N}{N-1}\delta\right) + \frac{n\bar{M}}{\mathrm{CV}_y^2} \sum_{i=1}^{N} \frac{w_i Q_i}{M_i^2} \left(\frac{D_i}{Y}\right)^2. \quad (4.12)$$

We observe that the design effect for $\hat{\bar{Y}}$ differs from that for $\hat{Y}$ in the second term containing $D_i = \sum_{j=1}^{M_i} (y_{ij} - \bar{Y})$ instead of Y_i. In addition, we note that the quantity $\delta = \delta_p(y)$ is the adjusted coefficient of determination (R_{adj}^2) in the regression analysis context. It may be called a homogeneity measure. For more discussion on δ, see Särndal *et al.* (1992, pages 130–131) and Lohr (1999, page 140).

Example 4.3 One-stage simple random sampling of clusters. In this example, if $p_i = 1/N$ for all $i = 1, \ldots, N$, the two design effects in (4.11) and (4.12) reduce, respectively, to

$$\mathrm{Deft}_p^2(\hat{Y}) = \left(\frac{N-1}{N}\right)\left(1 + \frac{M-N}{N-1}\delta\right) + \frac{1}{N \cdot \mathrm{CV}_y^2} \sum_{i=1}^{N} (M_i - \bar{M}) \left(\frac{M_i}{\bar{M}}\right)\left(\frac{\bar{Y}_i}{\bar{Y}}\right)^2 \quad (4.13)$$

and

$$\mathrm{Deft}_p^2(\hat{\bar{Y}}) \cong \left(\frac{N-1}{N}\right)\left(1 + \frac{M-N}{N-1}\delta\right) + \frac{1}{N \cdot \mathrm{CV}_y^2} \sum_{i=1}^{N} (M_i - \bar{M}) \left(\frac{M_i}{\bar{M}}\right)\left(\frac{\bar{D}_i}{\bar{Y}}\right)^2, \quad (4.14)$$

where $\bar{M} = M/N$. Since $\mathrm{Deft}_p^2(\hat{Y}) - \mathrm{Deft}_p^2(\hat{\bar{Y}}) \propto \sum_{i=1}^{N} M_i (M_i - \bar{M})(2\bar{Y}_i - \bar{Y})$, the inequality between design effects for $\hat{Y}$ and $\hat{\bar{Y}}$ depends on the joint distribution of $\bar{Y}_i$ and M_i.

Example 4.4 One-stage simple random sampling of clusters of equal-size. In this case, we have $M_i \equiv M_0$ and $p_i = 1/N$ for all $i = 1, \ldots, N$ and both design effects in (4.13) and (4.14) can be approximated by the same quantity given as

$$\left(\frac{N-1}{N}\right)\left[1 + \frac{N(M_0 - 1)}{N-1}\delta\right], \quad (4.15a)$$

since $M_i - \bar{M} = 0$ for all $i = 1, \ldots, N$.

To introduce the clustering effect on variance estimation, one often uses the simplest form of one-stage simple random cluster sampling as in example 4.4. For example, see Cochran (1977, section 9.4), Lehtonen and Pahkinen (1995, page 91), and Lohr (1999, section 5.2.2). Although these authors adopted a without-replacement sampling scheme, we compare their formulae with our formulae with the with – replacement sampling assumption for the sake of both simplicity and consistency. Furthermore, the comparison is valid because their formulae are defined with the finite population correction incorporated in both numerator and denominator so that its effect is basically cancelled out. Cochran (1977, section 9.4) derived

$$\mathrm{Deft}_p^2(\hat{\bar{Y}}) = \frac{NM_0 - 1}{M_0(N-1)}\left[1 + (M_0 - 1)\rho\right] \cong 1 + (M_0 - 1)\rho, \quad (4.15b)$$

where ρ is called the intracluster correlation coefficient defined by

$$\rho = \frac{2\sum_{i=1}^{N} \sum_{j>k=1}^{M_0} (y_{ij} - \bar{Y})(y_{ij} - \bar{Y})}{(M_0 - 1)\sum_{i=1}^{N} \sum_{j=1}^{M_0} (y_{ij} - \bar{Y})^2}. \quad (4.15c)$$

Rewriting $\sum_{i=1}^{N}[\sum_{j=1}^{M_0}(y_{ij}-\bar{Y})]^2 = M_0(N-1)S_{yB}^2$ and $\sum_{i=1}^{N}\sum_{j=1}^{M_0}(y_{ij}-\bar{Y})^2=(NM_0-1)S_y^2=(N-1)S_{yB}^2+N(M_0-1)S_{yW}^2$, it is easy to show that

$$2\sum_{i=1}^{N}\sum_{j>k=1}^{M_0}(y_{ij}-\bar{Y})(y_{ik}-\bar{Y})$$
$$=\sum_{i=1}^{N}\left[\sum_{j=1}^{M_0}(y_{ij}-\bar{Y})\right]^2-\sum_{i=1}^{N}\sum_{j=1}^{M_0}(y_{ij}-\bar{Y})^2$$
$$=(M_0-1)\left[(NM_0-1)S_y^2-NM_0S_y^2w\right]$$

and, thus, from (4.15c), $\rho=1-\{NM_0/(NM_0-1)\}(S_{yW}^2/S_y^2)\cong\delta$ assuming $M_i\equiv M_0$ for all $i=1,\ldots,N$, $NM_0/(NM_0-1)\cong 1$. Therefore, further assuming $(N-1)/N\cong 1$ and $(NM_0-1)M_0^{-1}(N-1)^{-1}\cong 1$, both design effect formulae (4.15a) and (4.15b) are approximately equivalent to $1+(M_0-1)\delta$. Other authors arrived at the same approximate formula. This is because δ and ρ essentially measure the same thing, which is the cluster homogeneity. Under this situation, two estimators $\hat{Y}$ and $\hat{\bar{Y}}$ have the same design effect as discussed in example 3.2. Note that this is a simple case of a self-weighting sampling design.

Särndal *et al.* (1992, section 8.7) compared the design effects for the two estimators under the setting of example 4.3. They also derived a simplified expression $1+(\bar{M}-1)\delta$ for (4.13) and (4.14), assuming the covariances of M_i with $M_i\bar{Y}_i^2$ and $M_i\bar{D}_i^2$ are ignorable. Their discussion on the difference between total and mean estimators boils down to Δ_a in example 3.2. They also noted that the design effect can be much more severe for the population total than for the population mean because more is lost through sampling of clusters when the total is estimated than when the mean is estimated.

A common practice to handle unequal cluster sizes is to use a more efficient sampling method that incorporates the size difference such as pps sampling of clusters. Expressions (4.11) and (4.12) can be applied to arbitrary selection probabilities p_i, where p_i are set to be proportional to some size measures $Z_i\geq 0$. The difference between the design effects for $\hat{Y}$ and $\hat{\bar{Y}}$ is explained by Δ_a in (3.9), or alternatively

$$\Delta_a=\frac{m'}{CV_y^2}\sum_{i=1}^{N}\frac{w_iQ_i}{M^2}\left[\left(\frac{\bar{Y}_i}{\bar{\bar{Y}}}\right)^2-\left(\frac{\bar{D}_i}{\bar{\bar{Y}}}\right)^2\right]. \qquad (4.16)$$

The term Q_i in (4.16) represents the effect of p_i on the variance estimation when size measures other than the actual cluster sizes M_i are used. Thomsen, Tesfu, and Binder (1986) considered the effect of an out-dated size measure among other factors under two-stage sampling with simple random sample of element at the second stage. We will come back to this in section 4.4.

4.3 Self-Weighting Designs

In a self-weighting sample, every sample element has the same weight. This leads to simple forms for both total and mean estimators. They are given by $\hat{Y}=y/f$ and $\hat{\bar{Y}}=y/m$, where $f=m/M$ is the overall sampling fraction and $y=\sum_{i=1}^{n}\sum_{j=1}^{m_i}y_{ij}$ is the sample total. Then just like simple random sampling as shown in (3.4), the two estimators have the same design effect.

A self-weighting sampling design can be implemented in various ways by synchronizing the first stage sampling method with the second stage sampling method (*e.g.*, Kish 1965, section 7.2). For example, if equal probability sampling is used for the first stage sampling, then the second stage should be sampled by an equal probability sampling method with a uniform sampling fraction for all PSUs. As a special case of this, where an srs of PSUs of equal size (*i.e.*, $M_i=M_0$ for all i) is selected, Hansen *et al.* (1953, Vol. II, pages 162 – 163) showed

$$CV_p^2(\hat{\bar{Y}})\cong\frac{1}{m}CV_y^2\left[1+\rho(\bar{m}-1)\right], \qquad (4.17)$$

where $CV_p^2(\hat{\bar{Y}})=V_p(\hat{\bar{Y}})/\bar{Y}^2$ is the relative variance of $\hat{\bar{Y}}$ under the sampling design p and ρ is the intracluster correlation coefficient as defined in (4.15c). Since the relative variance of $\hat{\bar{Y}}$ under srswr is $m^{-1}CV_{y\cdot}^2$, the well known approximate design effect formula for $\hat{\bar{Y}}$ under a self-weighting design follows immediately as

$$Deff_p^2(\hat{\bar{Y}})=1+\rho(\bar{m}-1). \qquad (4.18)$$

For one-stage cluster designs, we showed similar forms given in (4.15a) and (4.15b) (see also Yamane 1967, section 8.7). Hansen *et al.* (1953, Vol. II. page 204) further showed $CV_p^2(\hat{Y})=CV_p^2(\hat{\bar{Y}})$ for a sample design that employs simple random sampling at both stages. This implies that $\hat{Y}$ and $\hat{\bar{Y}}$ have the same design effect.

4.4 Two-Stage Unequal Probability Sampling

Let us first consider the following example.

Example 4.5 A two-stage sampling design where n PSUs are selected with replacement with probability p_i and an equal size simple random sample of $m_0\geq 2$ elements is selected with replacement from each selected PSU. With routine calculations and simplification, we can show that

$$Deff_p^2(\hat{\bar{Y}})\cong 1+(m_0-1)\tau+W_y^*, \qquad (4.19)$$

where

$$\tau=\frac{(N-1)S_{yB}^2+\sum_{i=1}^{N}(m_0-1)^{-1}S_{yi}^2}{(N-1)S_{yB}^2+\sum_{i=1}^{N}(M_i-1)S_{yi}^2}, \qquad (4.20)$$

$S_{yi}^2=(M_i-1)^{-1}\sum_{j=1}^{M_i}(y_{ij}-\bar{Y}_i)^2$, $W_y^*=W_y/V_{srswr}(\hat{Y}_{srs})=(m_0/CV_y^2)\sum_{i=1}^{N_i}(Q_i/p_iM^2)(\bar{Y}_i/\bar{Y})^2(1+CV_{yi}^2/m_0)$, and

$CV_{yi}^2 = S_{yi}^2 / \bar{Y}_i^2$ denotes the within-cluster relative variance of the y-variable. Similarly,

$$\text{Deft}_p^2(\hat{\bar{Y}}) \cong 1 + (m_0 - 1)\tau + W_d^*, \tag{4.21}$$

where $W_d^* = W_d / V_{\text{srswr}}(\hat{Y}_{\text{srs}}) = (m_0 / CV_y^2)\sum_{i=1}^N (Q_i / p_i M^2)(\bar{D}_i / \bar{Y})^2 (1 + CV_{di}^2 / m_0)$, and $\bar{D}_i$ and CV_{di}^2 are defined with the transformed variable $d(d_{ij} = y_{ij} - \bar{Y})$ analogously to $\bar{Y}_i$ and CV_{yi}^2, respectively. (Detailed derivations of expressions (4.19) and (4.21) are available from the authors.) For the case with $m_i = m_0$ for all i, the difference in the design effects given in (4.19) and (4.21) reduces to (3.7) or (4.16). There is no contribution from the second stage sampling to the difference.

Coming back to Thomsen *et al.* (1986) who studied the effect of using an outdated measure of size on the variance, the above discussion on $\hat{Y}$ parallels with their discussion. The only difference is that they assumed a without-replacement sampling scheme at the second stage. Note, however, that the definition of τ in Thomsen *et al.* (1986) is slightly different from (4.20) and from δ in section 4.2. However, there is a close connection between them. To see this, let us write the τ as a function of some quantities b_i's associated with PSUs as follows:

$$\tau(b_i) = \frac{(N-1)S_{yB}^2 - \sum_{i=1}^N b_i S_{yi}^2}{(N-1)S_{yB}^2 + \sum_{i=1}^N (M_i - 1)S_{yi}^2}.$$

Then the τ in Thomsen *et al.* (1986) is obtained with $b_i = 1$, the τ in example 4.5 with $-1/(m_0 - 1)$, and δ in section 4.2 with $(M_i - 1)/\{\sum_{i=1}^N (M_i - 1)/(N-1)\}$. Equating Kish's formula (4.18) for $\hat{\bar{Y}}$ to (4.19) for $\hat{Y}$, they obviously overlooked that the design effects for $\hat{Y}$ and $\hat{\bar{Y}}$ can be very different.

For more general cases, Kish (1987) proposed the following popular formula for $\hat{\bar{Y}}$:

$$\text{Deft}_{\text{Kish}}^2(\hat{\bar{Y}}) = \frac{n\sum_{g=1}^G n_g w_g^2}{\left(\sum_{g=1}^G n_g w_g\right)^2}[1 + \rho(\bar{m} - 1)]$$
$$= (1 + cv_w^2)[1 + \rho(\bar{m} - 1)].$$

This was obtained by applying (4.5) (or (4.6)) and (4.18) recursively to incorporate the effects of both clustering and unequal weights. Gabler, Haeder and Lahiri (1999) justified the above formula for $\hat{\bar{Y}}$ using a superpopulation model defined for the cross-classification of N clusters and G weighting classes. However, the difference between the design effects for $\hat{\bar{Y}}$ and $\hat{Y}$ cannot be exposed by such a model-based approach, since y_k is treated as a random variable while w_k as fixed. Under this approach, $\text{Deft}_p^2(\hat{Y})$ differs from $\text{Deft}_p^2(\hat{\bar{Y}})$ only by a factor of $(\hat{M}/M)^2$, although the actual difference can be much more pronounced as we have showed in this paper (*e.g.*, expressions (3.7) and (4.23)).

4.5 More General Cases

Weighting survey data involves not only sampling weights but also various weighting adjustments such as post-stratification, raking, and nonresponse compensation. We consider these general cases here.

We can rewrite the first-order Taylor approximation to the weighted mean estimator $\hat{\bar{Y}} = \hat{Y}/\hat{M}$ given in (3.2) as $(\hat{Y} - Y)/Y \cong (\hat{\bar{Y}} - \bar{Y})/\bar{Y} + (\hat{M} - M)/M$. Taking variance on both sides,

$$CV_p^2(\hat{Y}) \cong CV_p^2(\hat{\bar{Y}}) + CV_p^2(\hat{M}) + 2R_p(\hat{\bar{Y}}, \hat{M})\, CV_p(\hat{\bar{Y}})CV_p(\hat{M}), \tag{4.22}$$

where $CV_p^2(\hat{Y}), CV_p^2(\hat{\bar{Y}}), CV_p^2(\hat{M})$ are the relative variances of $\hat{Y}, \hat{\bar{Y}}$, and $\hat{M}$ respectively, and $R_p(\hat{\bar{Y}}, \hat{M})$ is the correlation coefficient of $\hat{\bar{Y}}$ and $\hat{M}$ with respect to the complex sampling design p and any weighting adjustments. Since the relative variances of simple sample total and mean $\hat{Y}_{\text{srs}}$ and $\bar{y}_{\text{srs}}$ are $CV_{\text{srswr}}^2(\hat{Y}_{\text{srs}}) = CV_{\text{srswr}}^2(\bar{y}_{\text{srs}}) = m^{-1}CV_y^2$ under srswr of size m, it follows from (4.22) that

$$\text{Deft}_p^2(\hat{Y}) \cong \text{Deft}_p^2(\hat{\bar{Y}}) + 2R_p(\hat{\bar{Y}}, \hat{M})\nabla_p(y)\text{Deft}_p(\hat{\bar{Y}}) + \nabla_p^2(y), \tag{4.23}$$

where $\nabla_p(y) = CV_p(\hat{M}) / CV_{\text{srswr}}(\bar{y}_{\text{srs}})$ is nonnegative. As an illustration, consider a binary variable y, where $CV_y^2 \cong (1 - \bar{Y})/\bar{Y}$ and, thus, $\nabla_p(y)$ can be arbitrarily large as $\bar{Y}$ approaches 1 or small as $\bar{Y}$ approaches zero assuming $CV_p(\hat{M}) \neq 0$. When $\nabla_p(y)$ is near zero, the two design effects are nearly equal. Otherwise, one is larger than the other depending on the values of $\nabla_p(y)$ and $R_p(\hat{\bar{Y}}, \hat{M})$. When the sampling weights are benchmarked to the known population size M, $\hat{Y}$ and $\hat{\bar{Y}}$ have the same design effect since $\hat{M} = M$ and $CV_p(\hat{M}) = 0$. In this case, $\hat{\bar{Y}}$ is not affected by the benchmarking but $\hat{Y} = M\hat{\bar{Y}}$, which is a ratio estimator. Note that poststratification or raking procedures may be used if population size information is available at subpopulation level and we also get equivalent design effects. In general, however, we have $\text{Deft}_p^2(\hat{Y}) \geq \text{Deft}_p^2(\hat{\bar{Y}})$ if

$$R_p(\hat{\bar{Y}}, \hat{M}) \geq -\frac{1}{2}\frac{\nabla_p(y)}{\text{Deft}_p(\hat{\bar{Y}})} \quad \text{or}$$

$$R_p(\hat{\bar{Y}}, \hat{M}) \geq -\frac{1}{2}\frac{CV_p(\hat{M})}{CV_p(\hat{\bar{Y}})}, \tag{4.24}$$

and vice versa.

It is illuminating to look at some specific situations. For example, if $R_p(\hat{\bar{Y}},\hat{M}) \geq 0$, then $\text{Deft}_p^2(\hat{Y}) > \text{Deft}_p^2(\hat{\bar{Y}})$, however, a negative correlation (*i.e.*, $R_p(\hat{\bar{Y}},\hat{M}) < 0$) doesn't necessarily lead to $\text{Deft}_p^2(\hat{Y}) \leq \text{Deft}_p^2(\hat{\bar{Y}})$. For a special case of $R_p(\hat{\bar{Y}},\hat{M}) = 0$, the difference is given by

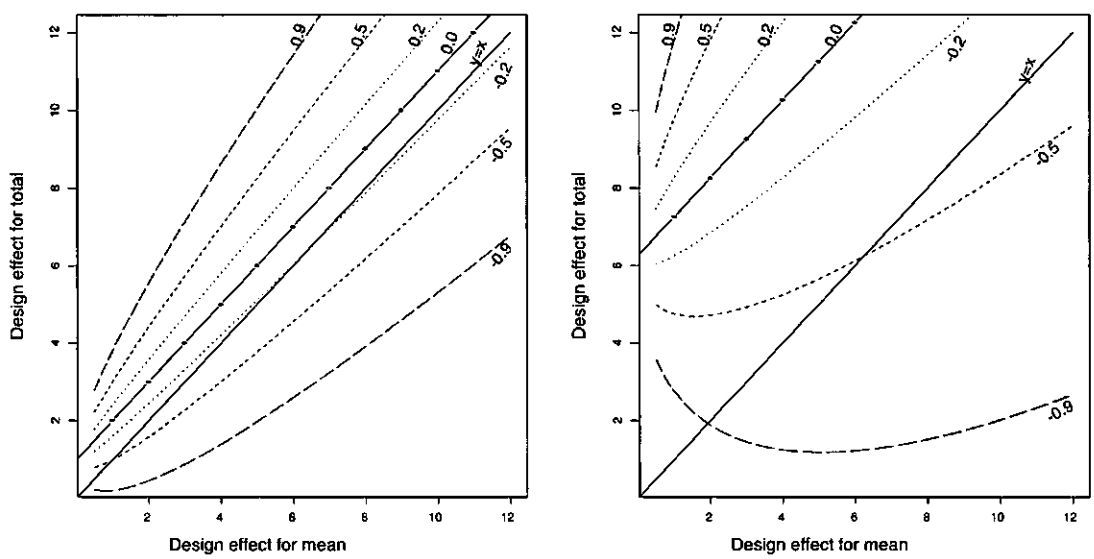
$$\text{Deft}_p^2(\hat{Y}) - \text{Deft}_p^2(\hat{\bar{Y}}) \cong \frac{\text{CV}_p^2(\hat{M})}{\text{CV}_{\text{srswr}}^2(\bar{y}_{\text{srs}})}. \qquad (4.25)$$

Figure 1 shows graphically the relation between the two design effects. The expression in (4.23) is plotted for some fixed values of $R_p(\hat{\bar{Y}},\hat{M})$ and $\nabla_p(y)$. The solid line passing through the origin which represents equal design effects is the reference line. As the graphs show, the comparison is not clear-cut. When $R_p(\hat{\bar{Y}},\hat{M}) < 0$, $\text{Deft}_p^2(\hat{Y}) \geq \text{Deft}_p^2(\hat{\bar{Y}})$ for small $\text{Deft}_p^2(\hat{\bar{Y}})$ but the relation flips over as $\text{Deft}_p^2(\hat{\bar{Y}})$ grows larger.

Hansen *et al.* (1953, Vol. I, pages 338 – 339) indicated that $R_p(\hat{\bar{Y}},\hat{M})$ would often be close to 0. Under this situation, expression (4.25) is also written as $\text{Deft}_p^2(\hat{Y}) \cong \text{Deft}_p^2(\hat{\bar{Y}})\,[1 + \text{CV}_p^2(\hat{M})/\text{CV}_p^2(\hat{\bar{Y}})]$, from which we get $\text{Deft}_p^2(\hat{Y}) \geq \text{Deft}_p^2(\hat{\bar{Y}})$. This special case was studied by Jang (2001). However, this doesn't seem necessary as can be seen in the following example.

Example 4.6 To illustrate the relationship between the design effects for $\hat{\bar{Y}}$ and $\hat{Y}$, we used a data set for the adults collected from the U.S. Third National Health and Nutrition Examination Survey (NHANES III), which is given as a demo file in WesVar version 4.0. NHANES III is a nationwide large-scale medical examination survey based on a stratified multistage sampling design, for which the Fay's modified balance repeated replication (BRR) method was employed for variance estimation. (See Judkins 1990 for more details on Fay's method.) We used only 19,793 records with complete responses to those characteristics listed in Table 1. Note that the weight in the demo file is different from the NHANES III final weight that was obtained by poststratification. For more detailed information on the demo file, see Westat (2001).

Table 1 presents the design effects for $\hat{\bar{Y}}$ and $\hat{Y}$, and component terms of (4.23) for the selected characteristics. Note that $\nabla_p(y)$ monotonically decreases in CV_y given that $m = 19{,}793$ and $\text{cv}_p(\hat{M}) = 3.2\%$. Although $\nabla_p(y)$ tends to be the determinant factor in the difference of the design effects, $R_p(\hat{\bar{Y}},\hat{M})$ can be important when it is negative. For example, for two race/ethnicity characteristics, African American and Hispanic, the negative values, –0.67 and –0.24 of $R_p(\hat{\bar{Y}},\hat{M})$ were responsible for $\text{Deft}_p^2(\hat{Y}) < \text{Deft}_p^2(\hat{\bar{Y}})$. Some design effects for $\hat{Y}$ are huge. This is not the case with the NHANES III poststratified final weights, with which $\hat{Y}$ and $\hat{\bar{Y}}$ have the same design effect. This illustrates the importance of benchmarking weight adjustments for total estimates.



(a) $\nabla_p(y) = 1.0$ (b) $\nabla_p(y) = 2.5$

Figure 1. Plots of $\text{Deft}_p^2(\hat{Y})$ versus $\text{Deft}_p^2(\hat{\bar{Y}})$ for (a) $\nabla_p(y) = 1.0$ (b) $\nabla_p(y) = 2.5$. The solid line corresponds to $\text{Deft}_p^2(\hat{\bar{Y}}) = \text{Deft}_p^2(\hat{Y})$. Other lines correspond to $R_p(\hat{\bar{Y}},\hat{M}) = -0.9, -0.5, -0.2, 0, 0.2, 0.5, 0.9$, respectively.

Table 1
Comparison of the design effects for the weighted total and mean using a subset of the adult data file from the U.S. Third National Health and Nutrition Examination Survey (NHANES III)

		Mean			Total						
Characteristic		Estimate	$Deft^2$	cv	Estimate	$Deft^2$	cv	cv_y	$r_p(\hat{\bar{Y}},\hat{M})$	$\nabla_p(y)$	$-\frac{cv_p(\hat{M})}{2cv_p(\hat{\bar{Y}})}$
Has smoked 100+ cigarettes in life?	Yes	0.53	4.13	0.014	98,397,795	31.31	0.038	0.944	0.20	4.83	−0.58
Has diabetes?	Yes	0.05	1.75	0.040	9,783,307	1.92	0.042	4.246	−0.34	1.07	−0.31
	No	0.95	1.75	0.002	176,341,218	393.47	0.033	0.236	0.34	19.35	−5.53
Has hypertension/ high blood pressure?	Yes	0.23	3.42	0.024	42,939,866	7.96	0.037	1.826	−0.18	2.50	−0.37
	No	0.77	3.42	0.007	143,184,660	78.44	0.034	0.548	0.18	8.32	−1.22
Race/Ethnicity	African American*	0.12	7.64	0.054	21,567,028	4.21	0.040	2.762	−0.67	1.65	−0.11
	Hispanic*	0.05	6.70	0.079	9,550,326	6.48	0.078	4.300	−0.24	1.06	−0.08
Gender	Male	0.48	1.40	0.009	88,725,967	19.18	0.033	1.048	−0.11	4.35	−1.55
	Female	0.52	1.40	0.008	97,398,559	25.39	0.034	0.954	0.11	4.77	−1.70
Number of cigarettes smoked per day	–	5.25	6.42	0.037	977,225,826	10.51	0.047	2.044	−0.09	2.23	−0.17
Population Size	–	–	–	–	186,124,526	–	0.032	–	–	–	–

Note: * denotes the cases where the design effect for $\hat{\bar{Y}}$ is smaller than that for $\hat{\bar{Y}}$.

5. CONCLUSION

We studied the design effects of the two most widely used estimators for the population mean and total in sample surveys under various with-replacement sampling schemes. We do not think the employment of with-replacement sampling is necessarily a serious limitation because we can see things more clearly without muddling the math with probably unnecessary complications with without-replacement sampling schemes. Furthermore, the effect of the finite population correction is largely canceled out in our formulation of the design effect and so the results are quite comparable with traditional design effects for without-replacement sampling. Therefore, our findings should be useful in practice. We summarize our key findings below.

Kish's well-known approximate formulae for the design effect for (ratio type) weighted mean estimators are not easily generalized in their form and concepts to more general problems, especially weighted total estimators contrary to what many people would perceive. In fact, $\hat{\bar{Y}}$ and $\hat{\bar{Y}}$ often have very different design effects unless the sampling design is self-weighting or the sampling weights are benchmarked to the known population size. In addition, the design effect is in general not free from the distribution of the study variable even for the mean estimator, let alone the total estimator. Furthermore, the correlation of the study variable with the weights used in estimation can be an important factor in determining the design effect. Therefore, apart from its original intention, the design effect measures not only the effect of a complex sampling design on a particular statistic but also the effects of the distribution of the study variable and its relations to the sampling design on the statistic. As complex survey software packages routinely produce the design effect, it seems appropriate to warn the user of the packages of these rather obscure facts about the design effect.

ACKNOLWEDGEMENT

The authors thank Louis Rizza at Westat, an associate editor, and two referees for their helpful comments and suggestions on an earlier version of this paper.

REFERENCES

APOSTOL, T.M. (1974). *Mathematical Analysis.* 2nd Ed. Reading, MA: Addison-Wesley.

BARRON, E.W., and FINCH, R.H. (1978). Design Effects in a complex multistage sample: The Survey of Low Income Aged and Disabled (SLIAD), *Proceedings of the Section on Survey Research Methods,* American Statistical Association. 400-405.

COCHRAN, W.G. (1977). *Sampling Techniques.* 3rd Ed. New York: John Wiley & Sons, Inc.

CORNFIELD, J. (1951). Modern methods in the sampling of human populations. *American Journal of Public Health.* 41, 654-661.

GABLER, S., HAEDER, S. and LAHIRI, P. (1999). A model based justification of Kish's formula for design effects for weighting and clustering. *Survey Methodology.* 25, 105-106.

HANSEN, M.H., HURWITZ, W.N. and MADOW, W.G. (1953). *Sample Survey Methods and Theory.* Vol. I, New York: John Wiley & Sons, Inc.

HANSEN, M.H., HURWITZ, W.N. and MADOW, W.G. (1953). *Sample Survey Methods and Theory,* Vol. II, New York: John Wiley & Sons, Inc.

JUDKINS, D.R. (1990). Fay's method for variance estimation, *Journal of Official Statistics*. 6, 223-239.

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

KISH, L. (1987). Weighting in Deft^2. *The Survey Statistician*. June 1987.

KISH, L. (1992). Weighting for unequal p_i. *Journal of Official Statistics*. 8, 183-200.

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

JANG, D. (2001). On procedures to summarize variances for survey estimates. *Proceedings of the Survey Research Methods of the American Statistical Association*. In CD-ROM.

LEHTONEN, R., and PAHKINEN, E.J. (1995). *Practical Methods for Design and Analysis of Complex Surveys*. New York: John Wiley & Sons, Inc.

LOHR, S.L. (1999). *Sampling: Design and Analysis*. Pacific Grove, CA: Brooks/Cole.

PARK, I., and LEE, H. (2001). The design effect: do we know all about it? *Proceedings of the Section on Survey Research Methods*, American Statistical Association. In CD-ROM.

PARK, I., and LEE, H. (2002). A revisit of design effects under unequal probability sampling. *The Survey Statistician*. 46, 23-26.

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

SPENCER, B.D. (2000). An approximate design effect for unequal weighting when measurements may correlate with selection probabilities. *Survey Methodology*. 26, 137-138.

THOMSEN, I., TESFU, D. and BINDER, D.A. (1986). Estimation of Design Effects and Intraclass Correlations When Using Outdated Measures of Size. *International Statistical Review*. 54, 343-349.

WESTAT (2001). *WesVar 4.0 User's Guide*. Rockville, MD: Westat, Inc.

YAMANE, T. (1967). *Elementary Sampling Theory*. New Jersey: Prentice-Hall.

Survey Methodology, December 2004
Vol. 30, No. 2, pp. 195-208
Statistics Canada

Robust Generalized Regression Estimation

JEAN-FRANÇOIS BEAUMONT and ASMA ALAVI [1]

ABSTRACT

The Best Linear Unbiased (BLU) estimator (or predictor) of a population total is based on the following two assumptions: i) the estimation model underlying the BLU estimator is correctly specified and ii) the sampling design is ignorable with respect to the estimation model. In this context, an estimator is robust if it stays close to the BLU estimator when both assumptions hold and if it keeps good properties when one or both assumptions are not fully satisfied. Robustness with respect to deviations from assumption (i) is called model robustness while robustness with respect to deviations from assumption (ii) is called design robustness. The Generalized Regression (GREG) estimator is often viewed as being robust since its property of being Asymptotically Design Unbiased (ADU) is not dependent on assumptions (i) and (ii). However, if both assumptions hold, the GREG estimator may be far less efficient than the BLU estimator and, in that sense, it is not robust. The relative inefficiency of the GREG estimator as compared to the BLU estimator is caused by widely dispersed design weights. To obtain a design-robust estimator, we thus propose a compromise between the GREG and the BLU estimators. This compromise also provides some protection against deviations from assumption (i). However, it does not offer any protection against outliers, which can be viewed as a consequence of a model misspecification. To deal with outliers, we use the weighted generalized M-estimation technique to reduce the influence of units with large weighted population residuals. We propose two practical ways of implementing M-estimators for multipurpose surveys; either the weights of influential units are modified and a calibration approach is used to obtain a single set of robust estimation weights or the values of influential units are modified. Some properties of the proposed approach are evaluated in a simulation study using a skewed finite population created from real survey data.

KEY WORDS: Design robustness; Model robustness; M-estimator; Outliers; Shrunk weights; Best linear unbiased predictor.

1. INTRODUCTION

In classical theory, sample data can be viewed as being randomly drawn from an infinite population and assumptions are made about the unknown distribution of the infinite population. In other words, a model is postulated and the interest lies in the estimation of model parameters. In this context, an estimator $\hat{\theta}$ of a model parameter θ is robust if it stays close to the maximum likelihood estimator of θ when the model assumptions hold and if it keeps good properties when the model assumptions are not fully satisfied. The unknown distribution of the infinite population is often assumed to be the normal distribution and, as a result, the maximum likelihood estimator reduces to the usual least-squares estimator.

The presence of outliers in the sample can be viewed as a consequence of a deviation from a model assumption. The majority of the sample could be assumed to come from the selected model but some units, called outliers, could be thought of as coming from a different model. Therefore, the presence of such outliers in the sample may introduce bias and increase the variance of the least-squares estimator of the selected model parameters. Outliers could also be the consequence of a highly skewed distribution. In this case, the least-squares estimator is not biased but may be highly inefficient due to a deviation from the usual normality assumption. The presence of outliers in the sample could also be the result of measurement errors. However, it is assumed in the rest of this paper that the data have been verified and corrected, if necessary, and that there is no measurement error left in the data. Outlier-robust estimation for infinite populations has been studied extensively (for a review, see Huber 1981; or Hampel, Ronchetti, Rousseeuw and Stahel 1986).

In survey sampling theory, the interest usually lies in the estimation of finite population parameters such as the total, $t_y = \sum_{k \in U} y_k$, of a variable of interest y for a finite population U of size N. Because it is usually not possible to observe the variable y for all population units, the usual practice consists of selecting from the finite population a random sample s of size n according to some probability sampling design $p(s \mid \mathbf{Z})$. The matrix of design information $\mathbf{Z}$ contains N rows with its k^{th} row equal to $\mathbf{z}_k'$, and $\mathbf{z}$ is a vector of auxiliary variables available at the design stage. This does not preclude the finite population itself to be assumed to come from a model, as it is explicitly the case when it is chosen to make model-based inferences. Under this type of inference, Royall (1976) derived the Best Linear Unbiased (BLU) estimator (or predictor) $\hat{t}_y^B$ of t_y (see also Valliant, Dorfman and Royall 2000, Chapter 2). It is based

[1] Jean-François Beaumont and Asma Alavi, Household Survey Methods Division, Statistics Canada, 16th floor, R.H. Coats Building, Ottawa, Ontario, Canada, K1A 0T6. E-mail: Jean-Francois.Beaumont@statcan.ca and Asma.Alavi@statcan.ca.

on the following two assumptions: i) the estimation model underlying the BLU estimator $\hat{t}_y^B$ is correctly specified and ii) the sampling design is ignorable with respect to the estimation model. In this context, an estimator $\hat{t}_y$ of the finite population total t_y is robust if it stays close to the BLU estimator $\hat{t}_y^B$ when both assumptions hold and if it keeps good properties when one or both assumptions are not fully satisfied. Robustness with respect to deviations from assumption (i) is called model robustness while robustness with respect to deviations from assumption (ii) is called design robustness.

Although we consider robust estimators that are constructed from a model-based viewpoint, we prefer evaluating their properties as much as possible with respect to the sampling design. This allows us to choose the constants on which robust estimators depend and to evaluate their quality without having to rely on a model and, more specifically, without having to rely on a model for the outliers. This also provides an objective framework for comparing estimators derived under different models. This preference of evaluating properties of model-based estimators with respect to the sampling design is also shared by Little (1983) who notes that design-based asymptotics may be more useful for assessing estimators than model-based asymptotics, particularly when the data set is large.

The Generalized Regression (GREG) estimator of t_y is often viewed as being robust since its property of being Asymptotically Design Unbiased (ADU) is not dependent on assumptions (i) and (ii); that is, the GREG estimator is bias-robust even though its form can be justified by an estimation model. However, if both assumptions hold, the GREG estimator may be far less efficient than the BLU estimator and, in that sense, it is not robust. The relative inefficiency of the GREG estimator as compared to the BLU estimator is caused by widely dispersed design weights. The fact that variable design weights may increase the variance of an estimator is well known (see, for example, Rao 1966; DuMouchel and Duncan 1983; Kish 1992; Pfeffermann 1993; Korn and Graubard 1999, Chapter 4; Elliott and Little 2000; and Kalton and Flores-Cervantes 2003) and is not uncommon in household surveys due to the presence of many weight adjustments before calibration (Kish 1992; and Kalton and Flores-Cervantes 2003). This problem is often treated by truncating the larger design weights (Potter 1988, 1990, 1993; and Stokes 1990).

To obtain a design-robust estimator when the design weights are highly variable, we propose a compromise between the GREG and the BLU estimators based on the weighted Least-Squares (LS) technique. This compromise estimator has a smaller design bias than the BLU estimator when the ignorability assumption is not satisfied and, at the same time, is more efficient than the GREG estimator when this assumption holds. It also provides some protection against deviations from model assumptions. Balanced sampling (Royall and Herson 1973) and nonparametric calibration (Chambers, Dorfman and Wehrly 1993) are other methods that provide protection against certain types of model misspecifications (see also Valliant, Dorfman and Royall 2000, Chapter 3, 4 and 11). However, none of these methods offer any protection against outliers, which can be viewed as a consequence of a model misspecification. In a model-based framework, the idea underlying the *M*-estimation technique has been proposed to develop outlier-robust alternatives to the BLU estimator (Chambers 1986; Lee 1991; and Welsh and Ronchetti 1998). In a design-based framework, the *M*-estimation technique has also been used to develop outlier-robust alternatives to the GREG estimator (Gwet and Rivest 1992; Hulliger 1995 1999; Duchesne 1999; and Zaslavsky, Schenker and Belin 2001). *M*-estimation is also discussed in the review paper by Lee (1995) and an empirical comparison of several outlier-robust estimators can be found in Gwet and Lee (2000).

Finite population parameters are often very sensitive to the presence of outliers in the population. This is to be contrasted to model (infinite population) parameters, which are usually insensitive to outliers. The problem of outlier robustness is therefore different for finite and infinite populations. As noted in Chambers (1986), it is the sampling error (or the prediction error in a model-based framework) of an estimator which must be insensitive to outliers in finite populations and not necessarily the estimator itself. For instance, when a simple random sampling design is used, the sample median is robust in the classical sense. As a result, its design variance is essentially unaffected by the presence of an outlier in the finite population, no matter how large is that outlier. However, the sampling error and the design bias of the sample median, when used as an estimator of the finite population mean, take an arbitrarily large value when one or more population unit takes an arbitrarily large value. This is explained by the fact that the finite population mean itself takes an arbitrarily large value in such a case. Unlike the sample median, the sample mean is design unbiased but it is not robust in the classical sense. The sampling error and the design variance of the sample mean can thus be very affected by the presence of an outlier in the finite population. This illustrates why outlier-robustness for finite populations is often viewed as a trade-off between bias and variance and why outliers must usually have an influence, at least to some extent, on estimators. The Mean Squared Error (MSE) is therefore a useful criterion for evaluating the quality of outlier-robust estimators of finite population parameters.

The real goal of this paper is to find a robust alternative to the commonly-used GREG estimator of t_y. However, it

is more natural to discuss robustness issues by first introducing the optimal (BLU) estimator. Therefore, the assumptions underlying the BLU estimator are discussed in section 2. We also give additional conditions under which the BLU estimator has a negligible asymptotic design bias. Section 3 deals with design robustness and the weighted LS estimator is introduced. In section 4, model robustness (more specifically, outlier robustness) is discussed and the weighted generalized *M*-estimation technique is suggested to reduce the influence of units with large weighted population residuals. The proposed estimator is census-consistent in the sense that it is equal to the finite population total t_y when a census is conducted. We propose two practical ways of implementing *M*-estimators for multipurpose surveys; either the weights of influential units are modified and a calibration approach is used to obtain a single set of robust estimation weights or the values of influential units are modified. Mean Squared Error (MSE) estimation is discussed in section 5. In section 6, some properties of the proposed approach are evaluated in a simulation study using a skewed finite population created from real survey data. Finally, some concluding remarks are made in the last section.

2. THE BEST LINEAR UNBIASED ESTIMATOR

Let us assume that we have a vector of auxiliary variables $\mathbf{x}$ available for all units of the sample s and for which population totals, $\mathbf{t}_x = \sum_{k\in U} \mathbf{x}_k$, are known. Let us also denote by $\mathbf{X}$, the matrix containing N rows with its k^{th} row equal to $\mathbf{x}_k'$. The vector $\mathbf{x}$ may or may not contain some variables in the vector $\mathbf{z}$ of design variables. Before discussing robustness, we first describe the two assumptions (see A1 and A2 below) with respect to which robustness is desired. Then, we briefly explain how to validate them.

A1) The following estimation model m holds: y_k given $\mathbf{X}$, for $k \in U$, are independently distributed with mean $\mathrm{E}_m(y_k \mid \mathbf{X}) = \mathbf{x}_k'\boldsymbol{\beta}$ and variance $\mathrm{V}_m(y_k \mid \mathbf{X}) = \sigma^2 v_k$, where $\boldsymbol{\beta}$ and σ^2 are unknown model parameters, $v_k = \mathbf{x}_k'\boldsymbol{\lambda}$ and $\boldsymbol{\lambda}$ is a vector of known constants. The subscript "m" indicates that expectations and variances are evaluated with respect to model m.

A2) The sampling design is independent of $\mathbf{y}$ after conditioning on $\mathbf{X}$; that is, $p(s \mid \mathbf{y}, \mathbf{X}) = p(s \mid \mathbf{X})$, where $\mathbf{y}$ is a vector containing N elements with its k^{th} element equal to y_k.

Assumption (A1) describes the estimation model m, which specifies the distribution of $\mathbf{y}$ conditional on $\mathbf{X}$. Standard techniques can be used to validate this model (see, for example, Draper and Smith 1980, Chapter 3). The linearity assumption $\mathrm{E}_m(y_k \mid \mathbf{X}) = \mathbf{x}_k'\boldsymbol{\beta}$ is an important assumption underlying the estimation model m. There are many ways of assessing the validity of this assumption. A graph of residuals $e_k = y_k - \mathbf{x}_k'\hat{\boldsymbol{\beta}}$ versus $\mathbf{x}_k'\hat{\boldsymbol{\beta}}$, for some m-unbiased estimator $\hat{\boldsymbol{\beta}}$ of $\boldsymbol{\beta}$, is often suggested for this purpose. Any trend in this graph is an indication that the relationship between $\mathbf{y}$ and $\mathbf{x}$ is not linear. To obtain robustness against a deviation from the linearity assumption, a poststratification model can be used when it is possible to partition the population into homogeneous and mutually exclusive groups. An example of the importance of careful modeling in sample surveys can be found in Hedlin, Falvey, Chambers and Kokic (2001).

Assumption (A2) is a sufficient condition for the ignorability (Rubin 1976) of the sampling design with respect to the distribution of $\mathbf{y}$ conditional on $\mathbf{X}$. In other words, it means that the distribution of $\mathbf{y}$ is independent of s after conditioning on $\mathbf{X}$. Using assumption (A1), $\mathbf{y}$ can be split into a fixed term $\mathbf{X}\boldsymbol{\beta}$ and a random error term $\boldsymbol{\varepsilon} = \mathbf{y} - \mathbf{X}\boldsymbol{\beta}$. Consequently, if the sampling design is independent of $\boldsymbol{\varepsilon}$ after conditioning on $\mathbf{X}$; that is, if $p(s \mid \boldsymbol{\varepsilon}, \mathbf{X}) = p(s \mid \mathbf{X})$, then assumption (A2) is satisfied and the sampling design is ignorable. Since we only consider sampling designs of the form $p(s \mid \mathbf{Z})$, an obvious way to make the sampling design ignorable is achieved by including all design variables $\mathbf{z}$ into the estimation model. Examples of such design variables may include the variables used to form the strata, the variable used as a size measure if probability-proportional-to-size sampling is used and so on. The design weights may also provide a useful summary of the design information. Note that it may not be necessary to include all design variables into the estimation model (see Sugden and Smith 1984). Design variables that are independent of $\mathbf{y}$ (or $\boldsymbol{\varepsilon}$) after conditioning on $\mathbf{X}$ should not be included. To assess the validity of assumption (A2), a graph of the residuals, $e_k = y_k - \mathbf{x}_k'\hat{\boldsymbol{\beta}}$, versus design weights w_k (or any design variable) may be useful (see Pfeffermann 1993). Any trend in this graph suggests that the design weights are correlated with the random error $\boldsymbol{\varepsilon}$ and that the sampling design is not ignorable with respect to the estimation model. More formal tests can also be performed to assess the validity of this assumption (see, for example, DuMouchel and Duncan 1983; Graubard and Korn 1993; and, for more references on this topic, Pfeffermann 1993).

Under the estimation model m and the ignorability assumption (A2), it is easy to show that the BLU estimator (Royall 1976) $\hat{t}_y^B$ of t_y takes the simple projection form $\hat{t}_y^B = \mathbf{t}_x'\hat{\mathbf{B}}^{\mathbf{B}}$, where $\hat{\mathbf{B}}^{\mathbf{B}}$ is implicitly defined by the equation

$$\sum_{k\in s}(y_k - \mathbf{x}_k'\hat{\mathbf{B}}^{\mathbf{B}})\frac{\mathbf{x}_k}{v_k} = \mathbf{0}. \qquad (2.1)$$

The BLU estimator can also be written as $\hat{t}_y^B = \sum_{k\in s} w_k^B y_k$, where the BLU estimation weights w_k^B are given by

$$w_k^B = \frac{\mathbf{x}_k'}{v_k}\left(\sum_{k\in s}\frac{\mathbf{x}_k\mathbf{x}_k'}{v_k}\right)^{-1}\mathbf{t}_x. \qquad (2.2)$$

The model variance $V_m\{(\hat{t}_y^B - t_y)\,|\,s,\mathbf{X}\}$ of $\hat{t}_y^B$ is the smallest for every possible sample among all linear m-unbiased estimators of t_y. A direct consequence of this result is that the anticipated variance $E_m\{E_p(\hat{t}_y^B - t_y)^2\,|\,\mathbf{X}\}$ of $\hat{t}_y^B$ is also the smallest among all linear m-unbiased estimators of t_y, where the subscript p indicates that the expectation is evaluated with respect to the sampling design. Under the additional assumption that y_k given $\mathbf{X}$ follows a normal distribution, $\hat{\mathbf{B}}^B$ is also the maximum likelihood estimator of the vector of model parameters $\boldsymbol{\beta}$.

In general, the BLU estimator $\hat{t}_y^B$ is not ADU. However, under the estimation model m, the ignorability assumption (A2) and the following additional assumption (A3), the BLU estimator has the property of being Asymptotically Design Unbiased in Probability (ADUP) in the sense that its relative design bias $E_p(\hat{t}_y^B - t_y)/t_y$ converges in probability to 0 as n and N increase without bound.

A3) $\sum_{k\in U} E_p\{(w_k^B)^2 I_k\}\sigma_k^2 = O(N)$, $\sum_{k\in U}\mathbf{x}_k'\boldsymbol{\beta} = O(N)$ and $\sum_{k\in U}\sigma_k^2 = O(N)$, where $\sigma_k^2 = \sigma^2 v_k$ and I_k is a dummy random variable indicating whether unit k is selected in the sample ($I_k = 1$) or not ($I_k = 0$).

Assumption (A3) describes the asymptotic behaviour of three population quantities. In particular, requiring that $\sum_{k\in U} E_p\{(w_k^B)^2 I_k\}\sigma_k^2 = O(N)$ essentially means that none of the BLU estimation weights becomes too large as the sample size and the population size increase. For instance, if $\mathbf{x}_k = v_k = 1$ and if a sampling design of fixed size n is used, then condition $\sum_{k\in U} E_p\{(w_k^B)^2 I_k\}\sigma_k^2 = O(N)$ is equivalent to assuming that the weights $w_k^B = N/n$ remain bounded as both n and N grow. The proof that $\hat{t}_y^B$ is ADUP is given in the appendix and does not require that $v_k = \mathbf{x}_k'\boldsymbol{\lambda}$. As a result, the BLU estimator is ADUP even when the model variance $V_m(y_k\,|\,\mathbf{X})$ is misspecified.

As pointed out above, the BLU estimator is efficient when the estimation model m and the normality assumption hold as well as the ignorability assumption (A2). Under these assumptions and the additional assumption (A3), the BLU estimator is also ADUP. Consequently, a first step towards robustness consists of selecting and validating an estimation model such that these assumptions are satisfied as much as possible. However, they are rarely fully satisfied in practice. For example, one can be reluctant to include all strata identifiers into the estimation model when the number of strata is very large. In such a case, the ignorability assumption might not fully hold. Also, the estimation model, including the normality assumption, may not hold for every variable of interest. Consequently, the non-critical use of the BLU estimator $\hat{t}_y^B$ of t_y is not always appropriate and robust estimators may be needed.

3. DESIGN ROBUSTNESS

Using the fact that $v_k = \mathbf{x}_k'\boldsymbol{\lambda}$, it can be easily shown (see Särndal, Swensson and Wretman, 1992, page 231) that t_y can be expressed as $t_y = \mathbf{t}_x'\mathbf{B}$, where $\mathbf{B}$ is implicitly defined by the equation

$$\sum_{k\in U}(y_k - \mathbf{x}_k'\mathbf{B})\frac{\mathbf{x}_k}{v_k} = \mathbf{0}. \qquad (3.1)$$

The vector $\mathbf{B}$ would be the LS estimator of $\boldsymbol{\beta}$, under the estimation model m, if a census could be conducted. Since $\mathbf{t}_x$ is known, the objective of finding an estimator of the population total t_y is thus equivalent to finding an estimator of $\mathbf{B}$. In the design-based theory, a natural estimator $\hat{\mathbf{B}}^G$ of $\mathbf{B}$ is implicitly defined by the equation

$$\sum_{k\in s} w_k(y_k - \mathbf{x}_k'\hat{\mathbf{B}}^G)\frac{\mathbf{x}_k}{v_k} = \mathbf{0}, \qquad (3.2)$$

where w_k, the design weight of unit k, equals to the inverse of the selection probability π_k. The use of $\hat{\mathbf{B}}^G$ leads to the GREG estimator $\hat{t}_y^G = \mathbf{t}_x'\hat{\mathbf{B}}^G$ of t_y. The GREG estimator $\hat{t}_y^G$ takes a simple projection form because $v_k = \mathbf{x}_k'\boldsymbol{\lambda}$ (see Särndal *et al.* 1992, page 231). It can also be written as $\hat{t}_y^G = \sum_{k\in s} w_k^G y_k$, where the GREG estimation weights w_k^G are given by

$$w_k^G = w_k\frac{\mathbf{x}_k'}{v_k}\left(\sum_{k\in s} w_k\frac{\mathbf{x}_k\mathbf{x}_k'}{v_k}\right)^{-1}\mathbf{t}_x. \qquad (3.3)$$

As pointed out in the introduction, the GREG estimator is bias-robust since its property of being ADU is not dependent on the validity of the estimation model m and the ignorability assumption. However, the GREG estimator is not variance-robust since it may be far less efficient than the BLU estimator when both assumptions hold. The inefficiency of the GREG estimator is due to widely dispersed design weights. In household surveys, this situation is not uncommon because of many weight adjustments before calibration. Also, practical considerations for the choice of a sampling design combined with limited information available at the design stage often lead to sampling designs that are approximately ignorable. In household surveys, for instance, geographic information is often the main auxiliary information available to construct the strata. Unless the number of strata is very large, such information is usually weakly correlated with quantitative variables of interest, such as *expenditures* or *income*, and their corresponding population residual variable $E = y - \mathbf{x}'\mathbf{B}$. As a result, the design weight variable w is also weakly correlated with E. This suggests that the ignorability assumption may approximately hold. This also suggests that the design weights act more or less as a random noise when estimating $\mathbf{B}$ using (3.2) and that their influence could be significantly reduced. To obtain a design-robust estimator when the

design weights are highly variable, we thus propose to shrink the design weights towards their mean and to use the LS estimator $\hat{t}_y^{LS} = \mathbf{t}_x' \hat{\mathbf{B}}^{LS}$, where $\hat{\mathbf{B}}^{LS}$ is implicitly defined by

$$\sum_{k \in s} \tilde{w}_k (y_k - \mathbf{x}_k' \hat{\mathbf{B}}^{LS}) \frac{\mathbf{x}_k}{v_k} = \mathbf{0} \qquad (3.4)$$

and where $\tilde{w}_k$ is the shrunk weight of unit k given by

$$\tilde{w}_k = \left(\frac{\sum_{k \in s} w_k}{\sum_{k \in s} g(w_k; \alpha)} \right) g(w_k; \alpha). \qquad (3.5)$$

The reason for the ratio in the right side of (3.5) is simply to ensure that $\sum_{k \in s} \tilde{w}_k = \sum_{k \in s} w_k$ and the role of the function $g(w_k; \alpha)$ is to obtain shrunk weights $\tilde{w}_k$ that are less variable than the design weights w_k. This function is assumed to be monotone in the constant α, with $1 \leq g(w_k; \alpha) \leq w_k$. The BLU and GREG estimators are therefore extreme special cases of the LS estimator obtained when α is such that $g(w_k; \alpha) = 1$ and $g(w_k; \alpha) = w_k$ respectively. To obtain a simple compromise between these two extreme estimators, we suggest using $g(w_k; \alpha) = w_k^{\alpha}$, with $0 \leq \alpha \leq 1$. The choice $\alpha = 0$ leads to the BLU estimator while the choice $\alpha = 1$ leads to the GREG estimator. In fact, this suggestion was proposed by Kish (1992, page 198). Other functions $g(w_k; \alpha)$ and other ways of reducing the variability of design weights can be found in the literature (see, for example, Elliott and Little 2000). Truncating large design weights ($g(w_k; \alpha) = \min(w_k, \alpha)$, with $\alpha > 0$) is a common approach that deals with this problem. This approach may be useful when assumptions (A1) and (A2) are not fully satisfied and when there are some abnormally large design weights. A better approach may be to truncate large weighted residuals. The weighted generalized M-estimation technique discussed in the next section can be used for this purpose.

The LS estimator $\hat{t}_y^{LS}$ can also be written as $\hat{t}_y^{LS} = \sum_{k \in s} w_k^{LS} y_k$, where the LS estimation weights w_k^{LS} are given by

$$w_k^{LS} = \tilde{w}_k \frac{\mathbf{x}_k'}{v_k} \left(\sum_{k \in s} \tilde{w}_k \frac{\mathbf{x}_k \mathbf{x}_k'}{v_k} \right)^{-1} \mathbf{t}_x. \qquad (3.6)$$

Note that the estimation weights w_k^{LS}, including w_k^B and w_k^G as special cases, are calibrated on the known population totals $\mathbf{t}_x$ in the sense that they satisfy the calibration equation $\sum_{k \in s} w_k^{LS} \mathbf{x}_k = \mathbf{t}_x$ (see Deville and Särndal 1992).

4. MODEL (OUTLIER) ROBUSTNESS

As pointed out in the introduction, the LS estimator $\hat{t}_y^{LS}$ provides some protection against deviations from the ignorability assumption and also against deviations from model assumptions. However, it does not offer any protection against outliers, which can be viewed as a consequence of a model misspecification, including a deviation from the normality assumption. For instance, the GREG estimator is ADU no matter the validity of the estimation model. However, its design variance may be very large in the presence of outliers in the finite population because they may greatly influence its sampling error when they are selected in the sample. This problem may be amplified when the design weights are widely dispersed. For the Horvitz-Thompson estimator, this was well illustrated in the circus example of Basu (1971). Of course, the use of efficient auxiliary variables at the estimation stage can control the impact of outliers on estimates. However, such auxiliary variables are often not available and outlier-robust estimators may provide significant gains over the LS estimator.

Using the Taylor linearization technique (see, for example, Särndal *et al.* 1992, page 235) and given that $t_y = \mathbf{t}_x' \mathbf{B}$, it is well known and easy to show that the sampling error of the GREG estimator can be approximated as follows: $\hat{t}_y^G - t_y \approx \sum_{k \in s} w_k E_k$, where $E_k = y_k - \mathbf{x}_k' \mathbf{B}$ is the population residual for unit k. As a result, a large design weight associated with a large population residual (or outlier) may have a substantial impact on the quality of the GREG estimator. Moreover, it is straightforward to show that the sampling error of the LS estimator can be expressed as $\hat{t}_y^{LS} - t_y = \sum_{k \in s} w_k^{LS} E_k$. Therefore, a large estimation weight associated with a large population residual may greatly influence the sampling error and the quality of the LS estimator. To deal with this problem, we use the Schweppe version (Hampel *et al.* 1986, pages 315 – 316) of the weighted generalized M-estimation technique to reduce the influence of units with large weighted population residuals. This leads to the M-estimator $\hat{\mathbf{B}}^M$ of $\mathbf{B}$, which is implicitly defined by

$$\sum_{k \in s} \tilde{w}_k \frac{1}{h_k} \psi\left(\frac{h_k \tilde{E}_k(\hat{\mathbf{B}}^M)}{Q} \right) \frac{\mathbf{x}_k}{\sqrt{v_k}} = \mathbf{0}, \qquad (4.1)$$

where $\tilde{E}_k(\hat{\mathbf{B}}^M) = (y_k - \mathbf{x}_k' \hat{\mathbf{B}}^M)/\sqrt{v_k}$, Q is a positive population scale parameter and h_k is a weight that may depend not only on $\mathbf{x}_k$ but also on $\mathbf{z}_k$. The role of the function $\psi(.)$ consists of reducing the influence of units with a large $h_k \tilde{E}_k(\mathbf{B})$. From the above considerations, $h_k = w_k^{LS} \sqrt{v_k}$ or $h_k = \tilde{w}_k \sqrt{v_k}$ is a natural choice. In the former case, the influence of large $w_k^{LS} E_k$ is reduced while, in the latter case, the influence of large $\tilde{w}_k E_k$ is reduced. The choice $h_k = w_k^{LS} \sqrt{v_k}$ may be preferred to $h_k = \tilde{w}_k \sqrt{v_k}$ when there are outliers in the auxiliary variables $\mathbf{x}$ or when α is not close to 1 (assuming $g(w_k; \alpha) = w_k^{\alpha}$). The main point here is that h_k should depend on survey weights w_k^{LS} or $\tilde{w}_k$ and that both choices suggested above should perform better than simpler choices that do not take into

account the auxiliary variables $\mathbf{z}$ such as $h_k = \sqrt{v_k}$ or $h_k = 1$, which reduce the influence of large unweighted residuals. Also, it should again be noted that the interest is in finding a robust estimator for the vector of population parameters $\mathbf{B}$ and not for the vector of model parameters $\boldsymbol{\beta}$. In fact, $\mathbf{B}$ is itself not robust (in the classical sense) for $\boldsymbol{\beta}$ since it may be highly affected by the presence of outliers in the finite population. As a result, outliers must have a certain influence on $\hat{\mathbf{B}}^{\mathbf{M}}$.

Equation (4.1) can be written in the weighted linear regression form:

$$\sum_{k\in s} \tilde{w}_k^*(\hat{\mathbf{B}}^{\mathbf{M}}, Q)\,(y_k - \mathbf{x}_k'\hat{\mathbf{B}}^{\mathbf{M}})\,\frac{\mathbf{x}_k}{v_k} = 0, \qquad (4.2)$$

where

$$\tilde{w}_k^*(\hat{\mathbf{B}}^{\mathbf{M}}, Q) = \tilde{w}_k\,\frac{\psi(r_k)}{r_k}$$

and

$$r_k = \frac{h_k \tilde{E}_k(\hat{\mathbf{B}}^{\mathbf{M}})}{Q}.$$

We propose the following modification of the popular function $\psi(.)$ of Huber (1964) that makes the adjusted weights $\tilde{w}_k^*(\hat{\mathbf{B}}^{\mathbf{M}}, Q)$ always greater than or equal to 1: $\psi(r_k) = r_k$, if $|r_k| \le \varphi$, and $\psi(r_k) = \text{sign}(r_k)\,\max(|r_k|/\tilde{w}_k, \varphi)$, otherwise, where φ is a positive constant. This leads to adjusted weights

$$\tilde{w}_k^*(\hat{\mathbf{B}}^{\mathbf{M}}, Q) = \begin{cases} \tilde{w}_k, & \text{if } |r_k| \le \varphi, \\ \max\left(1, \tilde{w}_k\,\dfrac{\varphi}{|r_k|}\right), & \text{otherwise}. \end{cases} \qquad (4.3)$$

The Iteratively Reweighted Least-Squares (IRLS) algorithm (Beaton and Tukey 1974) is often used to solve (4.2) and (4.3). At a given iteration i, the adjusted weights $\tilde{w}_k^*(\mathbf{B}^{(i-1)}, Q^{(i-1)})$ are first calculated using (4.3) and then $\mathbf{B}^{(i)}$ is obtained by solving (4.2) with $\tilde{w}_k^*(\hat{\mathbf{B}}^{\mathbf{M}}, Q)$ and $\hat{\mathbf{B}}^{\mathbf{M}}$ replaced by $\tilde{w}_k^*(\mathbf{B}^{(i-1)}, Q^{(i-1)})$ and $\mathbf{B}^{(i)}$ respectively. To obtain $\mathbf{B}^{(i)}$, an estimate of Q is usually calculated at each iteration of the IRLS algorithm. In the simulation study of section 6, we have used

$$Q^{(i-1)} = 1.483 \times \text{weighted sample median of} \left(\left|h_k \tilde{E}_k\left(\mathbf{B}^{(i-1)}\right)\right| ; k \in s\right), \qquad (4.4)$$

where the weighted sample median is calculated using the weights $\tilde{w}_k / h_k$. Equation (4.4) reduces to the proposal of Hulliger (1999) when $h_k = 1$ and $g(w_k; \alpha) = w_k$. We suggest using $\mathbf{B}^{(0)} = \hat{\mathbf{B}}^{\mathbf{LS}}$ as the vector of starting values since $\hat{\mathbf{B}}^{\mathbf{LS}}$ is easy to obtain. The iterative procedure is normally repeated until convergence is reached. To reduce computer time, especially if a resampling method is used for MSE estimation, a single iteration of the IRLS algorithm can be performed. In section 6, it is shown empirically that performing a single iteration yields an estimator of the population total that has properties similar to the fully-iterated estimator. This point has also been noted by Lee (1991).

The M-estimator of t_y is given by $\hat{t}_y^M = \mathbf{t}_{\mathbf{x}}'\hat{\mathbf{B}}^{\mathbf{M}}$. With the restriction that $\tilde{w}_k^*(\hat{\mathbf{B}}^{\mathbf{M}}, \hat{Q}) \ge 1$, where $\hat{Q}$ is an estimator of Q, the estimators $\hat{\mathbf{B}}^{\mathbf{M}}$ and $\hat{t}_y^M$ are census-consistent in the sense that they are exactly equal to $\mathbf{B}$ and t_y respectively, no matter the value of φ and α, when a census is conducted ($\pi_k = 1$, for $k \in U$). This restriction might be useful for controlling the design bias of $\hat{t}_y^M$ when there are shrunk weights $\tilde{w}_k$ close to 1. Note that the estimators $\hat{\mathbf{B}}^{\mathbf{M}}$ and $\hat{t}_y^M$ reduce to $\hat{\mathbf{B}}^{\mathbf{LS}}$ and $\hat{t}_y^{\mathrm{LS}}$ respectively when $\varphi = \infty$ ($\psi(r_k) = r_k$). The M-estimator $\hat{t}_y^M$ can also be expressed as $\hat{t}_y^M = \sum_{k\in s} w_k^M y_k$, where the M-estimation weights w_k^M are given by

$$w_k^M = \tilde{w}_k^*(\hat{\mathbf{B}}^{\mathbf{M}}, \hat{Q})\,\frac{\mathbf{x}_k'}{v_k}\left(\sum_{k\in s} \tilde{w}_k^*(\hat{\mathbf{B}}^{\mathbf{M}}, \hat{Q})\,\frac{\mathbf{x}_k\mathbf{x}_k'}{v_k}\right)^{-1}\mathbf{t}_{\mathbf{x}}. \qquad (4.5)$$

The estimation weights w_k^M are still calibrated on the known population totals $\mathbf{t}_{\mathbf{x}}$ ($\sum_{k\in s} w_k^M \mathbf{x}_k = \mathbf{t}_{\mathbf{x}}$).

In order to determine appropriate values for α and φ, the MSE of the M-estimator $\hat{t}_y^M$ can be estimated for different choices of α and φ using past or current sample data. Then, the values of α and φ that give the smallest estimated MSE can be chosen. Estimation of MSE is discussed in section 5. As noted in Hulliger (1995), choosing adaptively α and φ by minimizing the estimated MSE with current sample data leads to an estimator $\hat{t}_y^M$ that does not require estimating the scale parameter Q. Also, this procedure controls the magnitude of the design bias of $\hat{t}_y^M$ without requiring the use of additional constants. However, it is likely to provide less efficiency than using the optimal (although unknown) values of α and φ.

In multipurpose surveys, different values of α and φ are likely to be obtained for different variables of interest. If multiple sets of weights are to be avoided, some form of compromise is needed. As a first step towards a compromise, a common value of α, satisfactory for the most important variables of interest, can be determined. Then, we propose two practical ways of implementing the M-estimator $\hat{t}_y^M$ without having to find a compromise value for φ; either the weights of influential units are modified and a calibration approach is used to obtain a single set of robust estimation weights or the values of influential units

are modified. The former is discussed in section 4.1 while the latter is discussed in section 4.2.

4.1 Modification of the Weights of Influential Units

Let us now assume that it is desired to estimate the population totals of a vector of q variables of interest $\mathbf{y} = (y_1, y_2, ..., y_q)'$. A vector of q M-estimators $\hat{\mathbf{t}}_y^{\mathbf{M}} = (\hat{t}_{y_1}^M, \hat{t}_{y_2}^M, ..., \hat{t}_{y_q}^M)'$ of $\mathbf{t}_y = \sum_{k \in U} \mathbf{y}_k$ can be obtained, with potentially different values of φ for different variables. To simplify the notation, we denote the adjusted weights associated with variable y_i by $\tilde{w}_k^*(y_i)$, for $i = 1, 2, ..., q$. Since the adjusted weights $\tilde{w}_k^*(y_i)$ depend on the variable of interest y_i, we obtain q sets of weights, even if a common value of φ is chosen.

Gwet and Rivest (1992), Duchesne (1999) and Hulliger (1999) suggested using the adjusted weights $\tilde{w}_k^*(\mathbf{y}) = \min(\tilde{w}_k^*(y_1), \tilde{w}_k^*(y_2), ..., \tilde{w}_k^*(y_q))$ to obtain a unique set of weights. Then, estimation weights $w_k^M(\mathbf{y})$ are calculated by replacing $\tilde{w}_k^*(\hat{\mathbf{B}}^{\mathbf{M}}, \hat{Q})$ by $\tilde{w}_k^*(\mathbf{y})$ in (4.5) and $\mathbf{t}_y$ is estimated by $\sum_{k \in s} w_k^M(\mathbf{y})\, \mathbf{y}_k$. Although the estimation weights $w_k^M(\mathbf{y})$ are calibrated on the known population totals $\mathbf{t}_x$, they are not calibrated on the vector of estimates $\hat{\mathbf{t}}_y^{\mathbf{M}}$, which are believed to be our best estimates in the sense of minimizing the estimated MSE. Moreover, the use of $\sum_{k \in s} w_k^M(\mathbf{y})\, \mathbf{y}_k$ likely leads to a larger design bias than $\hat{\mathbf{t}}_y^{\mathbf{M}}$ although it controls the design variance. To cope with these issues, we propose computing the estimation weights $w_k^{M,A}(\mathbf{y})$ by replacing $\tilde{w}_k^*(\hat{\mathbf{B}}^{\mathbf{M}}, \hat{Q})$ by the adjusted weights $\tilde{w}_k^*(\mathbf{y})$ in (4.5), and by augmenting the vector of auxiliary variables $\mathbf{x}$ and the known population totals $\mathbf{t}_x$ using $\mathbf{y}$ and $\hat{\mathbf{t}}_y^{\mathbf{M}}$ respectively. As a result, the estimation weights $w_k^{M,A}(\mathbf{y})$ are calibrated on $\mathbf{t}_x$ and $\hat{\mathbf{t}}_y^{\mathbf{M}}$, and $\mathbf{t}_y$ is estimated by $\hat{\mathbf{t}}_y^{\mathbf{M}} = \sum_{k \in s} w_k^{M,A}(\mathbf{y})\, \mathbf{y}_k$. Of course, there may be a limit on the number of variables that can be used for calibration purposes. This may somewhat restrict the applicability of this method when q is very large.

4.2 Modification of the Values of Influential Units

Another way of implementing the M-estimator $\hat{\mathbf{t}}_y^{\mathbf{M}}$ in practice consists of modifying the values of the variables of interest $\mathbf{y}$ and using the LS estimation weights w_k^{LS} for all variables. This can be done separately for each variable of interest, so we return to the case of only one variable of interest in this section.

Let us first denote by s_o the random set of all sample units k for which $\tilde{w}_k^*(\hat{\mathbf{B}}^{\mathbf{M}}, \hat{Q}) \neq \tilde{w}_k$. In other words, s_o is the random set of units that have been detected as being influential. Let also $\hat{\mathbf{B}}^{\mathbf{M}^*}$ be implicitly defined by the equation

$$\sum_{k \in s} \tilde{w}_k (y_{*k} - \mathbf{x}_k' \hat{\mathbf{B}}^{\mathbf{M}^*}) \frac{\mathbf{x}_k}{v_k} = \mathbf{0}, \qquad (4.6)$$

where $y_{*k} = y_k$, if $k \in s - s_o$, and $y_{*k} = y_k^*$, otherwise. The quantity y_k^* is a modified value for the influential unit k that is used to replace y_k. Note that $\hat{\mathbf{B}}^{\mathbf{M}^*} = \hat{\mathbf{B}}^{\mathbf{LS}}$ if $y_{*k} = y_k$, for $k \in s$. The population total t_y can then be estimated by $\hat{t}_y^{M^*} = \mathbf{t}_x' \hat{\mathbf{B}}^{\mathbf{M}^*}$. It is also easy to show that $\hat{t}_y^{M^*} = \sum_{k \in s} w_k^{LS} y_{*k}$.

The idea here consists of finding modified values y_k^*, for $k \in s_o$, as close as possible to the original values y_k and that satisfy the constraint $\hat{\mathbf{B}}^{\mathbf{M}^*} = \hat{\mathbf{B}}^{\mathbf{M}}$. Under this constraint, it is obvious that $\hat{t}_y^{M^*} = \hat{t}_y^M$. A possible implementation of this idea is obtained by minimizing the distance function $\sum_{k \in s_o} \tilde{w}_k (y_k - y_k^*)^2 / v_k$ subject to the constraint $\hat{\mathbf{B}}^{\mathbf{M}^*} = \hat{\mathbf{B}}^{\mathbf{M}}$. This leads to the modified values

$$y_k^* = y_k + \mathbf{x}_k' \left(\sum_{k \in s_o} \frac{\tilde{w}_k}{v_k} \mathbf{x}_k \mathbf{x}_k' \right)^{-1} \left(\sum_{k \in s} \frac{\tilde{w}_k}{v_k} \mathbf{x}_k \mathbf{x}_k' \right) \left(\hat{\mathbf{B}}^{\mathbf{M}} - \hat{\mathbf{B}}^{\mathbf{LS}} \right). \qquad (4.7)$$

This idea is essentially equivalent to reverse calibration proposed by Ren and Chambers (2002), except that these authors used the constraint $\hat{t}_y^{M^*} = \hat{t}_y^M$ instead of $\hat{\mathbf{B}}^{\mathbf{M}^*} = \hat{\mathbf{B}}^{\mathbf{M}}$. We prefer the latter since it leads to modified values that better preserve the relationships between the variable of interest y and the auxiliary variables $\mathbf{x}$.

Other ways of determining modified values that satisfy the constraint $\hat{\mathbf{B}}^{\mathbf{M}^*} = \hat{\mathbf{B}}^{\mathbf{M}}$ can be found. For example, it is straightforward to show that this constraint is satisfied when the following modified values are used:

$$y_k^* = a_k y_k + (1 - a_k) \mathbf{x}_k' \hat{\mathbf{B}}^{\mathbf{M}}, \qquad (4.8)$$

where $a_k = \tilde{w}_k^*(\hat{\mathbf{B}}^{\mathbf{M}}, \hat{Q}) / \tilde{w}_k$. The modified values in equation (4.8) have a simple interpretation: they are a weighted average of the robust prediction $\mathbf{x}_k' \hat{\mathbf{B}}^{\mathbf{M}}$ and the observed value y_k. Less weight is given to the observed value y_k when it has a smaller value of a_k and, therefore, when it is highly influential.

5. MEAN SQUARED ERROR ESTIMATION

Estimation of the MSE of $\hat{t}_y^M$ can be used for three different purposes: i) finding appropriate values for α and φ using past or current sample data, ii) evaluating the quality of estimates and iii) making inferences about unknown population quantities. Using the fact that $\mathrm{E}_p(\hat{t}_y^G) \approx t_y$, it can be easily shown that the MSE of $\hat{t}_y^M$ can be approximated by

$$\mathrm{MSE}_p(\hat{t}_y^M) \approx \mathrm{V}_p(\hat{t}_y^M) + \mathrm{E}_p(\hat{t}_y^M - \hat{t}_y^G)^2 - \mathrm{V}_p(\hat{t}_y^M - \hat{t}_y^G). \qquad (5.1)$$

The last two terms of (5.1) are equal to $[\mathrm{E}_p(\hat{t}_y^M - \hat{t}_y^G)]^2$. They represent the square of the design bias of $\hat{t}_y^M$. As suggested in Gwet and Rivest (1992), a potential estimator of $\mathrm{MSE}_p(\hat{t}_y^M)$ is given by

$$\mathrm{mse}_p(\hat{t}_y^M) = \hat{\mathrm{V}}_p(\hat{t}_y^M) + \max\left(0\ ,\ (\hat{t}_y^M - \hat{t}_y^G)^2 - \hat{\mathrm{V}}_p(\hat{t}_y^M - \hat{t}_y^G)\right), \quad (5.2)$$

where $\hat{\mathrm{V}}_p(\hat{t}_y^M)$ and $\hat{\mathrm{V}}_p(\hat{t}_y^M - \hat{t}_y^G)$ are estimators of $\mathrm{V}_p(\hat{t}_y^M)$ and $\mathrm{V}_p(\hat{t}_y^M - \hat{t}_y^G)$ respectively.

Since the estimator $\hat{t}_y^M$ has a complex structure, resampling variance estimation methods provide a convenient way of estimating $\mathrm{V}_p(\hat{t}_y^M)$ and $\mathrm{V}_p(\hat{t}_y^M - \hat{t}_y^G)$. The jackknife, the bootstrap and the balanced repeated replications methods are described and evaluated in Rao, Wu and Yue (1992) for stratified multistage sampling designs, where the primary sampling units are assumed to have been selected with replacement. They have shown in an empirical study that the jackknife variance estimator can have a large bias when estimating the variance of a non-smooth estimator, such as the sample median. Therefore, the jackknife variance estimator might be more biased for estimating the variance of the M-estimator than the balanced repeated replication or the bootstrap method when, at each iteration of the IRLS algorithm, Q is estimated using a non-smooth estimator such as (4.4). Gwet and Lee (2000) studied empirically the performance of the jackknife and the bootstrap methods for some robust estimators. In general, they found encouraging results. It is important to note that the estimator $\hat{t}_y^M$ should be recomputed for each resample. This includes repeating the procedure used to estimate α and φ if they are estimated using current sample data.

When the goal of MSE estimation is only to find appropriate values for α and φ, it may be convenient to consider simplified MSE estimators in order to reduce computer time. We now propose four different ways of simplifying MSE estimation:

i) Only a single iteration of the IRLS algorithm could be done for each resample even if a fully-iterated M-estimator is used. This might yield reasonable variance estimates since the singly-iterated and fully-iterated M-estimators seem to have similar properties (see section 6.4).

ii) Some quantities could be assumed fixed (not random) for MSE estimation. This is likely to lead to an underestimation of the MSE but it may be useful if the goal of MSE estimation is only to find appropriate values for α and φ. For example, the adjusted weights $\tilde{w}_k^*(\hat{\mathbf{B}}^{\mathbf{M}}, \hat{Q})$ could be assumed fixed. This approximation was in fact suggested in Hulliger (1999). Alternatively, if the M-estimator is implemented using the methodology in section (4.2), the modified values in (4.7) or (4.8) could be treated as true values for MSE estimation.

iii) The term $\hat{\mathrm{V}}_p(\hat{t}_y^M - \hat{t}_y^G)$ in (5.2) could be omitted. This would lead to the MSE estimator: $\mathrm{mse}_p(\hat{t}_y^M) = \hat{\mathrm{V}}_p(\hat{t}_y^M) + (\hat{t}_y^M - \hat{t}_y^G)^2$. Note that this approach leads to an overestimation of the MSE.

iv) A combination of two of the above three propositions could be considered. For example, the adjusted weights $\tilde{w}_k^*(\hat{\mathbf{B}}^{\mathbf{M}}, \hat{Q})$ could be assumed fixed and the term $\hat{\mathrm{V}}_p(\hat{t}_y^M - \hat{t}_y^G)$ in (5.2) could be omitted. In such a case, an estimator for $\mathrm{V}_p(\hat{t}_y^M)$ could be obtained by noting that $\mathrm{V}_p(\hat{t}_y^M) = \mathbf{t}_{\mathbf{x}}' \mathbf{V}_p(\hat{\mathbf{B}}^{\mathbf{M}}) \mathbf{t}_{\mathbf{x}}$ and by using the well known Taylor linearization technique of Binder (1983) to estimate $\mathbf{V}_p(\hat{\mathbf{B}}^{\mathbf{M}})$. After some straightforward algebra, we obtain the MSE estimator

$$\mathrm{mse}_p(\hat{t}_y^M) = \sum_{k \in s} \sum_{l \in s} \frac{(\pi_{kl} - \pi_k \pi_l)}{\pi_{kl}} w_k^M (y_k - \mathbf{x}_k' \hat{\mathbf{B}}^{\mathbf{M}}) w_l^M (y_l - \mathbf{x}_l' \hat{\mathbf{B}}^{\mathbf{M}}) + (\hat{t}_y^M - \hat{t}_y^G)^2, \quad (5.3)$$

where π_{kl} is the joint probability of selection of units k and l.

6. SIMULATION STUDY

We performed a simulation study to evaluate some properties of the LS estimator and the M-estimator for a skewed finite population. In particular, we compared a version of the M-estimator that reduces the influence of large weighted population residuals to another one that reduces the influence of large unweighted population residuals. We also compared the performance of the singly- and fully-iterated M-estimators. Section 6.1 describes the population and the sampling design, and sections 6.2 to 6.4 discuss results from the simulation.

6.1 Population and Sampling Design

The data from Statistics Canada's 1998 Survey of Household Spending (SHS) are used to serve as the population. This survey uses a stratified multi-stage design and contains information about 15,457 households on several variables. The variable *Renovation/Repair* is chosen as the variable of interest y. This variable is considered for its greater potential of having very large values. A vector $\mathbf{x}$ of three binary auxiliary variables have been created by dividing the variable *Income* into three categories (*Income* $\leq$ 30,000, 30,000 < *Income* $\leq$ 60,000 and *Income* > 60,000) and we have chosen $v_k = 1$, for all $k \in U$. In other words, we have considered a poststratification estimation model, which should give us robustness against deviations from the linearity assumption. The population coefficient of

determination (R^2) for this estimation model is 0.13. This is a typical R^2 in household surveys.

From this population, 5,000 samples of expected sample size 300 have been selected using Poisson sampling. We wanted to give households quite dispersed probabilities of selection resulting in variable design weights. We thus assigned probabilities of selection such that they were proportional to the inverse of the SHS design weights (which include a nonresponse adjustment factor). The selection probabilities are thus given by $\pi_k = (300/\sum_{k\in U}\pi_k^*)\,\pi_k^*$, where π_k^*, for $k \in U$, is the reciprocal of the design weight (including a nonresponse adjustment factor) from the SHS data.

Table 6.1 gives some summary statistics for this population. We note that the population residuals are very skewed and that the skewness increases when the residuals are multiplied by the design weights. Figure 6.1 shows a graph of the population residuals versus the design weights. First, we note that there is a clear outlier with a residual greater than 50,000 and with a design weight not close to 1. Fortunately, the most extreme design weights are not associated with large population residuals. Also, although this graph may be misleading because of the huge number of points that are overlapping, there does not seem to be any clear relationship between the population residuals and the design weights. In fact, the coefficient of correlation between the design weights and the population residuals is 0.0049. Such a small coefficient of correlation is not atypical in household surveys, for reasons discussed in section 3, and suggests that the ignorability assumption may hold approximately.

Table 6.1
Summary Statistics about the Population

Variable	Mean	Standard Deviation	Skewness
Renovation/Repair	367	1,124	12.6
Population Residual	0	1,104	12.8
Design Weight	177	170	1.8
Weighted Population Residual	922	295,685	15.0

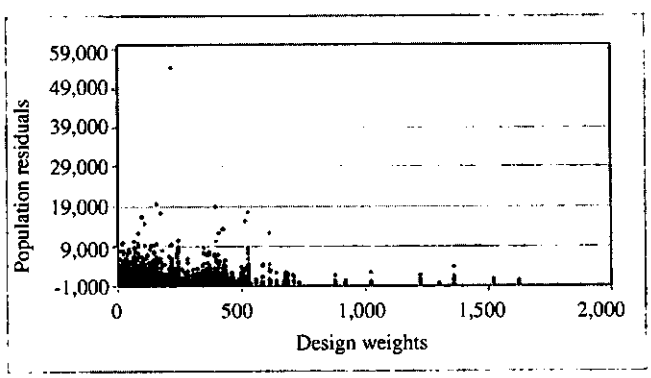


Figure 6.1. Graph of the population residuals versus the design weights

For each of the 5,000 samples, estimates of the population total for the *Renovation/Repair* variable have been calculated for both the LS estimator and two versions of the *M*-estimator; one that reduces the influence of large weighted population residuals ($h_k = \tilde{w}_k$) and another one that reduces the influence of large unweighted population residuals ($h_k = 1$). For the i^{th} sample, the relative error in percentage of any estimate $\hat{t}_{yi}$ of t_y is defined as $\Delta_i = 100\% \times (\hat{t}_{yi} - t_y)/t_y$. The Relative Bias (RB) and the Relative Root Mean Squared Error (RRMSE) of any estimator $\hat{t}_y$, expressed as a percentage of the population total, can thus be estimated by RB $= \sum_{i=1}^{5,000} \Delta_i/5{,}000$ and RRMSE $= \sqrt{\sum_{i=1}^{5,000} \Delta_i^2/5{,}000}$ respectively. Another measure of interest is the Maximum Absolute Relative Error (MARE) in percentage given by MARE $= \max(|\Delta_i|;\ i = 1, 2, \ldots, 5{,}000)$. This measure may be useful to assess the sensitivity of an estimator to the presence of influential units in the sample.

6.2 The LS Estimator: Design Robustness

In this section, we evaluate the properties of the LS estimator. Figure 6.2 illustrates the RB, RRMSE and MARE of the LS estimator for 11 values of α ($\alpha = 0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1$) when $g(w_k;\alpha) = w_k^\alpha$. On the one hand, the BLU estimator ($\alpha = 0$) has an RRMSE close to the minimum and the smallest MARE among these 11 values of α but, as expected, leads to the largest RB (in absolute value). Its RB is equal to -13.05%, which is not negligible. Given that a poststratification model is used, this suggests that the ignorability assumption is not fully satisfied even though the correlation between the design weights and the population residuals is small. On the other hand, the GREG estimator ($\alpha = 1$) has a very small RB but has the largest RRMSE and MARE due to the variability of the design weights. When $\alpha = 0.2$, the LS estimator is biased, with an RB of -9.11%, but has a value of MARE relatively close to the smallest value and has the smallest RRMSE (17.94%) among the values of α considered. This is a substantial reduction in comparison with the RRMSE of the GREG estimator (34.77%). In general, values of α between 0.2 and 0.5 provide a reasonable compromise estimator with respect to RB, RRMSE, and MARE. Note that, for larger expected sample sizes, we expect that the minimum MSE be reached for larger values of α because the bias of the LS estimator may dominate its variance.

We have also considered the LS estimator obtained by choosing adaptively, for each selected sample, the value of α that leads to the smallest estimated MSE among the set of 11 values of α considered above. The MSE has been estimated using equation (5.3). The average value of α over the 5,000 selected samples is 0.43. This is slightly larger

than the value of α (0.2) that leads to the smallest MSE (see figure 6.2). This may be due to the simplification made to obtain (5.3), which omits a component of the square design bias when estimating the MSE. Nevertheless, this LS estimator shows a significant improvement over the GREG estimator in terms of RRMSE (26.05%) and MARE (217.99%). This LS estimator shows also a significant improvement over the BLU estimator in terms of RB (−6.24%). Therefore, it seems that choosing adaptively the value of α leads to a useful compromise between the GREG and BLU estimators. However, there is a price to pay in terms of RRMSE by estimating α instead of using the optimal (although unknown) value of α.

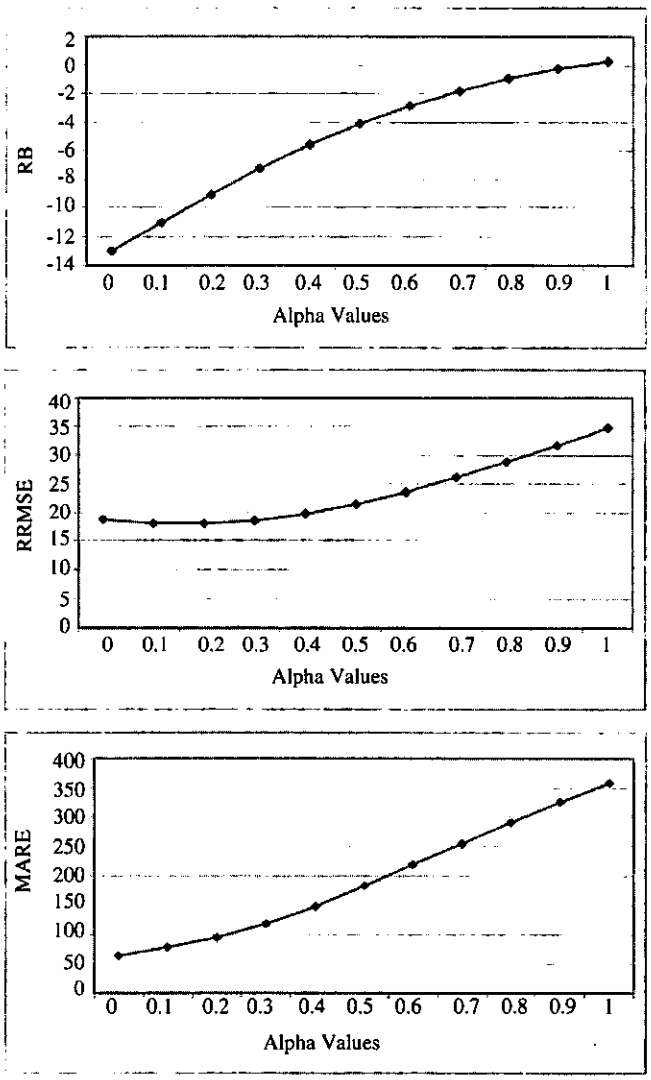


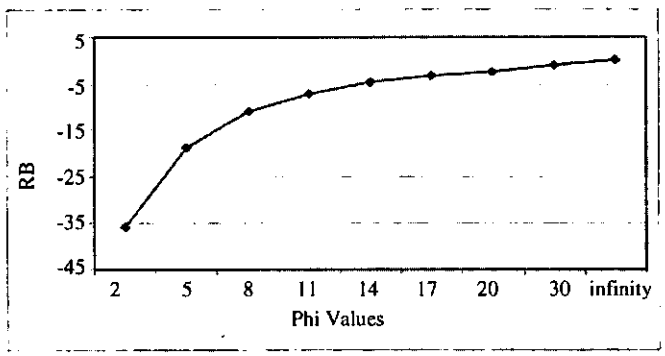
Figure 6.2. RB, RRMSE and MARE of the LS estimator

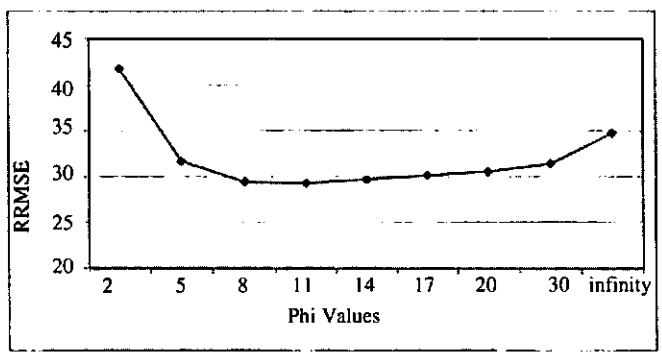
6.3 The *M*-estimator: Outlier robustness

We have compared two versions of the *M*-estimator; one that reduces the influence of large weighted population residuals ($h_k = \tilde{w}_k$) and another one that reduces the influence of large unweighted population residuals ($h_k = 1$). For the weighted version, we chose 7 values of φ ($\varphi = 10, 25, 50, 100, 150, 200, \infty$) and for the unweighted version, we chose 9 values of φ ($\varphi = 2, 5, 8, 11, 14, 17, 20, 30, \infty$). We have only considered the case $\alpha = 1$, as we did not want to confound the effects of changing the constant α with the effect of changing the constant φ. Of course, a more efficient estimator could be found by an appropriate choice of both constants. It is to be noted that the results are based on a single iteration of the IRLS algorithm using $\mathbf{B}^{(0)} = \hat{\mathbf{B}}^{G}$ as the vector of starting values.

It can be seen from figures 6.3 and 6.4 that the weighted version ($h_k = \tilde{w}_k$) has a better potential for reducing the RRMSE and the MARE of *M*-estimators than the unweighted version ($h_k = 1$). Both graphs of RRMSE present a *U*-shaped curve. The RRMSE curve for $h_k = \tilde{w}_k$ shows that a value of φ between 50 and 150 leads to an RRMSE between 25% and about 27%, while the RRMSE of the GREG estimator (last point on the graphs) is equal to 34.77%. The RRMSE curve for $h_k = 1$ shows that the RRMSE is around 30% for values of φ between 8 and 20. In the area where the RRMSE is close to its minimum value, the MARE is smaller when $h_k = \tilde{w}_k$. This suggests that $h_k = \tilde{w}_k$ may control influential units better than $h_k = 1$. As expected, the RB in both figures decreases as φ increases.

We have also considered the weighted and unweighted versions of the *M*-estimator obtained by choosing adaptively, for each selected sample, the value of φ that leads to the smallest estimated MSE (using equation 5.3) among the sets of values of φ considered above. The average value of φ over the selected samples is 72.34 for the weighted version and 10.58 for the unweighted version. Calculation of these averages excludes samples for which $\varphi = \infty$ (13 samples for $h_k = \tilde{w}_k$ and 1 sample for $h_k = 1$). Both averages are close to the optimal values of φ found in figures 6.3 and 6.4 (100 for $h_k = \tilde{w}_k$, and 11 for $h_k = 1$). The weighted version of the *M*-estimator has an RB of −10.24%, RRMSE of 28.07% and MARE of 197.86%. The unweighted version of the *M*-estimator has an RB of −8.26%, RRMSE of 28.18% and MARE of 232.57%. Therefore, both versions of the *M*-estimator lead to a significant improvement over the GREG estimator in terms of RRMSE and MARE at the expense of an increase in RB (around −10%). The MARE is smaller for the weighted version, which again indicates that it controls influential units better than the unweighted version. However, the difference in the RRMSE between these two estimators is very small. Curiously, it seems that there is no increase in MSE due to estimating φ instead of using the optimal value when the unweighted version is used. This observation is somewhat difficult to explain.





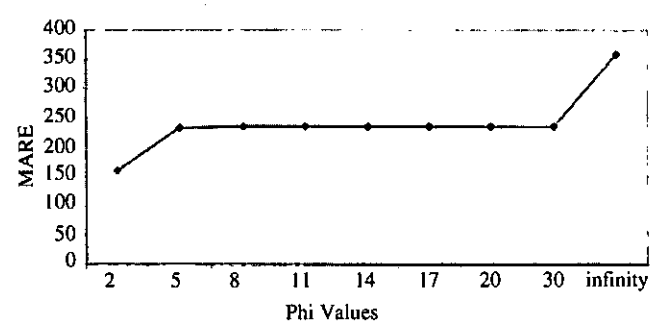
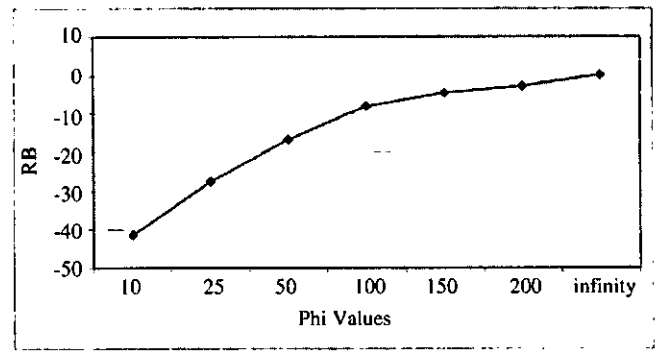
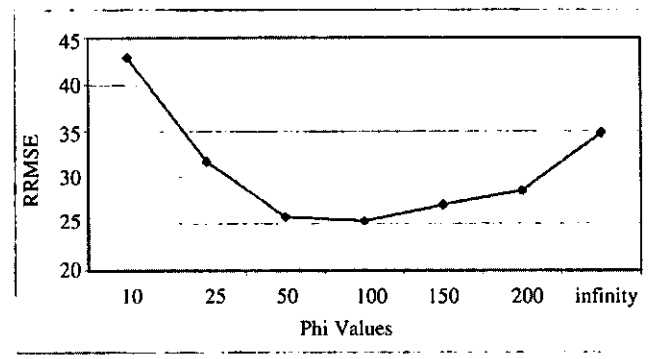


Figure 6.3. RB, RRMSE and MARE of the *M*-estimator when $h_k = 1$





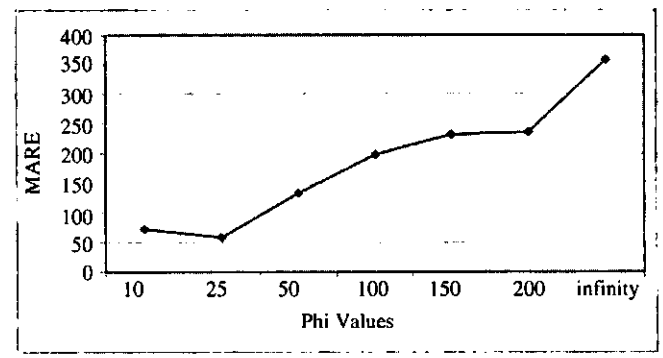


Figure 6.4. RB, RRMSE and MARE of the *M*-estimator when $h_k = \tilde{w}_k$

6.4 Comparison of the Singly-iterated and Fully-iterated *M*-estimators

We now compare the singly- and fully-iterated *M*-estimators when $\alpha = 1$. We only consider the following two cases: i) $h_k = 1$ and $\varphi = 11$; and ii) $h_k = \tilde{w}_k$ and $\varphi = 100$. Most of the time, the IRLS algorithm converged quickly in the fully-iterated case (average number of iterations for convergence is 7.53 for $h_k = 1$, and 7.29 for $h_k = \tilde{w}_k$), but in some of the 5,000 samples (64 for $h_k = 1$, and 75 for $h_k = \tilde{w}_k$) it did not converge. When this situation occurred, we kept the *M*-estimate from the last iteration of the IRLS algorithm. From table 6.2, it is evident that the RB, RRMSE and MARE of the singly- and fully-iterated *M*-estimators are very close to each other. A point worth noting is the slightly smaller RBs for singly-iterated *M*-estimators. This point has also been observed by Lee (1991) and is likely due to the fact that we used $\mathbf{B}^{(0)} = \hat{\mathbf{B}}^{G}$ as the vector of starting values for the IRLS algorithm, which is ADU for $\mathbf{B}$.

Table 6.2
Comparison of Singly- and Fully-iterated *M*-estimators

Estimator	Singly-iterated			Fully-iterated		
	RB	RRMSE	MARE	RB	RRMSE	MARE
M-estimator ($h_k = 1$, $\varphi = 11$)	−6.94%	29.28%	235.07%	−7.93%	29.27%	235.07%
M-estimator($h_k = \tilde{w}_k$, $\varphi = 100$)	−8.14%	25.36%	197.86%	−8.27%	25.33%	196.73%

7. CONCLUSION

In this paper, we considered robust alternatives to the optimal (BLU) estimator. We first proposed a compromise between the GREG and BLU estimators, the LS estimator, to deal with deviations from the ignorability assumption. The LS estimator is obtained by shrinking the design weights toward their mean. It is expected to be more stable than the GREG estimator when the ignorability assumption holds approximately and less biased than the BLU estimator when this assumption is not fully satisfied. This was confirmed in a simulation study using a population created from real survey data. The LS estimator also offers some protection against deviations from model assumptions.

To deal with outliers, we suggested using the weighted generalized *M*-estimation technique to reduce the influence of units with large weighted population residuals. We found in a simulation study that significant gains in MSE could be obtained with this method. We also found that an *M*-estimator obtained using a single iteration of the IRLS algorithm performed similarly to a fully-iterated *M*-estimator. Finally, we proposed implementing *M*-estimators for multipurpose surveys by modifying either the weights of influential units or their values. We believe that both approaches are useful and contribute to bridge a small gap between theory and practice.

ACKNOWLEDGEMENTS

The authors would like to sincerely thank the Associate Editor and three referees for their constructive remarks and suggestions. They would also like to thank Cynthia Bocci and Wesley Yung for their comments, which helped improve the clarity of the paper.

APPENDIX

In this proof, we remove the conditioning on **X** when taking expectations and variances with respect to model *m* in order to simplify the notation. Using Slutsky's theorem, to show that $\mathrm{E}_p(\hat{t}_y^B - t_y)/t_y$ converges in probability to 0, as the sample size n and the population size N tend to infinity, under assumptions (A1), (A2) and (A3), it suffices to show that:

a) $\mathrm{E}_p(t_y/\mathbf{t}_x'\boldsymbol{\beta}) = t_y/\mathbf{t}_x'\boldsymbol{\beta}$ converges in probability to 1 and

b) $\mathrm{E}_p(\hat{t}_y^B/\mathbf{t}_x'\boldsymbol{\beta})$ converges in probability to 1.

To show (a), note that

$$\mathrm{E}_m\left(\frac{t_y}{\mathbf{t}_x'\boldsymbol{\beta}}\right)=1$$

and

$$\mathrm{V}_m\left(\frac{t_y}{\mathbf{t}_x'\boldsymbol{\beta}}\right)=\frac{1}{N}\,\frac{1}{(\mathbf{t}_x'\boldsymbol{\beta}/N)^2}\,\sum_{k\in U}\sigma_k^2\Big/N\,.$$

By Chebychev's inequality, $t_y/\mathbf{t}_x'\boldsymbol{\beta}$ converges in probability to 1 under model m, as N increases, if $\mathbf{t}_x'\boldsymbol{\beta} = O(N)$ and $\sum_{k\in U}\sigma_k^2 = O(N)$ (assumption A3).

To show (b), we first note that $\mathrm{E}_m\mathrm{E}_p(.) = \mathrm{E}_p\mathrm{E}_m(.\,|\,s)$ provided that the set of all possible samples does not depend on which population was generated by model m. Consequently, if assumption (A2) holds, it is straightforward to show that $\mathrm{E}_m\mathrm{E}_p(\hat{t}_y^B/\mathbf{t}_x'\boldsymbol{\beta}) = 1$. Then, we note that

$$\mathrm{V}_{mp}\left(\frac{\hat{t}_y^B}{\mathbf{t}_x'\boldsymbol{\beta}}\right)=\mathrm{V}_m\mathrm{E}_p\left(\frac{\hat{t}_y^B}{\mathbf{t}_x'\boldsymbol{\beta}}\right)+\mathrm{E}_m\mathrm{V}_p\left(\frac{\hat{t}_y^B}{\mathbf{t}_x'\boldsymbol{\beta}}\right). \qquad (A.1)$$

As a result, $\mathrm{V}_m\mathrm{E}_p(\hat{t}_y^B/\mathbf{t}_x'\boldsymbol{\beta}) \le \mathrm{V}_{mp}(\hat{t}_y^B/\mathbf{t}_x'\boldsymbol{\beta})$ since the two terms on the right side of (A.1) are greater than or equal to 0. By the previous inequality and Chebychev's inequality, $\mathrm{E}_p(\hat{t}_y^B/\mathbf{t}_x'\boldsymbol{\beta})$ converges in probability to 1 under model m, as n and N increase, if $\lim_{n,N\to\infty}\mathrm{V}_{mp}(\hat{t}_y^B/\mathbf{t}_x'\boldsymbol{\beta}) = 0$. Using assumption (A2), it is straightforward to show that

$$\mathrm{V}_{mp}\left(\frac{\hat{t}_y^B}{\mathbf{t}_x'\boldsymbol{\beta}}\right)=\frac{1}{N}\,\frac{1}{(\mathbf{t}_x'\boldsymbol{\beta}/N)^2}\,\sum_{k\in U}\mathrm{E}_p\left\{\left(w_k^B\right)^2 I_k\right\}\sigma_k^2\Big/N\,.$$

Consequently, $\lim_{n,N\to\infty}\mathrm{V}_{mp}(\hat{t}_y^B/\mathbf{t}_x'\boldsymbol{\beta}) = 0$ if $\mathbf{t}_x'\boldsymbol{\beta} = O(N)$ and $\sum_{k\in U}\mathrm{E}_p\{(w_k^B)^2 I_k\}\,\sigma_k^2 = O(N)$ (assumption A3). This completes the proof.

REFERENCES

BEATON, A.E., and TUKEY, J.W. (1974). The fitting of power series, meaning polynomials, illustrated on band-spectroscopic data. *Technometrics*, 16, 147-185.

BASU, D. (1971). An essay on the logical foundations of survey sampling, part 1. In *Foundations of statistical inference*, (Eds. V.P. Godambe and D.A. Sprott), Toronto: Holt, Rinehart, and Winston, 203-233.

BINDER, D.A. (1983). On the variances of asymptotically normal estimators from complex surveys. *International Statistical Review*, 51, 279-292.

CHAMBERS, R.L. (1986). Outlier robust finite population estimation. *Journal of the American Statistical Association*, 81, 1063-1069.

CHAMBERS, R.L., DORFMAN, A.H. and WEHRLY, T.E. (1993). Bias robust estimation in finite populations using nonparametric calibration. *Journal of the American Statistical Association*, 88, 268-277.

DEVILLE, J.-C., and SÄRNDAL, C.-E. (1992). Calibration estimators in survey sampling. *Journal of the American Statistical Association*, 87, 376-382.

DRAPER, N., and SMITH, H. (1980). *Applied regression analysis, second edition*. New-York, John Wiley & Sons, Inc.

DUCHESNE, P. (1999). Robust calibration estimators. *Survey Methodology*, 25, 43-56.

DUMOUCHEL, W.H., and DUNCAN, G.J. (1983). Using sample survey weights in multiple regression analyses of stratified samples. *Journal of the American Statistical Association*, 78, 535-543.

ELLIOTT, M.R., and LITTLE, R.J.A. (2000). Model-based alternatives to trimming survey weights. *Journal of Official Statistics*, 16, 191-209.

GRAUBARD, B.I., and KORN, E.L. (1993). Hypothesis testing with complex survey data: the use of classical quadratic test statistics with particular reference to regression problems. *Journal of the American Statistical Association*, 88, 629-641.

GWET, J.-P., and LEE, H. (2000). An evaluation of outlier-resistant procedures in establishment surveys. In *The Second International Conference on Establishment Surveys*, American Statistical Association, Alexandria, Virginia, 707-716.

GWET, J.-P., and RIVEST, L.-P. (1992). Outlier resistant alternatives to the ratio estimator. *Journal of the American Statistical Association*, 87, 1174-1182.

HAMPEL, F.R., RONCHETTI, E.M., ROUSSEEUW, P.J. and STAHEL, W.A. (1986). *Robust Statistics: the Approach Based on Influence Functions*. New-York, John Wiley & Sons, Inc.

HEDLIN, D., FALVEY, H., CHAMBERS, R. and KOKIC, P. (2001). Does the model matter for GREG estimation? A business survey example. *Journal of Official Statistics*, 17, 527-544.

HUBER, P.J. (1964). Robust estimation of a location parameter. *Annals of Mathematical Statistics*, 35, 73-101.

HUBER, P.J. (1981). *Robust Statistics*. New-York, John Wiley & Sons, Inc.

HULLIGER, B. (1995). Outlier robust Horvitz-Thompson estimators. *Survey Methodology*, 21, 79-87.

HULLIGER, B. (1999). Simple and robust estimators for sampling. In *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 54-63.

KALTON, G., and FLORES-CERVANTES, I. (2003). Weighting methods. *Journal of Official Statistics*, 19, 81-97.

KISH, L. (1992). Weighting for unequal P_i. *Journal of Official Statistics*, 8, 183-200.

KORN, E.L., and GRAUBARD, B.I. (1999). *Analysis of Health Surveys*. New-York, John Wiley & Sons, Inc.

LEE, H. (1991). Model-based estimators that are robust to outliers. In *Proceedings of the Annual Research Conference*, Washington, DC, U.S. Bureau of the Census, 178-202.

LEE, H. (1995). Outliers in business surveys. In *Business Survey Methods*, (Eds. B.G. Cox, D.A. Binder, B.N. Chinnappa, A. Christianson, M.J. Colledge and P.S. Kott). Chapter 26, New-York, John Wiley & Sons, Inc.

LITTLE, R.J.A. (1983). Estimating a finite population mean from unequal probability sampling. *Journal of the American Statistical Association*, 78, 596-604.

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

POTTER, F. (1988). Survey of procedures to control extreme sampling weights. In *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 453-458.

POTTER, F. (1990). A study of procedures to identify and trim extreme sampling weights. In *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 225-230.

POTTER, F. (1993). The effect of weight trimming on nonlinear survey estimates. In *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 758-763.

RAO, J.N.K. (1966). Alternative estimators in PPS sampling for multiple characteristics. *Sankhyā*, Series A, 28, 47-60.

RAO, J.N.K., WU, C.F.J. and YUE, K. (1992). Some recent work on resampling methods for complex surveys. *Survey Methodology*, 18, 209-217.

REN, R., and CHAMBERS, R.L. (2002). Outlier robust imputation of survey data via reverse calibration. Southampton Statistical Sciences Research Institute Methodology Working Paper M03/19, University of Southampton.

ROYALL, R.M. (1976). The linear least-squares prediction approach to two-stage sampling. *Journal of the American Statistical Association*, 71, 657-664.

ROYALL, R.M., and HERSON, J. (1973). Robust estimation in finite populations I. *Journal of the American Statistical Association*, 68, 880-889.

RUBIN, D.B. (1976). Inference and missing data. *Biometrika*, 63, 581-592.

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

STOKES, L. (1990). A comparison of truncation and shrinking of sampling weights. In *Proceedings of the* 1990 *Annual Research Conference*, Washington, DC: Bureau of the Census, 463-471.

SUGDEN, R.A., and SMITH, T.M.F. (1984). Ignorable and informative designs in survey sampling inference. *Biometrika*, 71, 495-506.

VALLIANT, R., DORFMAN, A. and ROYALL, R.M. (2000). *Finite population sampling: a prediction approach.* New-York, John Wiley & Sons, Inc.

WELSH, A.H., and RONCHETTI, E. (1998). Bias-calibrated estimation from sample surveys containing outliers. *Journal of the Royal Statistical Society*, Series B, 60, 413-428.

ZASLAVSKY, A.M., SCHENKER, N. and BELIN, T.R. (2001). Downweighting influential clusters in surveys: application to the 1990 post enumeration survey. *Journal of the American Statistical Association*, 96, 858-869.

Survey Methodology, December 2004
Vol. 30, No. 2, pp. 209-218
Statistics Canada

Penalized Spline Nonparametric Mixed Models for Inference about a Finite Population Mean from Two-Stage Samples

HUI ZHENG and RODERICK J.A. LITTLE [1]

ABSTRACT

Samplers often distrust model-based approaches to survey inference because of concerns about misspecification when models are applied to large samples from complex populations. We suggest that the model-based paradigm can work very successfully in survey settings, provided models are chosen that take into account the sample design and avoid strong parametric assumptions. The Horvitz-Thompson (HT) estimator is a simple design-unbiased estimator of the finite population total. From a modeling perspective, the HT estimator performs well when the ratios of the outcome values and the inclusion probabilities are exchangeable. When this assumption is not met, the HT estimator can be very inefficient. In Zheng and Little (2003, 2004) we used penalized splines (p-splines) to model smoothly – varying relationships between the outcome and the inclusion probabilities in one-stage probability proportional to size (PPS) samples. We showed that p-spline model-based estimators are in general more efficient than the HT estimator, and can provide narrower confidence intervals with close to nominal confidence coverage. In this article, we extend this approach to two-stage sampling designs. We use a p-spline based mixed model that fits a nonparametric relationship between the primary sampling unit (PSU) means and a measure of PSU size, and incorporates random effects to model clustering. For variance estimation we consider the empirical Bayes model-based variance, the jackknife and balanced repeated replication (BRR) methods. Simulation studies on simulated data and samples drawn from public use microdata in the 1990 census demonstrate gains for the model-based p-spline estimator over the HT estimator and linear model-assisted estimators. Simulations also show the variance estimation methods yield confidence intervals with satisfactory confidence coverage. Interestingly, these gains can be seen for a common equal-probability design, where the first stage selection is PPS and the second stage selection probabilities are proportional to the inverse of the first stage inclusion probabilities, and the HT estimator leads to the unweighted mean. In situations that most favor the HT estimator, the model-based estimators have comparable efficiency.

KEY WORDS: Weighting; REML; Empirical Bayes estimation.

1. INTRODUCTION

In a sample survey, let y_i denote the value of an outcome Y for unit i, and let S denote the set of sampled units. The Horvitz-Thompson (HT) estimator (Horvitz and Thompson 1952) $\hat{Y}_{HT} = \sum_{i \in S} y_i / \pi_i$, where π_i is the probability of selection of unit i, is a design-unbiased estimator of the finite population total (and of the mean when divided by the known population count N). It can also be regarded as a model-based projective estimator (Firth and Bennett 1998) for the following linear model relating y_i to π_i:

$$y_i = \beta \pi_i + \pi_i \varepsilon_i,$$

where ε_i is assumed to be i.i.d. normally distributed with mean zero and variance σ^2.

In Zheng and Little (2003, 2004), we proposed a nonparametric model

$$y_i = f(\pi_i) + \varepsilon_i, \varepsilon_i \sim \text{ind } N\left(0, \pi_i^{2k} \sigma^2\right),$$

using penalized splines to model mean of outcome y_i as a smoothly-varying function f of the inclusion probabilities π_i. We showed in Zheng and Little (2003) that the nonparametric model-based estimators are more efficient than HT for general one-stage probability-proportional-to-size (PPS) samples and not much less efficient than HT when the data are generated using a model that favors HT.

In this article we consider two-stage sampling. In the first stage, a subset of m primary sampling units (PSUs) is drawn from a population with H PSUs with unequal probabilities $\pi_{1,h}$, $h = 1, ..., H$. Let us number the included PSUs from 1 to m. In the second stage, a simple random sample (srs) of n_h out of N_h secondary sampling units (SSUs) is drawn from the sampled PSU labeled h with probability $\pi_{2,h}$. The overall selection probability for unit i in PSU h is $\pi_h = \pi_{1,h} \pi_{2,h}$, and the HT estimator of the mean of an outcome Y is $\bar{y}_w = \sum_{h=1}^{m} \sum_{i=1}^{n_h} y_{hi} / (\pi_{1,h} \pi_{2,h}) / N$, where y_{hi} is the value of Y for unit i in PSU h and N is the known total number of units (SSUs) in the whole population. In a commonly adopted design, the first stage selection probability is proportional to an estimate of the PSU size, and the second stage inclusion probabilities are proportional to the inverse of the first stage inclusion probabilities so that the overall inclusion probabilities π_h are equal for all SSUs.

[1] Hui Zheng, Department of Health Care Policy, Harvard Medical School, 180 Longwood Avenue, Boston, MA 02115. E-mail: zheng@hcp.med.harvard.edu; Roderick J.A. Little, Department of Biostatistics, University of Michigan, 1420 Washington Heights, Ann Arbor, MI 48109. E-mail: rlittle@umich.edu.

The inverse probability weighted mean in this case equals the simple sample mean $\bar{y} = \sum_{h=1}^{m}\sum_{i=1}^{n_h} y_{hi} / \sum_{h=1}^{m} n_h$.

We assume throughout this article that the selection probabilities $\pi_{1,h}$ are known for all the PSUs $h = 1, \ldots, H$. In sections 2 and 3, we assume the PSU counts N_h are also known for all the PSUs in the population. In section 4, we discuss the common situation where N_h is only known for sampled PSUs, but the N_h for nonsampled PSUs can be estimated using a regression model based on auxiliary variables known for all PSUs in the population.

Särndal, Swensson and Wretman (1992) discussed model-assisted alternatives to the HT estimator for two-stage samples with auxiliary information available at the PSU or SSU level. In the first case, let x_h denote a vector of PSU-level auxiliary variables for PSU h. The PSU totals $t_h = \sum_{i=1}^{N_h} y_{hi}$ are assumed to be related to x_h according to a linear model:

$$E(t_h \mid x_h) = x_h^T\beta,\ \mathrm{Var}(t_h) = \sigma_h^2,\ h = 1, \ldots, H$$

(Särndal *et al.* 1992). β is estimated by the probability-weighted regression

$$\hat{B} = \left(\sum_{h=1}^{m} x_h x_h^T / (\sigma_h^2 \pi_{1,h})\right)^{-1} \sum_{h=1}^{m} x_h t_h^* / (\sigma_h^2 \pi_{1,h}),$$

where $t_h^* = \sum_{i=1}^{n_h} y_{hi} / \pi_{2,h}$, leading to the projected totals $\hat{t}_h = x_h^T \hat{B}$, $h = 1, \ldots H$. In practice, estimates $\hat{\sigma}_h^2$, either simply assumed (*e.g.*, σ_h proportional to a measure of size of stratum h) or estimated, replace σ_h^2 in the above formula. The generalized regression (GR) estimator of the grand total is

$$\hat{T}_A = \sum_{i=1}^{H} \hat{t}_h + \sum_{h=1}^{m} \frac{(t_h^* - \hat{t}_h)}{\pi_{1,h}},$$

and the estimate for the mean is $\hat{T}_A / N$. The term $\sum_{h=1}^{m} (t_h^* - \hat{t}_h) / \pi_{1,h}$ is a bias calibration term that makes the estimator design- consistent.

In the second case where auxiliary information is known at the SSU level, let x_{hi} denote the set of auxiliary variables for SSU i in PSU h, $h = 1, \ldots, H; i = 1, \ldots, N_h$. The relationship between the outcome and the auxiliary information is modeled by

$$E(y_{hi} \mid x_{hi}) = x_{hi}^T \beta, \ldots, \mathrm{Var}(y_{hi}) = \sigma_{hi}^2,\ h = 1, \ldots, H,\ i = 1, \ldots, N_h.$$

The probability weighted regression estimate for β is

$$\hat{B} = \left(\sum_{h=1}^{m}\sum_{i=1}^{n_h} x_{hi} x_{hi}^T / (\sigma_{hi}^2 \pi_{hi})\right)^{-1} \sum_{h=1}^{m}\sum_{i=1}^{n_h} x_{hi}\, y_{hi} / (\sigma_{hi}^2 \pi_{hi}),$$

where π_{hi} is the probability for unit (h, i) to be included in the sample. The GR estimator for the grand total is

$$\hat{T}_B = \sum_{h=1}^{H}\sum_{i=1}^{N_h} \hat{y}_{hi} + \sum_{h=1}^{m}\sum_{i=1}^{n_h} \frac{(y_{hi} - \hat{y}_{hi})}{\pi_{hi}},$$

where $\hat{y}_{hi} = x_{hi}^T \hat{B}$. The estimator for the mean is $\hat{T}_B / N$.

These two methods do not account for the within-PSU correlations of outcome. These correlations can be modeled by treating PSU means as random effects in a hierarchical model. For the case where PSU-level information x_h is available for all PSUs, one such model is:

$$y_{hi} \mid \mu_h \overset{\text{ind}}{\sim} N(\mu_h,\ \sigma^2)$$
$$\mu \sim N_H(\varphi,\ D) \qquad (1)$$

where $\mu = (\mu_1, \ldots, \mu_H)^T$, $\varphi = (\varphi_1, \ldots, \varphi_H)^T$ where μ_h is the mean outcome in PSU h, $\varphi_h = x_h^T\beta$, and D is the covariance matrix of the PSU means. The model-based estimator of $\bar{Y}$ is given by

$$\hat{E}(\bar{Y} \mid \mathbf{y}, x_h) = \frac{1}{N}\left(\sum_{h=1}^{m} [n_h \bar{y}_h + (N_h - n_h)\,\hat{\mu}_h] + \sum_{h=m+1}^{H} N_h \hat{\mu}_h\right),$$

where $\hat{\mu}_h = \hat{E}(y_{hi} \mid \mathbf{y}, x_h)$, and $\mathbf{y}$ is the vector of outcomes in the sample.

Different assumptions about φ and D in (1) lead to the following models:

Exchangeable random effects (XRE): (Holt and Smith 1979; Ghosh and Meeden 1986; Little 1991; Lazzaroni and Little 1998): $\varphi_h \equiv \mu_o, h = 1, \ldots, H$ and $D = \tau^2 I_H$;

Autoregressive (AR1): (Lazzaroni and Little 1998): $\varphi_h \equiv \mu_o, h = 1, \ldots, H$ and $D = r^2\{\rho^{|i-j|}\}$;

Linear (LIN): (Lazzaroni and Little 1998): $\varphi_h = \alpha + \beta x_h, h = 1, \ldots, H$ and $D = \tau^2 I_H$;

Nonparametric: (Elliott and Little 2000): $\varphi_h = f(x_h), h = 1, \ldots, H$ and $D = 0$.

The nonparametric models in Elliott and Little (2000) assume nonparametric mean function relating the outcome to the design variables. By assuming $D = 0$, the PSU means are modeled to equal the mean function f instead of varying around it. Nonparametric mixed models relax the assumptions on D (*e.g.*, $D = \tau^2 I_H$) and serve as a natural extension of the Elliott and Little (2000) model and linear mixed models with a parametric mean structure.

It is worth pointing out that some estimators in the above family of models correspond to standard design-based estimators. For example, in an equal-probability design where n_h are approximately constant across PSUs, the unweighted mean corresponds to the special model-based estimator that assumes φ_h is constant.

2. ESTIMATION FOR THE *P*-SPLINE MIXED MODEL

The linear structure of φ in LIN model is subject to misspecification when the actual mean structure is non-linear. The non-linearity problem can be partially solved by adding polynomial terms (*e.g.*, quadratic or cubic terms) to the fixed effects part in the LIN model. *P*-spline nonparametric mixed models (Lin and Zhang 1999; Brumback, Ruppert and Wand 1999; Coull, Schwartz and Wand 2001) are even more flexible, since they replace polynomials by smooth nonparametric functions. We propose the following *p*-spline nonparametric mixed model for inference about the population mean:

***P*-spline nonparametric mixed model** (PMM):

$$\varphi_h = f(x_h),\ h = 1, ..., H,\ D = \tau^2 I_H,$$

where f is a nonparametric degree p spline function:

$$f(x;\beta) = \beta_0 + \sum_{j=1}^{p} \beta_j x^j + \sum_{l=1}^{K} \beta_{l+p} (x - \kappa_l)_+^p,$$

where $\kappa_1 < ... < \kappa_K$ are K fixed knots, $\beta_0, ..., \beta_{p+K}$ are coefficients to be estimated and $(x)_+^p = x^p \mathbf{I}(x \geq 0)$.

A naive way of estimating $\beta_0, ..., \beta_{p+K}$ is to treat them as fixed and estimate them together with the variance components σ^2 and τ^2 by fitting a linear mixed model. However this method can yield estimates of f with too much roughness and variability. To avoid overfitting, the roughness of the estimation $\hat{f}$ can penalized by adding a penalty term to the sum of squared deviations, so that the solution $\hat{\beta}_0, ..., \hat{\beta}_p$ is minimizes

$$\sum_{h=1}^{m} (\hat{f}(x_h) - \hat{\mu}_h)^2 + \alpha \sum_{l=1}^{K} \beta_{l+p}^2 .$$

This is achieved in the context of the model by assigning $\beta_0, ..., \beta_p$ flat priors, $(\beta_{p+1}, ..., \beta_{p+K})$ a normal prior $N_m(0, \sigma_\beta^2)$, and letting $\alpha = \tau^2 / \sigma_\beta^2$. The result is a penalized spline (p-spline) model.

When $p = 1$, $\hat{f}$ is piecewise linear and the coefficients $\beta_0, ..., \beta_{K+1}$ and σ^2, σ_β^2 and τ^2 are estimated by fitting the linear mixed model:

$$y = X_1 \beta + X_2 u + \varepsilon, \qquad (2)$$

where $y = (y_{11}, y_{12}, ..., y_{mn_m})^T$, $\beta = (\beta_0, \beta_1)^T$, $u = (\beta_2, ..., \beta_{K+1}, u_1, ..., u_m)^T$,

$$X_1 = \begin{bmatrix} 1 & x_1 \\ 1 & x_1 \\ . & . \\ . & x_1 \\ . & x_2 \\ . & . \\ . & . \\ . & x_2 \\ . & . \\ . & . \\ . & . \\ 1 & x_m \end{bmatrix},$$

$$X_2 = \begin{bmatrix} (x_1 - \kappa_1)_+ & ... & (x_1 - \kappa_K)_+ & 1 & 0 & . & 0 \\ . & ... & . & . & . & . & . \\ . & ... & . & . & . & . & . \\ (x_1 - \kappa_1)_+ & ... & (x_1 - \kappa_K)_+ & 1 & 0 & . & . \\ (x_2 - \kappa_1)_+ & ... & (x_2 - \kappa_K)_+ & 0 & 1 & . & . \\ . & ... & . & . & . & . & . \\ . & ... & . & . & . & . & . \\ (x_2 - \kappa_1)_+ & ... & (x_2 - \kappa_K)_+ & 0 & 1 & . & 0 \\ . & ... & . & 0 & 0 & . & 1 \\ . & ... & . & . & . & . & . \\ . & ... & . & . & . & . & . \\ (x_m - \kappa_1)_+ & ... & (x_m - \kappa_K)_+ & 0 & 0 & . & 1 \end{bmatrix},$$

where x_h in X_1 and $(x_h - \kappa_l)_+$ in X_2 are both repeated n_h times. The random terms u and ε are mutually independent with

$$u = (\beta_2, ..., \beta_{K+1}, u_1, ..., u_m)^T \sim N_{K+m}(0, G),$$

$$G = \begin{bmatrix} \sigma_\beta^2 I_K & 0 \\ 0 & \tau^2 I_m \end{bmatrix}.$$

Variance components σ^2, σ_β^2 and τ^2 can be estimated by fitting model (2) by restricted maximum likelihood (REML).

The predicted means of PSUs included in the sample are given by: $\hat{\mu} = X_1 \hat{\beta} + X_2 \hat{u}$, where $\hat{\beta} = (X_1^T \hat{V}^{-1} X_1)^{-1} X_1^T \hat{V}^{-1} \bar{y}$, $\hat{u} = \hat{G} X_2^T \hat{V}^{-1} (\bar{y} - X_1 \hat{\beta})$, where $\hat{V} = X_2 \hat{G} X_2^T + \hat{\sigma}^2 \Sigma$, $\Sigma = \text{diag}[\{1/n_h\}_{h=1}^m]$ and $\bar{y} = (\bar{y}_1, ..., \bar{y}_m)^T$. The predicted mean for a PSU h that is not selected in the first stage is $\hat{\mu}_h = x_h^T \hat{\beta}^*$, where

$$x_h = [1 \; x_h \; (x_h - \kappa_1)_+ \ldots (x_h - \kappa_K)_+]^T$$

and

$$\hat{\beta}^* = [\hat{\beta}_0 \; \hat{\beta}_1, \; \ldots, \; \hat{\beta}_{K+1}]^T.$$

Combining the predictions, we obtain the model-based estimator of the population mean

$$\hat{E}(\bar{Y} \mid y, x_h) = \frac{1}{N}\left(\sum_{h=1}^{m}[n_h \bar{y}_h + (N_h - n_h)\,\hat{\mu}_h] + \sum_{h=m+1}^{H} N_h \hat{\mu}_h\right).$$

3. VARIANCE ESTIMATION METHODS

3.1 Empirical Bayes Model-based Variance

Model (2) can be interpreted as a Bayes model in which the parameters $u = (\beta_2, \ldots, \beta_{K+1}, u_1, \ldots, u_m)^T$ have multivariate normal prior $N_{K+m}(0, G)$, and β_0, β_1, σ^2, σ_β^2 and τ^2 all have the flat priors. This leads to the Bayes posterior variance for the vector $(\beta_0, \beta_1, \ldots, \beta_{K+1}, u_1, \ldots, u_m)^T$ conditional on σ^2, σ_β^2 and τ^2 as

$$\mathrm{Var}((\beta_0, \beta_1, \ldots, \beta_{K+1}, u_1, \ldots, u_m)^T \mid \sigma^2, \sigma_\beta^2, \tau^2, y) = \sigma^2 (X^T X + \Delta)^{-1}$$

where $X = [X_1 \; X_2]$ and

$$\Delta = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & \sigma^2/\sigma_\beta^2 \, I_K & 0 \\ 0 & 0 & 0 & \sigma^2/\tau^2 \, I_m \end{bmatrix},$$

where I_K and I_m are $(K \times K)$ and $(m \times m)$ identity matrices, respectively.

The empirical Bayes posterior variance for $(\beta_0, \beta_1, \ldots, \beta_{K+1}, u_1, \ldots, u_m)^T$ is calculated by replacing σ^2, σ_β^2 and τ^2 with their maximum likelihood (ML) or restricted maximum likelihood (REML) estimates $\hat{\sigma}^2, \hat{\sigma}_\beta^2$ and $\hat{\tau}^2$, respectively. The empirical Bayes method underestimates the true posterior variance, but the underestimation is not severe for the sample sizes encountered in many survey settings. A fully Bayes solution is also possible, but is not covered here.

The predicted population mean is $\hat{T}_{\text{pred}}/N$, where $\hat{T}_{\text{pred}} = T_1 + \hat{T}_2$, $T_1 = \sum_{h=1}^{m} n_h \bar{y}_h$ is the sample total, and $\hat{T}_2$ is the estimated total for units not included in the sample,

$$\hat{T}_2 = \sum_{h=1}^{m}(N_h - n_h)\,\hat{\mu}_h + \sum_{h=m+1}^{H} N_h \hat{\mu}_h = N_P X_P \left[\hat{\beta}_0 \; \hat{\beta}_1 \ldots \hat{\beta}_{K+1} \; \hat{\mu}_1 \ldots \hat{\mu}_m\right]^T, \quad (3)$$

where

$$N_P = [(N_1 - n_1)\ldots(N_m - n_m)\, N_{m+1} \ldots N_H],$$

and

$$X_P = \begin{bmatrix} 1 & x_1 & (x_1 - \kappa_1)_+ & \cdots & (x_1 - \kappa_K)_+ & 1 & 0 & \cdots & 0 \\ . & . & . & \cdots & . & 0 & 1 & 0 & 0 \\ . & . & . & \cdots & . & . & . & 0 & . \\ . & . & . & \cdots & . & 0 & 0 & 1 & 0 \\ 1 & x_m & (x_m - \kappa_1)_+ & \cdots & (x_m - \kappa_K)_+ & 0 & \cdots & 0 & 1 \\ 1 & x_{m+1} & (x_{m+1} - \kappa_1)_+ & \cdots & (x_{m+1} - \kappa_K)_+ & 0 & \cdots & \cdots & 0 \\ . & . & & \cdots & . & . & \cdots & \cdots & . \\ . & . & & \cdots & . & . & \cdots & \cdots & . \\ . & . & & \cdots & . & . & \cdots & \cdots & . \\ 1 & x_H & (x_H - \kappa_1)_+ & \cdots & (x_H - \kappa_K)_+ & 0 & \cdots & \cdots & 0 \end{bmatrix}.$$

The empirical Bayes posterior variance for $\hat{\bar{Y}} = \hat{T}_{\text{pred}}/N$ is

$$\mathrm{Var}(\hat{\bar{Y}} \mid \sigma^2, \sigma_\beta^2, \tau^2, X, X_P) = \sigma^2 (N_P X_P (X^T X + \Delta)^{-1} X_P^T N_P^T)/N^2.$$

3.2 The Jackknife Method

A jackknife variance estimator is developed for the PMM estimator. The jackknife replicates are constructed by dividing the set of PSUs into G equal-sized subgroups and computing the g^{th} pseudovalue as $\hat{\bar{Y}}_g = G\hat{\bar{Y}} - (G-1)\hat{\bar{Y}}_{(g)}$, where $\hat{\bar{Y}}$ is the original PMM estimator and $\hat{\bar{Y}}_{(g)}$ is the same estimator calculated from the reduced sample obtained by excluding the elements from the PSUs in the g^{th} subgroup.

The jackknife variance estimate of $\hat{\bar{Y}}$ is

$$v(\hat{\bar{Y}}) = \frac{1}{G(G-1)} \sum_{g=1}^{G} (\hat{\bar{Y}}_g - \hat{\bar{\bar{Y}}})^2,$$

where $\hat{\bar{\bar{Y}}} = \sum_{g=1}^{G} \hat{\bar{Y}}_g / G$. In order to balance the distribution of the selection probabilities across the subgroups, sampled units are stratified into n/G strata each of size G with similar first stage inclusion probabilities, and the G subgroups are constructed by randomly selecting one element from each stratum. To save computation, estimates $\hat{\sigma}^2, \hat{\sigma}_\beta^2$ and $\hat{\tau}^2$ are not recomputed for each replicate. That is, we compute pseudovalues of $(\beta_0, \beta_1, \ldots, \beta_{K+1}, u_1, \ldots, u_m)^T$ based on the variance components estimated from the whole sample.

Miller (1974) and Shao and Wu (1987, 1989) proved asymptotic properties of the jackknife estimator and jackknife variance estimation in the case of multiple linear

regression. Zheng and Little (2004) provided a theoretical justification for the jackknife method for the p-spline model-based estimator in the case of one-stage designs. Numerical simulations in section 4 suggest the above described jackknife method also works well for the two-stage design. Improved performance might be achieved using the weighted jackknife proposed by Hinkley (1977).

3.3 The Balanced Repeated Replication Method

The BRR method can be applied in stratified designs with two units sampled in each stratum. For designs with one PSU per stratum, strata are often collapsed (Kalton 1977) for BRR variance estimation. In our application we assume the PSUs are sampled systematically from a randomly ordered list. This can be viewed approximately as a stratified design with n strata each consisting of PSUs with cumulative measures of approximate size $\sum_{i=1}^{H} z_i / n$, where z_i are the measures of size for the PSUs . One PSU is sampled from each of the n strata. Assuming n is even, the design can be approximated by a stratified design with $n/2$ strata with measures of size $2\sum_{i=1}^{N} z_i / n$, and two units sampled per stratum. Balanced repeated half samples are constructed by selecting one PSU from each stratum, with the selection scheme based on Hadamard matrices (Plackett and Burman 1946). Let $\hat{\bar{Y}}_b$ be the p-spline estimator computed from the b^{th} half sample, using the same knots as used in the computation using the full sample – the number and placement of knots needs to allow the spline model to be fitted on each half-sample. The BRR estimator is given by $v_{\text{BRR}}(\hat{\bar{Y}}) = 1/B \sum_{b=1}^{B} (\hat{\bar{Y}}_b - \hat{\bar{Y}})^2$. This estimate of the variance is subject to some bias, because it treats the design as if it was stratified with two PSUs per stratum.

4. WHEN SOME PSU COUNTS ARE NOT KNOWN

In sections 2 and 3 we assumed that the PSU counts N_h are known for sampled and non-sampled PSUs. In this section we discuss the situation where N_h is only known exactly for the sampled PSUs (labeled 1 through m). We also assume that values $M_h, h = 1, ..., H$ of an auxiliary variable predictive of N_h are known for the whole population. For example, the M_h may be PSU counts estimated from outside sources such as a census. We conduct a regression of N_h on M_h using the sampled PSUs and replace the counts N_h in (3) for nonsampled PSUs with predictions $\hat{N}_h, h = m+1, ..., H$ from this regression. The resulting estimate of the total is

$$\tilde{T} = T_1 + \sum_{h=1}^{m} (N_h - n_h)\hat{\mu}_h + \sum_{h=m+1}^{H} \hat{N}_h \hat{\mu}_h .$$

The variance estimate of $\tilde{T}$ needs to incorporate the additional variability in $\hat{N}_h$. In particular, a model-based variance for $\tilde{T}$ is

$$\begin{aligned}\text{Var}(\tilde{T} \mid \pi_h, M_h) = {} & \text{Var}(E(\tilde{T} \mid \hat{N}_h, \pi_h, M_h)) \\ & + E(\text{Var}(\tilde{T} \mid \hat{N}_h, \pi_h, M_h)),\end{aligned}$$

where

$$E(\tilde{T} \mid \hat{N}_h, \pi_h, M_h) = \sum_{h=1}^{m} (N_h - n_h)\,\mu_h + \sum_{h=m+1}^{H} \hat{N}_h\,\mu_h$$

and

$$\text{Var}(\tilde{T} \mid \hat{N}_h, \pi_h, M_h) \approx \sigma^2 (\tilde{N}_P X_P (X^T X + \Delta)^{-1} X_P^T \tilde{N}_P^T),$$

$\tilde{N}_P = [(N_1 - n_1) ... (N_m - n_m)\, \hat{N}_{m+1} ... \hat{N}_H]$, and X, X_P and Δ are defined as in (3).

If the models for μ_h and N_h are both correctly specified, the above variance can be estimated according to the corresponding models.

5. SIMULATIONS

5.1 Simulation Design

Two simulations are conducted to compare the inverse probability weighting method, the model-assisted method (Särndal *et al.* 1992) and the PMM method in the case of two-stage samples.

In our first simulation, artificial populations are generated with different mean functions $f(\pi_{1,h})$ of the first stage inclusion probabilities. Four different mean functions are simulated: 1) NULL, a constant function; 2) LINDOWN, a linearly decreasing function; 3) EXP, an exponentially increasing function; and 4) SINE, a sine function.

Two combinations of values for variance components are simulated: 1) $\sigma = 0.1$ and $\tau = 0.2$; 2) $\sigma = 0.2$ and $\tau = 0.1$. Only normal errors around the mean functions are simulated while both normal and lognormal within-PSU errors are simulated.

The population consists of 500 PSUs, and in the first stage 48 PSUs are sampled systematically with probability proportional to size (PPS) from a randomly-ordered list. The PSU sizes are uniformly distributed with values ranging from 4 to about 400. The SSU count in each PSU is generated from a distribution with mean equal to 1.05 times the measure of size and log-normal errors with standard deviation 30.

Two types of second-stage sampling plans are studied: 1) within-PSU simple random sampling (srs) with inclusion probabilities proportional to the inverse of the first stage inclusion probabilities, resulting in an equal inclusion probability for all SSUs.; 2) within-PSU simple random sampling with the same sampling rate across sampled PSUs, so that the resulting inclusion probabilities for the SSUs in PSU h are proportional to $\pi_{1,h}$.

For each sample drawn under both sampling plans, the following methods are applied:

A. The HT estimator.
B. The model-assisted estimation method. We use a linear model regressing the outcome y_{hi} on the first stage inclusion probabilities, which are treated as element-level information. The GR estimator is computed by the formula given in section 1.
C. The PMM method, with the first-stage inclusion probabilities $\pi_{1,h}$ as the covariate. We use 20 equal percentiles of $\pi_{1,h}$ of the sampled PSUs as the knots for p-spline regression.
D. The PMM method with the PSU means μ_h estimated the same way as in C, but using estimated PSU counts from a simple linear regression of N_h on the measures of size, which are proportional to $\pi_{1,h}$. This part of the simulation is conducted to study the method described in section 4.

Estimates of $\bar{Y}$ from methods A-D are calculated for each of the 500 samples drawn repeatedly from the artificial populations (each artificial population is generated only once). For the PMM estimator, we compute the empirical Bayes, the jackknife ($K = 8$) and BRR variance estimators for each repeated sample. The mean estimate for the variance of PMM and the coverage rate of the corresponding 95% confidence interval are used to judge the quality of inference. For method D, we study the empirical bias of the model-based variance estimator described in section 4, together with coverage rates of associated confidence intervals.

In the second simulation study, we draw samples of household income data from the 5% public use microdata sample (PUMS) for the State of Michigan in the 1990 US Census, which we treat as a finite population. This simulation is more realistic than the previous simulation in that the outcome values are drawn from a real rather than simulated distribution. The PSUs we simulate are based on the natural geographical clusters called "Public Use Microdata Areas" (PUMAs),which are typically counties and places. There are 67 PUMAs in the Michigan 5% PUMS, with counts of families ranging from around 1,300 to over 10,000. We increase the number of available PSUs by dividing each PUMA into 5, resulting in 335 PSUs. The PSU counts ranges from 134 to 3,058. Figure 1 gives the scatter plot of one sample of the average household income versus sampled PSU sizes together with the regression curve $\hat{f}(x)$.

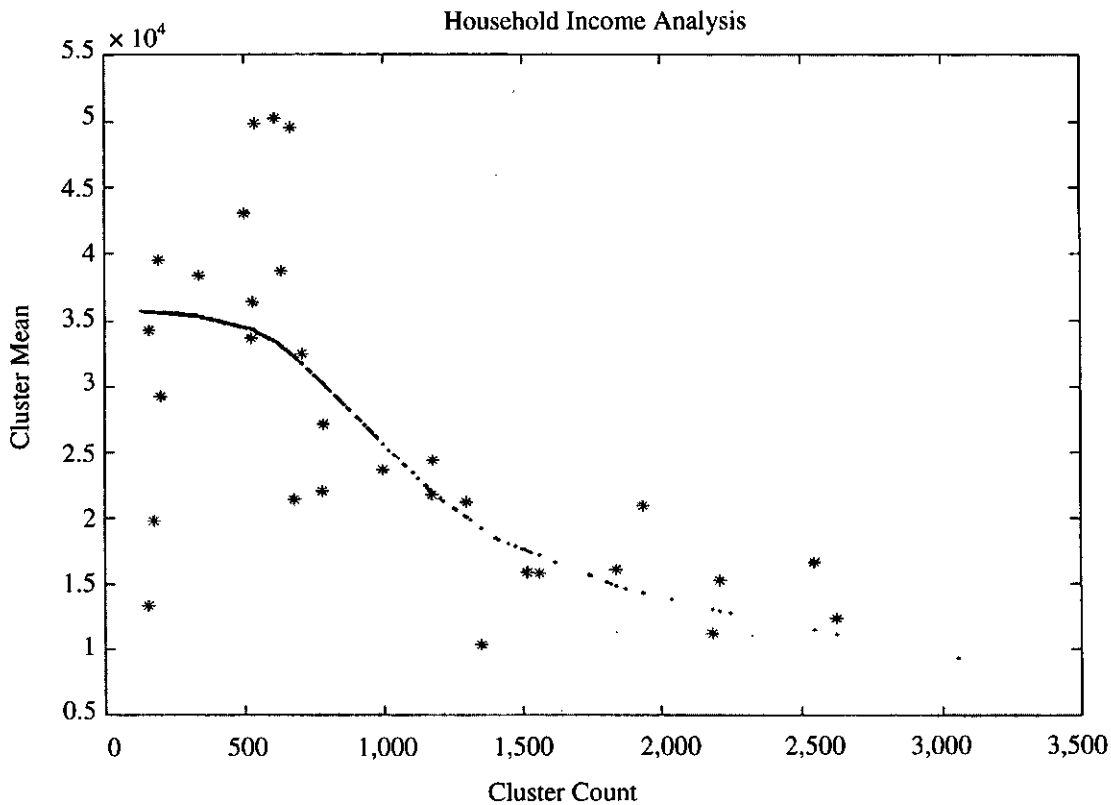


Figure 1. *P*-spline Regression Curve (dotted line) and the Average Household Income (stars) in Sampled PSUs

Five hundred two-stage samples are drawn, each consisting of 30 PSUs and 20 SSUs (families) from each selected PSU. The first stage sampling is systematic PPS where the measures of size are equal to the PSU counts. The second stage sample is simple random sampling with inclusion probabilities proportional to the inverse of the first stage inclusion probabilities. In the estimation of the mean, we use the true PSU counts as variable x_h, with values proportional to the first-stage inclusion probabilities. We apply the p-spline nonparametric mixed model formulated in (2). We use 10 equally spaced sample percentiles of the PSU counts as the knots in the p-spline.

5.2 Results

Table 1 gives the empirical bias and root mean squared error (RMSE) from four estimation methods of the finite population mean applied to equal probability sample from populations generated with both normal and log-normal within-PSU errors and two (σ, τ) combinations. The empirical bias and RMSE are estimated by the mean bias and squared error from the 500 repeated samples.

Table 1 suggests the PMM based methods give estimators with small biases. In the case of equal probability sampling, the PMM estimator is roughly as efficient as HT estimator when the mean function f is constant. In the more general cases such as NULL and LINDOWN, where f is linear but not constant, the linear model-assisted and PMM method are comparable and both are more efficient than the HT estimator in terms of root mean squared error. For populations EXP and SINE, whose mean functions are not linear, the PMM method is superior to both the HT and the linear model-assisted estimators. The improvement of efficiency requires the knowledge of complete design information including probabilities $\pi_{1,h}$ and PSU counts N_h for the whole population. When using estimated PSU counts $\hat{N}_h$ in the place of N_h, the resulting estimator is less efficient than in the case with known N_h, but the PMM estimator can still outperform the HT when the mean function is non-constant. Comparisons on populations with normal or log-normal within-PSU errors result in similar findings.

Similar gains for the PMM method are seen in Table 2, for the case of unequal probability sampling. This suggests that the key to improved efficiency is the better prediction given by the nonparametric models. Tables 1 and 2 both suggest that the p-spline model-based estimators have very small empirical design-biases. We believe this is because the flexible mean functions yield good predictions of the PSU means.

Table 3 compares point estimation and coverage of 95% confidence intervals from three variance estimation methods for PMM: the empirical Bayes model-based method, the Jackknife method and the BRR method. The empirical Bayes method is generally satisfactory but tends to underestimate the true variance of PMM estimator, resulting in under-coverage in some cases. The jackknife and the BRR methods tend to yield more robust estimates for the variance. In general, PMM yields estimates with improved efficiency over the traditional HT and linear model-assisted estimators and satisfactory design-based inferences.

Table 1
Empirical Biases and RMSE of PMM, HT, GR and PMM with Estimated N_h for Samples Under Equal Probability Designs

		PMM		Horvitz-Thompson		Linear Model-Assisted		PMM with Estimated N_h	
	$(\times 10^{-3})$	BIAS	RMSE	BIAS	RMSE	BIAS	RMSE	BIAS	RMSE
Normal	NULL	1.1	29.7	0.8	30.0	0.8	29.9	1.3	30.1
Errors	LINDOWN	3.5	30.7	3.6	36.4	3.7	30.7	2.3	30.4
$\tau = 0.2$	EXP	−4.4	29.1	−9.4	53.0	−9.5	36.7	−4.3	29.1
$\sigma = 0.1$	SINE	4.8	32.5	2.1	42.0	−0.3	35.9	5.2	34.3
Normal	NULL	5.7	22.0	6.6	22.5	6.6	22.1	5.5	22.3
Errors	LINDOWN	0.5	20.4	−0.6	27.1	−0.3	20.5	1.6	20.6
$\tau = 0.1$	EXP	0.9	23.1	1.9	50.3	−4.2	31.7	0.4	23.4
$\sigma = 0.2$	SINE	7.0	22.3	6.5	34.9	3.8	26.4	8.0	26.4
Log-normal	NULL	1.7	32.3	0.9	32.3	0.7	32.3	1.5	32.5
Errors	LINDOWN	2.9	31.9	3.8	39.4	2.7	32.1	3.2	32.0
$\tau = 0.2$	EXP	−0.6	28.4	−5.9	51.5	−6.9	36.4	−0.3	28.5
$\sigma = 0.1$	SINE	6.9	33.8	1.5	43.7	−1.9	39.0	−3.1	35.0
Log-normal	NULL	8.5	30.5	9.6	31.3	9.2	31.0	9.1	30.8
Errors	LINDOWN	3.6	32.3	1.9	37.5	3.6	32.1	6.4	33.1
$\tau = 0.1$	EXP	3.9	29.0	6.8	53.8	1.0	34.4	3.7	29.4
$\sigma = 0.2$	SINE	−2.9	30.1	−8.9	44.7	−12.0	38.4	−3.8	35.9

Table 2
Empirical Biases and RMSE of PMM, HT, GR and PMM with Estimated N_h for Samples Under Unequal Probability Designs

		PMM		Horvitz-Thompson		Linear Model-Assisted		PMM with Estimated N_h	
	($\times 10^{-3}$)	BIAS	RMSE	BIAS	RMSE	BIAS	RMSE	BIAS	RMSE
Normal	NULL	−4.5	29.3	−3.7	33.6	−3.2	30.5	−4.5	29.3
Errors	LINDOWN	−0.9	27.0	3.7	35.5	1.8	27.7	−0.7	26.9
$\tau = 0.2$	EXP	5.8	32.0	1.9	56.8	0.4	39.4	14.1	34.4
$\sigma = 0.1$	SINE	7.1	30.1	6.1	39.5	3.6	32.8	5.3	30.4
Normal	NULL	−7.7	21.3	−7.7	24.9	−6.6	21.1	−7.6	21.2
Errors	LINDOWN	1.1	20.7	3.2	30.6	1.2	20.7	3.5	21.1
$\tau = 0.1$	EXP	−2.3	20.9	−6.5	53.3	−7.2	30.0	−3.0	20.9
$\sigma = 0.2$	SINE	5.6	20.9	6.9	36.2	4.0	28.6	4.3	21.1
Log-normal	NULL	−0.5	28.5	−2.0	30.6	−2.1	29.5	−0.3	28.5
Errors	LINDOWN	5.4	32.6	5.0	39.0	3.7	34.1	6.0	32.7
$\tau = 0.2$	EXP	−1.3	28.6	−7.6	62.6	−7.1	36.8	−9.3	30.3
$\sigma = 0.1$	SINE	3.7	31.2	2.3	43.1	0.1	36.1	1.6	31.0
Log-normal	NULL	3.6	22.8	5.7	28.8	5.7	24.2	3.6	22.7
Errors	LINDOWN	6.0	26.8	9.3	37.5	7.5	27.3	2.5	26.0
$\tau = 0.1$	EXP	0.8	26.3	−2.3	50.8	−3.5	33.1	11.5	29.0
$\sigma = 0.2$	SINE	3.7	26.9	2.9	37.6	−0.1	30.2	2.2	27.8

Table 3
Variance Estimation and Empirical Coverage Rates of 95% C.I. Using the Model-based, Jackknife and BRR Methods

		Empirical variance	Empirical Bayes Model-based		Jackknife($K = 8$)		BRR	
			Estimate		Estimate		Estimate	
	Shape	($\times 10^{-5}$)	($\times 10^{-5}$)	%	($\times 10^{-5}$)	%	($\times 10^{-5}$)	%
Normal	NULL	88	74	92.8	94	96.4	96	94.4
Errors	LINDOWN	94	73	89.6	94	94.6	98	94.2
$\tau = 0.2$	EXP	85	70	91.4	88	94.6	85	93.4
$\sigma = 0.1$	SINE	83	67	91.6	90	95.8	85	94.4
Normal	NULL	48	45	93.8	48	96.0	49	93.8
Errors	LINDOWN	42	45	96.8	51	96.2	51	96.8
$\tau = 0.1$	EXP	53	54	95.0	61	97.2	59	95.2
$\sigma = 0.2$	SINE	44	46	95.8	55	96.6	49	96.0
Log-normal	NULL	104	83	91.8	104	94.8	100	93.6
Errors	LINDOWN	102	98	93.6	106	95.6	107	95.0
$\tau = 0.2$	EXP	81	77	93.4	97	96.4	89	94.8
$\sigma = 0.1$	SINE	92	99	94.8	97	95.2	92	93.4
Log-normal	NULL	93	97	94.2	100	96.2	99	95.2
Errors	LINDOWN	104	101	93.6	106	96.0	102	92.8
$\tau = 0.1$	EXP	84	81	94.6	84	95.2	82	95.0
$\sigma = 0.2$	SINE	110	96	94.4	98	95.6	92	93.0

Tables 4 and 5 give the empirical variance of the PMM estimator when the non-sampled PSU counts N_h are estimated. They also give the mean estimated variance of this estimator and corresponding coverage rates by the 95% C.I. The confidence intervals are calculated by the usual normal theory intervals based on our point and variance estimators. These two tables show the inference method discussed in section 5 tends to underestimate the true variance of PMM estimator using $\hat{N}_h$, giving in occasion under-coverage of the population mean. It remains to be studied in the future whether the JRR and BRR methods also yield satisfactory inferences for this method.

For the simulation study using 5% PUMS data, the simple mean has bias = −50.9 and RMSE = 2,600 and the p-spline nonparametric mixed model based method has bias = −41.9 and RMSE = 2,153.Thus both methods have small bias and the model-based estimator has a RMSE 17% less than the RMSE of the simple mean. This improved efficiency is due to the fact that the average household income decreases for as the number of families in the PSUs increases (figure 1). The PMM method exploits this relationship in its predictions.

Table 4
Variance Estimation and Empirical Coverage Rates of 95% C.I. Using *P*-spline and Estimated PSU Counts, Population Simulated with Normal Errors

	$\sigma=0.1$ and $\tau=0.2$			$\sigma=0.2$ and $\tau=0.1$		
	Empirical Variance ($\times10^{-5}$)	Estimated Variance ($\times10^{-5}$)	Coverage Rate	Empirical Variance ($\times10^{-5}$)	Estimated Variance ($\times10^{-5}$)	Coverage Rate
NULL	90	76	91.8	50	46	93.2
LINDOWN	93	74	90.4	43	46	95.6
EXP	85	72	93.0	55	56	96.2
SINE	110	98	94.8	50	55	97.6

Table 5
Variance Estimation and Empirical Coverage Rates of 95% C.I. Using *P*-spline and Estimated PSU Counts, Population Simulated with Log-normal Errors

	$\sigma=0.1$ and $\tau=0.2$			$\sigma=0.2$ and $\tau=0.1$		
	Empirical Variance ($\times10^{-5}$)	Estimated Variance ($\times10^{-5}$)	Coverage Rate	Empirical Variance ($\times10^{-5}$)	Estimated Variance ($\times10^{-5}$)	Coverage Rate
NULL	105	84	91.8	95	99	94.8
LINDOWN	103	98	94.4	110	102	94.4
EXP	81	79	94.6	87	83	94.2
SINE	110	150	96.4	91	130	95.8

6. DISCUSSION

Previous parametric model-based inference methods have been criticized mainly for their potentially large design biases when the model is misspecified. In our nonparametric models, the linearity assumption is replaced by a much weaker assumption of a smoothly-varying relationship. As a result, the model-based estimators are more robust, having small biases for a variety of population shapes.

Design information such as inclusion probabilities plays a key role in the model-based inference. Inverse-probability weighted methods imply simple assumptions about the relationship between the outcome variables and the design variables. With the method we propose, the gain in efficiency is realized by applying nonparametric models that relax these assumptions.

Our study has an interesting finding that the model-based estimators can be more efficient than the simple mean for an equal probability design. In other studies, we also find gains in efficiency from *p*-spline nonparametric mixed model in estimating post-stratum means in post-stratified samples.

The empirical Bayes method, the jackknife and BRR methods all give good confidence coverage with confidence intervals that are narrower than those given by the traditional methods. However, we expect the empirical Bayes method to be sensitive to model assumptions on the variance components (*e.g.*, constant within-PSU variances). When the PSU counts are not known for the sample but not for the whole population, model-based estimates of the unknown counts can still provide sound estimates of the population mean, if the model tracks the true PSU counts precisely enough. The model relating these counts to the auxiliary variable was treated parametrically here, but this could also be specified nonparametrically without much difficulty.

We believe *p*-spline nonparametric mixed models can be applied to more complex designs such as stratified and multi-stage designs. We also believe without much more effort our methods can be generalized for binary or ordinal outcomes.

ACKNOWLEDGEMENTS

This research was supported by grant DMS 0106914 from the National Science Foundation.

REFERENCES

BRUMBACK, B.A., RUPPERT, D. and WAND, M.P. (1999). Comment to variable selection and function estimation in additive nonparametric regression using data-based prior. *Journal of the American Statistical Association*, 94, 794-797.

COULL, B.A., SCHWARTZ, J. and WAND, M.P. (2001) Respiratory health and air pollution: Additive mixed model analyses. *Biostatistics*, 2(3), 337-349.

ELLIOTT, M.R., and LITTLE, R.J.A. (2000). Model-based alternatives to trimming survey weights. *Journal of Official Statistics*, 16, 191-209.

FIRTH, D., and BENNETT, K.E. (1998). Robust models in probability sampling. *Journal of the Royal Statistical Society*, B, 60, 3-21.

GHOSH, M., and MEEDEN, G. (1986). Empirical Bayes estimation of means from stratified samples. *Journal of the American Statistical Association*, 81, 1058-1062.

HINKLEY, D.V. (1977). Jackknifing in unbalanced situations. *Technometrics*, 19, 285-292.

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

HORVITZ, D.G., and THOMPSON, D.J. (1952). A generalization of sampling without replacement from a finite universe. *Journal of American Statistical Association*, 47, 663-685.

KALTON, G. (1977). Practical methods for estimating survey sampling errors. *Bulletin of the International Statistical Institute*, 47, 495-514.

LAZZARONI, L.C., and LITTLE, R.J.A. (1998). Random effects models for smoothing poststratification weights. *Journal of Official Statistics*, 14, 61-78.

LIN, X., and ZHANG, D. (1999). Inference in generalized additive mixed models using smoothing splines. *Journal of the Royal Statistical Society*, B, 61, 381-400.

LITTLE, R.J.A. (1991). Inference with survey weights. *Journal of Official Statistics*, 7, 405-424.

MILLER, R.G. (1974). An unbalanced jackknife. *Annals of Statistics*, 2, 880-891.

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

SHAO, J., and WU, C.F.J. (1987). Heteroscedasticity-robustness of jackknife variance estimators in linear models. *Annals of Statistics*, 15, 1563-1579.

SHAO, J., and WU, C.F.J. (1989). A general theory for jackknife variance estimation. *Annals of Statistics*, 17, 1176-1197.

U.S. CENSUS (1990). Dept. of Commerce. Census of Population and Housing, [United States]: public use microdata sample: 5- percent sample Computer file]. 3rd release. Washington, DC: U.S. Dept. of Commerce, Bureau of the Census [producer], 1995. Ann Arbor, MI: Inter-University Consortium for Political and Social Research [distributor], 1996.

ZHENG, H., and LITTLE, R.J.A. (2003). Penalized spline model-based estimation of the finite population total from probability-proportional-to-size samples. *Journal of Official Statistics*, 19, 99-117.

ZHENG, H., and LITTLE, R.J.A. (2004). Inference for the population total from probability-proportional-to-size samples based on predictions from a penalized spline nonparametric model. To appear in *Journal of Official Statistics*.

Survey Methodology, December 2004
Vol. 30, No. 2, pp. 219-234
Statistics Canada

A Finite Population Estimation Study with Bayesian Neural Networks

FAMING LIANG and ANTHONY YUNG CHEUNG KUK [1]

ABSTRACT

In this article, we study the use of Bayesian neural networks in finite population estimation.We propose estimators for finite population mean and the associated mean squared error. We also propose to use the student t-distribution to model the disturbances in order to accommodate extreme observations that are often present in the data from social sample surveys. Numerical results show that Bayesian neural networks have made a significant improvement in finite population estimation over linear regression based methods.

KEY WORDS: Bayesian model averaging; Bayesian neural network; Evolutionary Monte Carlo; Finite population; Markov Chain Monte Carlo; Prediction.

1. INTRODUCTION

Regression estimation is widely used in sample surveys for incorporating auxiliary population information (Cochran 1977) with the underlying model

$$y_t = \beta_0 + x_{t1}\beta_1 + \ldots + x_{tp}\beta_p + \epsilon_t, \quad t = 1, 2, \ldots, n, \quad (1)$$

where y_t is the survey variable for the t^{th} element of a population, $\boldsymbol{x}_t = (x_{t1}, \ldots, x_{tp})$ is the vector of auxiliary variables associated with $y_t, \beta_0, \beta_1, \ldots, \beta_p$ are the regression coefficients, and ϵ_t is the independent disturbance with zero mean and common variance. Although this model generally performs well, it has several inherent limitations. First, the model is specified linearly and thus can not capture some types of nonlinear relationship, which may be essential in some applications. Second, the least squares estimate, which is widely used for the model (1), may not be reliable in the presence of collinearity among the auxiliary variables. In this case, techniques, such as condition number reduction (Bankier 1990), ridge regression (Bardsley and Chambers 1984), and various variable selection procedures (Silva and Skinner 1997), have to be used to improve the poor prediction performance of the model. Third, in the presence of outliers, the least squares estimate may be severely affected by the outliers.

There are attempts to lessen the dependence of estimators on the linear model (1). Firth and Bennett (1998) identify a sufficient "internal bias calibration" condition under which a model-based estimator is automatically design consistent, regardless of how well the underlying model fits the population. The condition is met by certain estimators based on linear models, certain canonical link generalized linear models and nonparametric regression estimators constructed from them by a particular style of local likelihood fitting. Bias can also be calibrated externally, if not internally. Chambers, Dorfman and Wehrly (1993) start with a predictor of the population mean based on a heteroscedastic linear model and adjust for its bias using nonparametric regression. Kuk and Welsh (2001) propose a robustified model-based approach whereby a working model is first fitted using robust methods and subsequently the conditional distributions of the residuals given $\boldsymbol{x}$ are estimated nonparametrically to account for local model departure or outliers in localized regions.

Another way of incorporating auxiliary information into an estimator into an estimator in a design consistent manner is the model-calibrated approach first proposed by Deville and Särndal (1992). The basic idea is to choose weights that satisfy certain calibration equations and are closest to the normal Horvitz-Thompson design weights according to some distance measure. Theberge (1999) applies the calibration technique to estimate population parameters other than the means. More recently, Wu and Sitter (2001) extends the calibration approach to deal with nonlinear as well as generalized linear models by using the fitted values under these working models to set up the calibration equations. The model-calibration approach can be classified as "model-assisted" because while the efficiency of the model-calibrated estimator depends on the validity of the model, consistency does not.

There is certainly a growing trend in the survey literature in using nonlinear and nonparametric regression. Instead of model (1), one considers,

$$y_t = g(\boldsymbol{x}_t) + \epsilon_t,$$

where the regression function $g(\cdot)$ can be any arbitrary smooth function. Dorfman (1992) estimates g using the Nadaraya-Watson kernel estimator $\hat{g}$ to result in the

[1] Faming Liang, Department of Statistics, Texas A&M University, College Station, TX77843-3143. E-mail: fliang@stat.tamu.edu; Anthony Yung Cheung Kuk, Department of Statistics and Applied Probability, National University of Singapore, Singapore 117543. E-mail: stakuka@nus.edu.sg.

following model-based estimator or predictor of the finite population mean,

$$\hat{y}_K = N^{-1}\left\{\sum_{t=1}^{n} y_t + \sum_{t=n+1}^{N} \hat{g}(\boldsymbol{x}_t)\right\},$$

where it is assumed without loss of generality that the sample consists of the first n elements of the population. Kuk (1993) makes use of kernel method to estimate the conditional distribution of y given $\boldsymbol{x}$ as a way of incorporating auxiliary information in the estimation of the finite population distribution of y. For the case of scalar x, Breidt and Opsomer (2000) estimates g using local polynomial regression with design weights incorporated to account for the sampling design used and propose a generalized difference estimator,

$$\hat{y}_{LP} = N^{-1}\left\{\sum_{t=1}^{n} \frac{y_t - \hat{g}(x_t)}{\pi_t} + \sum_{t=1}^{N} \hat{g}(x_t)\right\} = N^{-1}\left\{\sum_{t=1}^{n} w_t y_t\right\},$$

where π_t is the sample inclusion probability. It can be shown that the weights w_t are calibrated to match the totals of $\boldsymbol{x}$ up to the q^{th} order, where q is the order of the local polynomial. As a consequence, $\hat{y}_{LP}$ is exactly model-unbiased if the true regression function is a polynomial of degree q or less. Breidt and Opsomer (2000) also show that $\hat{y}_{LP}$ is asymptotically design-unbiased and consistent under mild conditions. For more discussions on nonlinear and nonparametric methods, see Valliant, Dorfman and Royall (2000) (chapter 11).

In this paper, another nonlinear regression method, Bayesian neural network (BNN), is applied to the problem. BNN has an important advantage of being able to handle multivariate auxiliary variables and model selection with ease, which is not the case for many other nonlinear and nonparametric techniques. BNNs were first introduced by Buntine and Weigend (1991) and MacKay (1992), and were further developed by Neal (1996), Müller and Insua (1998), Marrs (1998), Holmes and Mallick (1998), and Liang and Wong (2001). But the BNN proposed in this paper is different from those cited above in one important respect: A prior is put on each network connection, instead of only on the number of hidden units as done in the literature. This allows us to treat the selection of network structure and the selection of input variables (auxiliary variables) uniformly. The network is trained by sampling from the joint posterior of the network structure and connection weights. The sampled network has often a sparse structure, which effectively prevents the data from being overfitted. A heavy tail distribution, such as the student t-distribution, is proposed to model the disturbances of the data with outliers. Numerical results show that BNN models have offered a significant improvement over the linear regression based models in finite population estimation.

The remaining part of this article is organized as follows. In section 2, we describe the BNN models and the associated estimators for finite populations. In section 3, we present our numerical results for one finite population example with two choices of auxiliary variables and comparisons with various linear regression based models. In section 4, we present our numerical results for another finite population example demonstrate how a cross-validation procedure can be applied to determine the parameter setting for BNN models. In section 5, we conclude the paper with a brief discussion.

2. FINITE POPULATION ESTIMATION WITH BAYESIAN NEURAL NETWORKS

2.1 Bayesian Neural Network Models

Suppose we have data pairs $D=\{(\boldsymbol{x}_1, y_1),\ldots,(\boldsymbol{x}_n, y_n)\}$, which were generated from the relationship

$$y_t = g(\boldsymbol{x}_t) + \epsilon_t, \tag{2}$$

where $y_t \in R^1, \boldsymbol{x}_t = (x_{t1},\ldots,x_{tp}) \in R^p, g(\cdot)$ is the true regression function of unknown form, and $\epsilon_t/\sigma \sim t(\nu)$ with $\nu > 2$ being a known degree of freedom of the t-distribution. Here $g(\cdot)$ may be highly nonlinear, and σ is an unknown scale parameter. We use the student t-distribution, instead of the Gaussian distribution as usual, to model the disturbances in order to accommodate extreme observations that are often present in the data from social sample surveys.

Before describing our BNN model, we first give a brief description for feed-forward neural networks. Figure 1 illustrates a one-hidden layer feed-forward neural network. It consists of four types of units, bias units, input units, hidden units, and output units. The unit to which the input features are presented is referred to as the input unit. The bias unit is a special type of input units with a constant input, say, 1. The unit where the network output is formed is referred to as the output unit. The hidden unit is so called because its input and output are only used for internal connections and are unavailable to the outside world. In a feed-forward neural network, each hidden unit independently processes the values fed to it by the units in the preceding layer and then presents its output to the units in the next layer for further processing. It has been shown by several authors (Cybenko 1989; Funahashi 1989; Hornik, Stinchcombe and White 1989) that neural networks are universal approximators in that a one-hidden layer feed-forward neural network with linear output units can approximate any continuous functions arbitrarily well on compact sets by increasing the number of hidden units. To survey regression, this is an important advantage of neural network models over other regression models. In the survey regression literature, whether model-assisted or model-based,

there is usually considerable attention paid to the consequences of model misspecification. The neural network model avoids this consideration partially due to its specific property of universal approximation. In section 2.2.1, we show that as the sample size is large, the unknown regression function $g(\cdot)$ in (2) can be well approximated by BNN models, regardless of the true function form of $g(\cdot)$. Essentially, BNN falls into the class of data-driven methods.

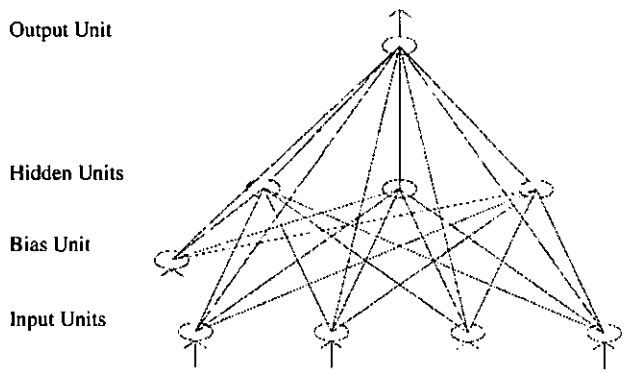


Figure 1. A fully connected one hidden layer feed-forward neural network with 4 input units, 3 hidden units and 1 output unit. The arrows indicate the direction of data feeding.

In our BNN model, the function $g(\cdot)$ in model (2) is approximated by a function of the form

$$\hat{g}(\boldsymbol{x}_t, \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma}) = \alpha_0 I_{\alpha_0} + \sum_{i=1}^{p} x_{ti}\alpha_i I_{\alpha_i} + \sum_{j=1}^{M} \beta_j I_{\beta_j} \psi\left(\sum_{i=1}^{p} x_{ti}\gamma_{ji} I_{\gamma_{ji}} + \gamma_{j0} I_{\gamma_{j0}}\right), \quad (3)$$

where I_ζ is an indicator function which indicates the effectiveness of the connection ζ; M denotes the maximum number of hidden units which is specified by users; α_0 denotes the bias term of the output unit, $\alpha_1, \ldots, \alpha_p$ denote the weights on the connections from the input units to the output unit; $\beta_1, \ldots, \beta_M$ denote the weights on the connections from hidden units to the output unit; γ_{j0} denotes the bias term of the j^{th} hidden unit, $\gamma_{j1}, \ldots, \gamma_{jp}$ denote the weights on the connections from the input units to the j^{th} hidden unit; and $\psi(\cdot)$ denotes the activation function. Sigmoid and hyperbolic tangent functions are two popular choices for the activation function. We set $\psi(z) = \tanh(z)$ for all examples of this paper.

Let Λ be the vector consisting of all indicators of model (3). Note that Λ specifies the structure of the corresponding network. Let $\boldsymbol{\alpha} = (\alpha_0, \alpha_1, \ldots, \alpha_p)$, $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_M)$, $\boldsymbol{\gamma}_j = (\gamma_{j0}, \ldots, \gamma_{jp})$, $\boldsymbol{\gamma} = (\boldsymbol{\gamma}_1, \ldots, \boldsymbol{\gamma}_M)$, and $\boldsymbol{\theta} = (\boldsymbol{\alpha}_\Lambda, \boldsymbol{\beta}_\Lambda, \boldsymbol{\gamma}_\Lambda, \sigma^2)$, where $\boldsymbol{\alpha}_\Lambda, \boldsymbol{\beta}_\Lambda$ and $\boldsymbol{\gamma}_\Lambda$ denote the non-zero subsets of $\boldsymbol{\alpha}, \boldsymbol{\beta}$ and $\boldsymbol{\gamma}$, respectively. Thus, the model (3) is completely specified by the tuple $(\boldsymbol{\theta}, \Lambda)$. For simplicity, in the following we will use $\boldsymbol{\theta}_\Lambda$ to denote a BNN model and use $\hat{g}(\boldsymbol{x}_t, \boldsymbol{\theta}_\Lambda)$ to re-denote the function $\hat{g}(\boldsymbol{x}_t, \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma})$. Also, we let $\boldsymbol{\theta}_\Lambda = (\boldsymbol{\theta}, \Lambda)$, and use $\boldsymbol{\theta}_\Lambda$ and $(\boldsymbol{\theta}, \Lambda)$ exchangeably. To conduct a Bayesian analysis for model (3), we have the following prior distributions: $\alpha_i \sim N(0, \sigma_\alpha^2)$ for $\alpha_i \in \boldsymbol{\alpha}_\Lambda$; $\beta_j \sim N(0, \sigma_\beta^2)$ for $\beta_j \in \boldsymbol{\beta}_\Lambda$; $\gamma_{ji} \sim N(0, \sigma_\gamma^2)$ for $\gamma_{ji} \in \boldsymbol{\gamma}_\Lambda$; and $f(\sigma^2) \sim 1/\sigma^2$. The total number of effective connections in Λ is $m = \sum_{i=0}^{p} I_{\alpha_i} + \sum_{j=1}^{M} I_{\beta_j} \delta(\sum_{i=0}^{p} I_{\gamma_{ji}}) + \sum_{j=1}^{M} \sum_{i=0}^{p} I_{\beta_j} I_{\gamma_{ji}}$, where $\delta(z) = 1$ if $z > 0$ and 0 otherwise. The model Λ is subject to a prior probability that is proportional to the mass put on m by a truncated Poisson (λ) with rate λ,

$$P(\Lambda) = \begin{cases} \dfrac{1}{Z}\dfrac{\lambda^m}{m!}, & m = 3, 4, \ldots, U \\ 0, & \text{otherwise} \end{cases}$$

where $U = (M+1)(p+1) + M$ is the number of connections of the full model in which all $I_\zeta = 1$; and $Z = \sum_{\Lambda \in \Omega} \lambda^m / m!$. Here we let Ω denote the set of all possible models with $3 \le m \le U$. We set the minimum number of m to three based on our views: neural networks are usually used for complex problems, and three has been a small enough number as a limiting network size. In these prior distributions, $\sigma_\alpha^2, \sigma_\beta^2, \sigma_\gamma^2$ and λ are hyper-parameters to be specified by users (discussed below). Furthermore, we assume that these prior distributions are independent *a priori*. Thus, we have the following log-posterior (up to an additive constant),

$$\begin{aligned} \log \pi(\boldsymbol{\theta}_\Lambda | D) = \text{Constant} &- \left(\frac{n}{2} + 1\right) \log \sigma^2 - \frac{\nu+1}{2} \sum_{t=1}^{n} \log\left(1 + \frac{(y_t - \hat{g}(\boldsymbol{x}_t, \boldsymbol{\theta}_\Lambda))^2}{\nu \sigma^2}\right) \\ &- \frac{1}{2} \sum_{i=0}^{p} I_{\alpha_i} \left(\log \sigma_\alpha^2 + \frac{\alpha_i^2}{\sigma_\alpha^2}\right) - \frac{1}{2} \sum_{j=1}^{M} I_{\beta_j} \delta\left(\sum_{i=0}^{p} I_{\gamma_{ji}}\right) \left(\log \sigma_\beta^2 + \frac{\beta_j^2}{\sigma_\beta^2}\right) \\ &- \frac{1}{2} \sum_{j=1}^{M} \sum_{i=0}^{p} I_{\beta_j} I_{\gamma_{ji}} \left(\log \sigma_\gamma^2 + \frac{\gamma_{ji}^2}{\sigma_\gamma^2}\right) - \frac{m}{2} \log(2\pi) \\ &+ m \log \lambda - \log(m!). \end{aligned} \quad (4)$$

Our BNN model is different from other BNN models existing in the literature in two important respects. First, the input variables of our BNN model are selected automatically by sampling from the joint posterior of the network structure and weights. Second, the structure of our BNN model is usually sparse and its performance less depends on

the initial specification for the input patterns and the number of hidden units. The sparse is in the sense that only a small number of connections are active in the network. So our BNN model avoids the problem of overfitting in a more natural way.

For data preparation and hyperparameter setting, we have the following suggestions. To avoid some weights that are trained to be extremely large or small (in absolute value) to accommodate different scales of input and output variables, we suggest that all input and output variables be normalized before feeding to the networks. In all examples of this article, the data is normalized by $(y_i - \bar{y})/S_y$, where $\bar{y}$ and S_y denote the mean and standard deviation of the training data, respectively. Based on the belief that a network with a large weight variation usually has a poor generalization performance, we suggest that $\sigma_\alpha^2, \sigma_\beta^2$ and σ_γ^2 are chosen for moderate values to penalize a large weight variation. For example, we set $\sigma_\alpha^2 = \sigma_\beta^2 = \sigma_\gamma^2 = 5$ for all examples of this article. The setting should also be fine for the other problems. The value of λ reflects our belief on the network size needed for the data under consideration. Here we follow the suggestion of Weigend, Huberman and Rumelhart (1990) to choose λ such that the number of connection weights is about one tenth of the size of the training sample. In one simulation, we assessed the influence of λ on BNN model size and predictionability. The numerical results suggest that the prediction ability of BNN models is rather robust to the variation of λ, although the BNN model size increases slowly as λ increases.

To sample from the posterior (4), a Monte Carlo algorithm, so called the reversible jump evolutionary Monte Carlo (RJEMC) algorithm, is developed. This algorithm extends the evolutionary Monte Carlo algorithm (Liang and Wong 2001) to sample from a variable dimensional space by incorporating some reversible jump moves proposed in Green (1995). For details of the algorithm, please refer to the support documents and software for the paper. They are available at http://www.stat.tamu.edu/~fliang.

2.2 Finite Population Estimation with Bayesian Neural Networks

2.2.1 Bayesian Model Averaging

In this subsection, we review some basic results of Bayesian model averaging and show one theorem for BNN models, which form the theoretical basis for the estimators described in section 2.2.2. Suppose that we are interested in estimating the quantity $\rho(\boldsymbol{\theta}_\Lambda)$, which is a function of both Λ and $\boldsymbol{\theta}$. The Bayesian estimator of $\rho(\boldsymbol{\theta}_\Lambda)$ can be written as

$$E_\pi \rho(\boldsymbol{\theta}_\Lambda) = \sum_{k=0}^{K} P(\Lambda_k | D) \int \rho(\theta_k, \Lambda_k) \pi(\boldsymbol{\theta}_k | \Lambda_k, D) d\boldsymbol{\theta}_k, \quad (5)$$

where K denotes the total number of models under consideration, $\boldsymbol{\theta}_k$ denotes the parameters associated with model Λ_k, and $\pi(\boldsymbol{\theta}_k | \Lambda_k, D)$ denotes the posterior density of $\boldsymbol{\theta}_k$ conditional on model Λ_k. Madigan and Raftery (1994) argued for this estimator that Bayesian model averaging (averaging over all the models in this fashion) accounts for the model uncertainty, and provides better predictive ability, as measured by the logarithmic scoring rule, than using any single model Λ_k. See Hoeting, Madigan, Raftery and Volinsky (1999) for a tutorial on Bayesian model averaging.

Suppose that samples $(\boldsymbol{\theta}_1, \Lambda_1), \ldots, (\boldsymbol{\theta}_M, \Lambda_M)$ have been drawn from the posterior distribution $\pi(\boldsymbol{\theta}_\Lambda | D)$ by a MCMC algorithm, then $\rho(\boldsymbol{\theta}_\Lambda)$ can be estimated by

$$\hat{\rho}(\boldsymbol{\theta}_\Lambda) = \frac{1}{M} \sum_{i=1}^{M} \rho(\boldsymbol{\theta}_{\Lambda_i}), \quad (6)$$

where $\boldsymbol{\theta}_{\Lambda_i} = (\boldsymbol{\theta}_i, \Lambda_i)$. Applying the standard Markov chain theory (Tierney 1994; Roberts and Casella 1999), under regularity conditions we have the following results. If $E_\pi |\rho(\boldsymbol{\theta}_\Lambda)| < \infty$, then

$$\frac{1}{M} \sum_{i=1}^{M} \rho(\boldsymbol{\theta}_{\Lambda_i}) \to E_\pi \rho(\boldsymbol{\theta}_\Lambda), \quad \text{a.s.}, \quad (7)$$

as $M \to \infty$. Furthermore, if $E_\pi |\rho(\boldsymbol{\theta}_\Lambda)|^{2+\delta} < \infty$ for some $\delta > 0$, then

$$M^{1/2} \left\{ \frac{1}{M} \sum_{i=1}^{M} \rho(\boldsymbol{\theta}_{\Lambda_i}) - E_\pi \rho(\boldsymbol{\theta}_\Lambda) \right\} \to N(0, \tau^2), \quad (8)$$

for some positive constant τ^2 as $M \to \infty$, and the convergence is in distribution.

Similar to (7) and (8), we have the following theorem for BNN models, of which proof is presented in Appendix.

Theorem 2.1 *Let $D = \{(x_1, y_1), \ldots, (x_n, y_n)\}$ denote a simple random sample drawn from a population which can be modeled by model (2). Let $(\boldsymbol{\theta}_1, \Lambda_1), \ldots, (\boldsymbol{\theta}_M, \Lambda_M)$ denote the sample drawn from the posterior distribution $\pi(\boldsymbol{\theta}_\Lambda | D)$, given in (4), by a MCMC method. Then, for any x_0 drawn from the same distribution with the observations D, we have*

(a)

$$E_\pi |\hat{g}(x_0, \boldsymbol{\theta}_\Lambda)|^{2+\delta} < \infty, \quad (9)$$

for some $\delta > 0$, as $n \to \infty$.

(b)

$$\frac{1}{M} \sum_{i=1}^{M} \hat{g}(x_0, \boldsymbol{\theta}_{\Lambda_i}) \to g(x_0), \quad a.s., \quad (10)$$

as $n \to \infty$ and $M \to \infty$.

(c)

$$M^{1/2} \left[\frac{1}{M} \sum_{i=1}^{M} \hat{g}(x_0, \boldsymbol{\theta}_{\Lambda_i}) - g(x_0) \right] \to N(0, \tau_*^2), \quad (11)$$

for some positive constant τ_*^2 as $n \to \infty$ and $M \to \infty$, and the convergence is in distribution.

To show some properties of moments of $1/M\sum_{i=1}^{M}\hat{g}(\boldsymbol{x}_0,\boldsymbol{\theta}_{\Lambda_i})$, we need the following theorem (Billingsley 1986, page 348, Corollary),

Theorem 2.2 *Let r be a positive integer. If $X_M \to X$ in distribution and $\sup_m E|X_m|^{r+\delta}<\infty$, where $\delta>0$, then $E|X|^r<\infty$ and $EX_m^r \to EX^r$.*

Following from (9), (11) and Theorem 2.2, we know

$$ME\left[\frac{1}{M}\sum_{i=1}^{M}\hat{g}(\boldsymbol{x}_0,\boldsymbol{\theta}_{\Lambda_i})-g(\boldsymbol{x}_0)\right]^2 \to \tau_*^2, \qquad (12)$$

as $n\to\infty$ and $M\to\infty$. It implies that

$$E\left[\frac{1}{M}\sum_{i=1}^{M}\hat{g}(\boldsymbol{x}_0,\boldsymbol{\theta}_{\Lambda_i})-g(\boldsymbol{x}_0)\right]^2 = \frac{\tau_*^2}{M}+o\left(\frac{1}{M}\right) \qquad (13)$$

holds as n and M are both large.

Note we have shown that (11) and (13) hold as the sample size $n\to\infty$. In the context of finite population, especially for a small finite population, a more precise expression for (11) and (13) would be

$$M^{1/2}\left[\frac{1}{M}\sum_{i=1}^{M}\hat{g}(\boldsymbol{x}_0,\boldsymbol{\theta}_{\Lambda_i})-E(y_0|D,\boldsymbol{x}_0)\right]\to N(0,\tau_*^2), \qquad (14)$$

and

$$E\left[\frac{1}{M}\sum_{i=1}^{M}\hat{g}(\boldsymbol{x}_0,\boldsymbol{\theta}_{\Lambda_i})-E(y_0|D,\boldsymbol{x}_0)\right]^2 = \frac{\tau_*^2}{M}+o\left(\frac{1}{M}\right), \qquad (15)$$

where $E(y_0|D,\boldsymbol{x}_0)$ denotes the prediction of y_0 which is the survey variable corresponding to $\boldsymbol{x}_0$. The equations (14) and (15) take into accounts the possible bias of the sample D. In the case that the population constitutes many exact copies of the sample $D, E(y_0|D,\boldsymbol{x}_0)=g(\boldsymbol{x}_0)$ holds, and equations (14) and (15) are reduced to (11) and (13), respectively.

2.2.2 BMA Estimators in Finite Populations

Consider a finite population of N distinguishable elements. Associated with the i^{th} elements are the survey variable y_i and the auxiliary variables $\boldsymbol{x}_i$. The values $\boldsymbol{x}_1,\ldots,\boldsymbol{x}_N$ are known for the entire population, while y_i is known only if the i^{th} unit is selected in the sample. Suppose a simple random sample $D=\{(\boldsymbol{x}_1,y_1),\ldots,(\boldsymbol{x}_n,y_n)\}$ has been drawn from the finite population, a BNN model has been built for the sample, and $(\boldsymbol{\theta}_1,\Lambda_1),\ldots,(\boldsymbol{\theta}_M,\Lambda_M)$ have been drawn from the posterior distribution of the BNN model, the BMA estimator for the mean of the finite population is

$$\bar{y}_{\text{BNN}} = f\,\bar{y}+\frac{1}{MN}\sum_{i=1}^{M}\sum_{t=n+1}^{N}\hat{g}(\boldsymbol{x}_t,\boldsymbol{\theta}_{\Lambda_i}),$$

where $\bar{y}$ is the sample mean of $y_1,\ldots,y_n$, and $f=n/N$ is the sample fraction. About this estimator, we have the following comments. First, $\bar{y}_{\text{BNN}}$ is a model-based estimator, so that all the inference is with respect to the model for the y_i's, not the survey design. As long as the model holds, the BNN estimator will have the mean squared error properties described below for any ignorable sampling design. Second, this estimator is identical to that proposed in Dorfman (1992), except that the BNN is replaced by a kernel-based regression. Third, this estimator can be used to estimate the mean of a finite population as long as each of the unsampled elements has the same distribution as the sample D.

The accuracy of an estimate can be measured by its mean squared error $E(\bar{y}_{\text{BNN}}-\bar{Y})^2$, where $\bar{Y}$ denotes the true population mean. To estimate $E(\bar{y}_{\text{BNN}}-\bar{Y})^2$, we first consider

$$\begin{aligned}
&E\left[(\bar{y}_{\text{BNN}}-\bar{Y})^2\,|\,D,\boldsymbol{X}_{n+1}^N\right]\\
&=E\left[\left\{\begin{array}{l}\dfrac{1}{MN}\displaystyle\sum_{i=1}^{M}\sum_{t=n+1}^{N}\hat{g}(\boldsymbol{x}_t,\boldsymbol{\theta}_{\Lambda_i})\\ -\dfrac{1}{N}\displaystyle\sum_{t=n+1}^{N}(g(\boldsymbol{x}_t)+\in_t)\end{array}\right\}^2\Bigg|D,\boldsymbol{X}_{n+1}^N\right]\\
&=\frac{(N-n)^2}{N^2}E\left[\left\{\begin{array}{l}\dfrac{1}{M(N-n)}\displaystyle\sum_{i=1}^{M}\sum_{t=n+1}^{N}\\ \hat{g}(\boldsymbol{x}_t,\boldsymbol{\theta}_{\Lambda_i})-\dfrac{1}{N-n}\displaystyle\sum_{t=n+1}^{N}g(\boldsymbol{x}_t)\end{array}\right\}^2\Bigg|D,\boldsymbol{X}_{n+1}^N\right]\\
&\quad+\frac{N-n}{N^2}\operatorname{var}(\in_t)\\
&=\frac{(N-n)^2}{N^2}E\left[\left\{\begin{array}{l}\dfrac{1}{M(N-n)}\displaystyle\sum_{i=1}^{M}\sum_{t=n+1}^{N}\hat{g}(\boldsymbol{x}_t,\boldsymbol{\theta}_{\Lambda_i})\\ -E(\bar{y}_u\,|\,D,\boldsymbol{X}_{n+1}^N)\\ +E(\bar{y}_u\,|\,D,\boldsymbol{X}_{n+1}^N)-E(\bar{y}_u)\end{array}\right\}^2\Bigg|D,\boldsymbol{X}_{n+1}^N\right]\\
&\quad+\frac{N-n}{N^2}\operatorname{var}(\in_t)\\
&\approx\frac{\tau_D^2}{M}+(1-f)^2\left\{E(\bar{y}_u)\,|\,D,\boldsymbol{X}_{n+1}^N)-E(\bar{y}_u)\right\}^2\\
&\quad+\frac{1-f}{N}\operatorname{var}(\in_t), \qquad (16)
\end{aligned}$$

where $\boldsymbol{X}_{n+1}^N=(\boldsymbol{x}_{n+1},\ldots,\boldsymbol{x}_N)$ denotes the set of auxiliary vectors of the unsampled elements; $\bar{y}_u$ denotes the averaged survey value of the unsampled elements, and

$$E(\bar{y}_u)=\frac{1}{N-n}\sum_{t=n+1}^{N}g(\boldsymbol{x}_t).$$

The last approximation of (16) follows from (15), that is, as M is large,

$$E\left\{\frac{1}{MN}\sum_{i=1}^{M}\sum_{t=n+1}^{M}\hat{g}(\boldsymbol{x}_t,\boldsymbol{\theta}_{\Lambda_i})-(1-f)E(\bar{y}_u|D,X_{n+1}^N)\right\}^2\approx\frac{\tau_D^2}{M},$$

for some positive constant τ_D^2. The term $E(\bar{y}_u|D,X_{n+1}^N)-E(\bar{y}_u)$ is the prediction bias due to the randomness or sampling bias of D. Following from (16), we have

$$E(\bar{y}_{\mathrm{BNN}}-\bar{Y})^2\approx\frac{E\tau_D^2}{M}+(1-f)^2$$
$$E\left\{E(\bar{y}_u\mid D,X_{n+1}^N)-E(\bar{y}_u)\right\}^2+\frac{1-f}{N}\mathrm{var}(\in_t). \qquad (17)$$

The quantity τ_D^2 can be estimated by the batch means method (Roberts 1996) as follows. Run the Markov chain for $M=rs$ iterations, where s is the batch size and is assumed suffciently large such that

$$\bar{y}_{\mathrm{BNN},k}=f\,\bar{y}+\frac{1}{sN}\sum_{i=(k-1)s+1}^{ks}\sum_{t=n+1}^{N}\hat{g}(\boldsymbol{x}_t,\boldsymbol{\theta}_{\Lambda_i}),$$

is approximately independently $N(f\,\bar{y}+(1-f)E(\bar{y}_u|D,X_{n+1}^N),\tau_D^2/s$. Therefore τ_D^2 can be approximated by

$$\hat{\tau}_D^2=\frac{s}{r-1}\sum_{k=1}^{r}(\bar{y}_{\mathrm{BNN},k}-\bar{y}_{\mathrm{BNN}})^2, \qquad (18)$$

which can be substituted into (17) in lieu of $E\tau_D^2$. Under the assumption $\in_t/\sigma\sim t(v)$, the BMA estimator $\mathrm{var}(\in_t)$ is

$$\hat{\mathrm{var}}(\in_t)=\frac{v}{v-2}\frac{1}{M}\sum_{i=1}^{M}\hat{\sigma}_i^2. \qquad (19)$$

Under the assumption that the population is made up of exact copies of the training data, we have $E(\bar{y}_u|D,X_{n+1}^N)-E(\bar{y}_u)\approx\bar{\hat{y}}-\bar{y}$, where $\bar{\hat{y}}$ denotes the fitted sample mean, and

$$E(\bar{\hat{y}}-\bar{y})^2=E\left\{\frac{1}{n}\sum_{t=1}^{n}\hat{\in}_t\right\}^2=\frac{1}{N}\mathrm{var}(\hat{\in}_t), \qquad (20)$$

where $\hat{\in}_t=\sum_{i=1}^{M}\hat{g}(\boldsymbol{x}_t,\boldsymbol{\theta}_{\Lambda_i})/M-y_t$ is the residual of the t^{th} element of D, and $\hat{\in}_t$'s are assumed to be iid and $E(\hat{\in}_t)=0$. Under the true model, we have $\mathrm{var}(\hat{\in}_t)\approx\mathrm{var}(\in_t)$. Hence, we suggest $E\{E(\bar{y}_u|D,X_{n+1}^N)-E(\bar{y}_u)\}^2$ be estimated by

$$\hat{\mathrm{Bias}}^2=\frac{1}{n}\hat{\mathrm{var}}(\in_t). \qquad (21)$$

In summary, $E(\bar{y}_{\mathrm{BNN}}-\bar{Y})^2$ can be estimated by

$$\hat{E}(\bar{y}_{\mathrm{BNN}}-\bar{Y})^2=\frac{\hat{\tau}_D^2}{M}+(1-f)^2\hat{\mathrm{Bias}}^2$$
$$+\frac{1-f}{N}\hat{\mathrm{var}}(\in_t)=\frac{\hat{\tau}_D^2}{M}+\frac{1-f}{n}\hat{\mathrm{var}}(\in_t). \qquad (22)$$

As $M\to\infty$ we have

$$\hat{E}(\bar{y}_{\mathrm{BNN}}-\bar{Y})^2=\frac{1-f}{n}\hat{\mathrm{var}}(\in_t). \qquad (23)$$

We note that this estimate is identical in form to that given by Cochran (1977) for the linear regression estimator.

3. FIRST SIMULATION STUDY

3.1 The Data

Our simulation population comprises 426 records for heads of household surveyed using the sample (long) questionnaire during the 1988 Test Population Census of Limeira, in São Paulo state, Brasil. This test was carried out as a pilot survey during the preparation for the 1991 Brazilian Population Census. For a detailed description for the test census, see Silva and Skinner (1997). We followed Silva and Skinner (1997) to consider the total monthly income as the main survey variable (y) together with 11 potential auxiliary variables, namely,

x_1	indicator of sex of head of household equal male;
x_2	indicator of age of head of household less than or equal to 35;
x_3	indicator of age of head of household greater than 35 and less than or equal to 55;
x_4	total number of rooms in household;
x_5	total number of bathrooms in household;
x_6	indicator of ownership of household;
x_7	indicator that household type is house;
x_8	indicator of ownership of at least one car in household;
x_9	indicator of ownership of color TV in household;
x_{10}	years of study of head of household;
x_{11}	proxy of total monthly income of head of household.

Figure 2, the scatter plots of y versus the 11 auxiliary variables, shows that a linear regression model is not appropriate for the data. Although y and x_{11} are strongly linearly correlated, the scatter plots of y versus some other auxiliary variables, say x_4, x_5 and x_{10}, suggest that their relationships can not be well modeled by a linear regression. In addition, if the data is modeled by a linear regression, the outlier, the 53^{th} element, may have a high influence on fitting and prediction of the model. More precisely, if the element is included in the training data, the fitted response curve will have a up-drift comparing to the true curve and as a result the finite population mean will be overestimated; if

the element is not included in the training data, prediction will proceed as though there were not outliers and as a result the finite population mean will be underestimated. The presence of the strong influence element also mounts a great challenge on BNN models and other data analysis strategies.

We followed Silva and Skinner (1997) to construct two alternative sets of auxiliary variables for simulations. The first set contains $x_1,\ldots,x_4$ and x_{11}, which includes the proxy variable x_{11} and has a reasonable explanatory power in predicting y. The second set contains $x_1,\ldots,x_{10}$, which has a weaker explanatory power than the first one due to the exclusion of x_{11}. So these two sets illustrate the predictive performances of BNN models with strong and weak auxiliary variables, respectively. As in Silva and Skinner (1997), 1,000 sample replicates of size 100 from this simulation population are selected by simple random sampling without replacement. The following computation were performed on the 1,000 replicates.

For each replicate, say k, it was analyzed by BNN models and various linear regression based strategies (reviewed below). For any strategy, the population mean estimate and its estimated mean squared error for the replicate k are denoted by $\bar{y}(k)$ and $V(\bar{y}(k))$, respectively. The computational results were summarized by computing the mean (MEAN), bias (BIAS), mean square error (MSE) and average of mean squared error estimates (AVMSE) from the set of the 1,000 replicates, given respectively by

$$\text{MEAN} = \sum_{k=1}^{S} \bar{y}(k)/S;$$

$$\text{BIAS} = \text{MEAN} - \bar{Y};$$

$$\text{MSE} = \sum_{k=1}^{S} \left[\bar{y}(k) - \bar{Y}\right]^2 / S;$$

$$\text{AVMSE} = \sum_{k=1}^{S} V(\bar{y}(k))/S,$$

where S is the total number of sample replicates under consideration, and $\bar{Y}$ = 194.34 for the simulation population. Empirical coverage rates for 95% confidence intervals based on asymptotic normal theory were also computed for each strategy and these rates, expressed as percentages, are presented in the last columns of Tables 1 and 3.

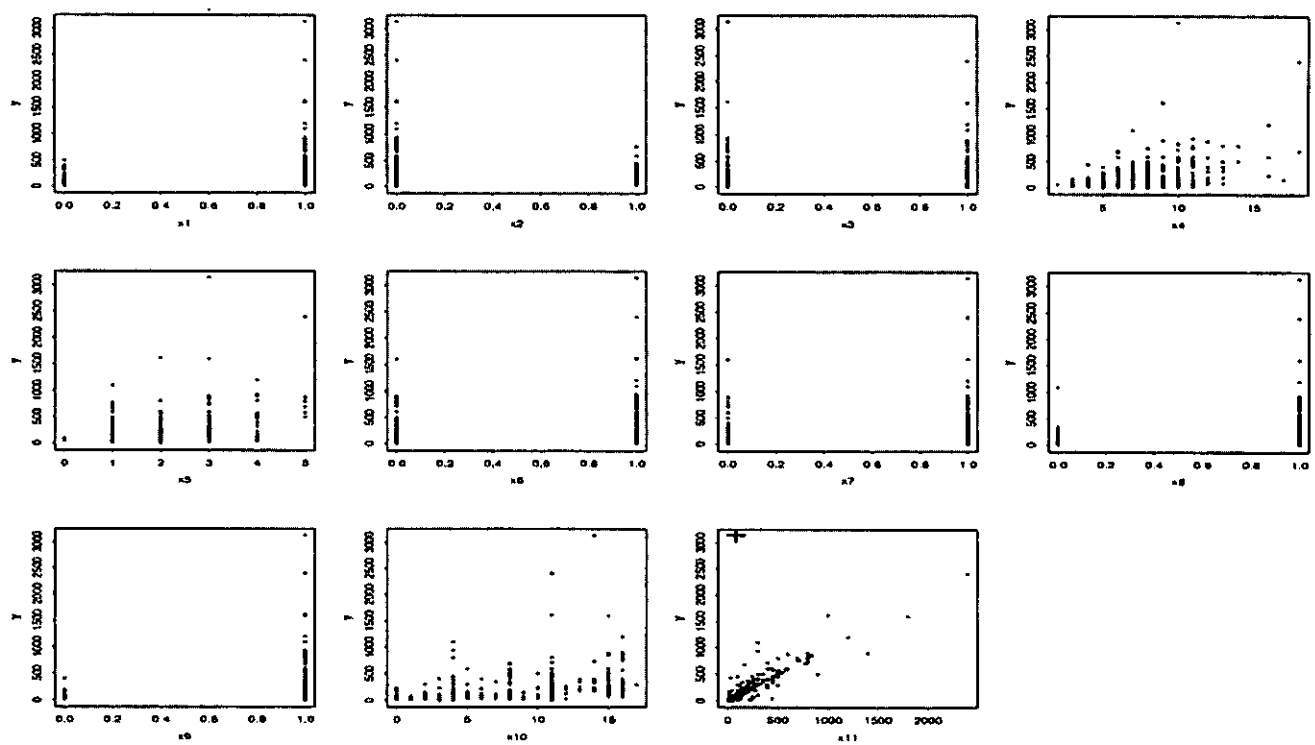


Figure 2. Scatter plots of the response variable y versus the auxiliary variables. In the plot of y versus x_{11} the "+" represents the 53rd element of the population.

3.2 Review of the Linear Regression Based Strategies

The linear regression based strategies that have been considered by Silva and Skinner (1997) are listed as follows.

SM)	Sample mean estimator, with no auxiliary variables $(\bar{y}, V_s)$.
Fs)	Forward selection of auxiliary variables with $(\bar{y}_r, V_s)$.
Fd)	Forward selection of auxiliary variables with $(\bar{y}_r, V_d)$.
Fg)	Forward selection of auxiliary variables with $(\bar{y}_r, V_g)$.
Bs)	Best subset selection from all subsets of auxiliary variables with $(\bar{y}_r, V_s)$.
Bd)	Best subset selection from all subsets of auxiliary variables with $(\bar{y}_r, V_d)$.
Bg)	Best subset selection from all subsets of auxiliary variables with $(\bar{y}_r, V_g)$.
Fl)	Fixed subset of auxiliary variable with $(\bar{y}_r, V_s)$.
SS)	Saturated subset of auxiliary variable with $(\bar{y}_r, V_s)$.
FR)	Forward subset selection using SAS PROC REG, with $(\bar{y}_r, V_s)$.
CN)	Condition number reduction subset selection procedure with $(\bar{y}, V_s)$.
RI)	Ridge regression estimator proposed by Dunstan and Chambers (1986).

To facilitate the description for the above strategies, we define the following notations. Let $U = \{1, \ldots, N\}$ denote a finite population of N distinguishable elements, $D \subset U$ denote a sample replicate of n elements drawn from U by simple random sampling without replacement, $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{ip})'$ be the vector of auxiliary variables associated with the i^{th} element, and $\beta = (\beta_1, \ldots, \beta_p)$. Let $\bar{\boldsymbol{X}} = N^{-1}\sum_{i\in u}\boldsymbol{x}_i$ be the vector of population means, $\bar{\boldsymbol{x}} = n^{-1}\sum_{i\in D}\boldsymbol{x}_i$ be the vector of sample means, $\bar{y} = n^{-1}\sum_{i\in D} y_i$ be the sample mean of the response variable, $\hat{S}_x = n^{-1}\sum_{i\in D}(\boldsymbol{x}_i - \bar{\boldsymbol{x}})(\boldsymbol{x}_i - \bar{\boldsymbol{x}})'$, $\hat{S}_{xy} = n^{-1}\sum_{i\in D}(\boldsymbol{x}_i - \bar{\boldsymbol{x}})(y_i - \bar{y})$, $g_i = 1 + (\bar{\boldsymbol{X}} - \bar{\boldsymbol{x}})'\hat{S}_x^{-1}(\boldsymbol{x}_i - \bar{\boldsymbol{x}})$ the so-called g-weights (Särndal, Swensson and Wretman 1989), and $\hat{\boldsymbol{\beta}} = \hat{S}_x^{-1}\hat{S}_{xy}$ the least squares estimator of $\boldsymbol{\beta}$ The regression estimator of $\bar{Y}$ is

$$\bar{y}_r = \bar{y} + (\bar{\boldsymbol{X}} - \bar{\boldsymbol{x}})\hat{\boldsymbol{\beta}}.$$

The V_s, V_d is and V_g are three estimators of the mean squared error of $\bar{y}_r$. The V_s is given by Cochran (1977, page 195),

$$V_s = \frac{1-f}{n(n-p-1)}\sum_{i\in D}\hat{\epsilon}_i^2,$$

where $\hat{\epsilon}_i = (y_i - \bar{y}) - (\boldsymbol{x}_i - \bar{\boldsymbol{x}})'\hat{\boldsymbol{\beta}}$ and $f = n/N$ is the sample fraction. The V_d is generalized (from $p=1$ to $p>1$) from one estimator studied by Deng and Wu (1987) and it is expected to have a smaller bias than V_s (Silva 1996),

$$V_d = \frac{1-f}{n(n-1)}\sum_{i\in D}\alpha_i\hat{\epsilon}_i^2,$$

where

$$\alpha_i = (g_i^2 - 2g_i f + f)/\left\{(1-f)\left[1 - (\boldsymbol{x}_i - \bar{\boldsymbol{x}})'\hat{S}_x^{-1}(\boldsymbol{x}_i - \bar{\boldsymbol{x}})/(n-1)\right]\right\}.$$

The V_g is modified from one estimator given by Särndal *et al.* (1989), and it has a similar performance to V_d,

$$V_g = \frac{1-f}{n(n-p-1)}\sum_{i\in D} g_i^2\hat{\epsilon}_i^2.$$

The best subset selection strategy (Bs, Bd and Bg) is to choose one subset which has the smallest mean squared error estimate among all 2^p possible subsets. The forward selection strategy (Fs, Fd and Fg) starts with the sample mean as an estimator, then adds the variable which minimizes the mean squared error estimate, and the procedure is repeated until the mean squared error estimate starts to increase. Refer to Silva and Skinner (1997) for details of the implementations of the strategies CN and RI.

3.3 Illustration on One Sample Replicate

To understand the behavior of $\bar{y}_{\text{BNN}}$ in presence of outliers and the role played by ν in robust inference, we focus on one particular sample. The training data comprises the first 100 elements of the population, and the auxiliary variables include $x_1, \ldots, x_4$ and x_{11} as the first explanatory set. Note that the 53^{th} element has been included in the training data.

For BNN models, we set $\lambda = 5$ and $M = 8$ which produces 62 connections for the full BNN model, and tried $\nu = 25$, 50, 100, 200 and $+\infty$, where $\nu = +\infty$ is equivalent to the assumption $\epsilon_i \sim N(0, \sigma^2)$. For each setting, RJEMC was run as follows: the network connections were first set to some random numbers drawn from $N(0, 0.01)$, and then were updated for 1,000 iterations in the parameter space of the full model, *i.e.*, all indicator variables are set to 1 in those iterations. After the initialization process, 4,000 iterations of RJEMC were run, and 800 samples were collected from these iterations at the lowest temperature level with an equal time space. The convergence of RJEMC can be diagnosed using the Gelman-Rubin statistic $\hat{R}$ (Gelman and Rubin 1992) based on multiple independent runs. Figure 3 shows $\hat{R}$ values computed from 10 independent runs. For each sample replicate of the simulation population, RJEMC converges ($\hat{R} < 1.1$) very fast, usually within the first 500 iterations (100 BNN samples). We discarded the first 200 samples for the burn-in process, and used the remaining 600 samples for the further inference.

For comparison, the linear regression model (1) was also applied to this sample replicate.

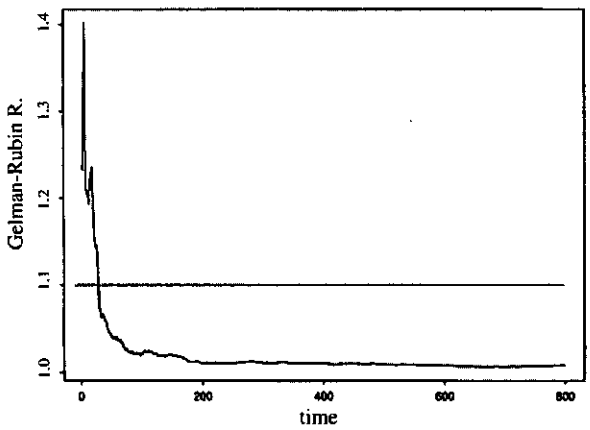


Figure 3. Gelman-Rubin statistic $\hat{R}$. The curve was computed based on 10 independent runs of RJEMC. The random errors are assumed to be distributed according to $t(100)$.

Figure 4 shows the original data together with the fitted and predicted values produced by various models. The BNN results were all obtained in one run of RJEMC. It can be seen that the linear regression model is not appropriate for this population as some fitted and predicted values produced by the model are negative for this sample replicate. Also, the fitted response curve (the solid curve in Figure 4(a) and 4(b)) is strongly influenced by the 53th element and lies above almost two-thirds of the data points. A similar phenomenon occurs for the prediction of unsampled values, see Figure 4(c) and 4(d). As a result, the population mean is overestimated (Figure 5). Comparing to that of the linear regression model, the results of the BNN models are less affected by the 53th element, especially for those computed with small values of ν. Figure 5 shows that as ν decreases, the estimated population mean by BNN models gets closer and closer to the true value, and the estimated 95% confidence interval of the population mean becomes narrower and narrower. It indicates that the influence of the 53th element on these estimates becomes weaker and weaker as ν decreases. This is not surprising as the use of a heavily tailed error distribution is known to make the inference more robust.

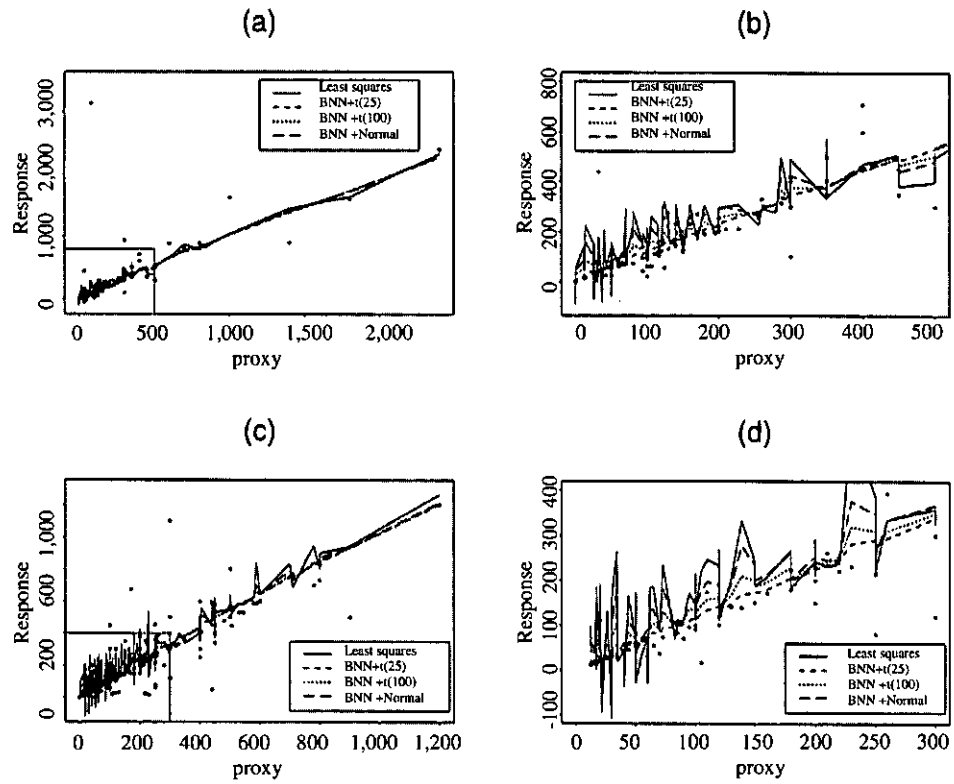


Figure 4. Fitted and predicted response curves by various models. The curves are plotted against the proxy variable, and the true response values are shown by points. (a) The fitted response curves for the sampled elements. (b) The amplification of the square region of (a). (c) The predicted response curves for the unsampled elements. (d) The amplification of the square region of (c), and for clearness only every fourth elements are plotted in the order of sorted proxy values.

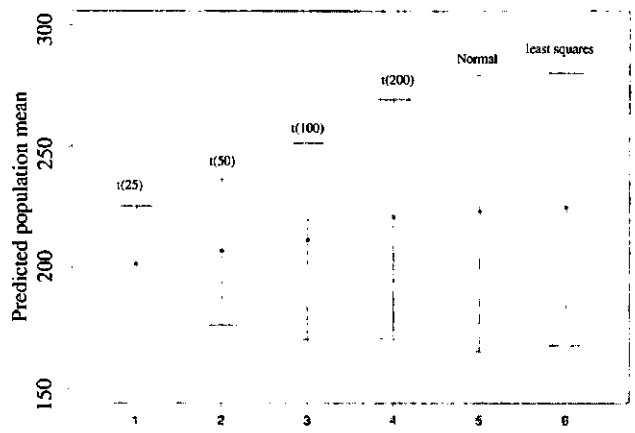


Figure 5. Estimated population mean and the associated 95% confidence interval by various models.The dotted line shows the true population mean which is 194.34.

3.4 Numerical Results on More Sample Replicates

BNN models were applied to analyze the 1,000 sample replicates. For each sample replicate of the first explanatory set, we set $\nu = 100$, $\lambda = 5$ and $M = 8$ which produces 62 connections for the full BNN model. RJEMC was run as described in section 3.3. In each run 600 BNN samples were obtained for the inference. The computational results were summarized in Table 1. It shows that BNN models have made a significantly improvement over the linear regression based models in population mean estimation for the first explanatory set. Although the BNN estimate is slightly biased (The relative bias is about 2.5% in terms of absolute values and is still acceptable.), it has the smallest MSE value among all estimates in Table 1 and the highest nominal coverage probability among the estimates with smaller MSE values (the boldfaced rows). As discussed in the last subsection, we expect $\bar{y}_{\text{BNN}}$ to behave differently for samples containing and not containing the outlying element 53. When averaged over only those samples that contain element 53, $\bar{y}_{\text{BNN}}$ with $\nu = 50$ performs very well with bias 1.51 and 99.6% coverage. The result is obviously not as good as for those samples not containing element 53 due to the inevitable underestimation of the finite population mean. Frankly, there is not much one can do if there are outliers in the population but none in the sample. No statistical method based on sample information alone will be able to predict the occurrence of outliers in the non-sample. We believe that $\bar{y}_{\text{BNN}}$ will perform very well for populations without outliers due to the universal approximation property of neural networks and the technique of Bayesian model averaging.

Let $\bar{x}_{11}$ denote the average of proxy values of the elements in one sample replicate. To see how the performance of the BNN models varied with $\bar{x}_{11}$, we ordered the 1,000 sample replicates according to their values of $\bar{x}_{11}$ and divided them into 20 groups of 50 replicates, the first group containing the 50 replicates whose $\bar{x}_{11}$ are smallest, and so forth. For each group, we calculated MEAN, MSE and AVMSE. Figure 6 shows these conditional values. From Figure 6(a) it is easy to see that BNN models possess one good property, namely, the population mean estimate is not sensitive to the value of $\bar{x}_{11}$. From Figure 6(b) it is easy to see that AVMSE provides an essentially unbiased estimate for MSE regardless of averaged proxy values.

To assess the influence of ν, M and λ on BNN model size and prediction ability for the first explanatory set, we conducted three groups of experiments. In the first group of experiments, we fixed $M = 8$ and $\lambda = 5$, and varied the value of ν, $\nu = 50$, 100 and 150. In the second group of experiments, we fixed $\nu = 100$ and $\lambda = 5$, and varied the value of M, $M = 6$, 8 and 10. In the third group of experiments, we fixed $\nu = 100$ and $M = 8$, and varied the value of λ, $\lambda = 4$, 5 and 6. For each setting, RJEMC was run as described in section 3.3 for the 1,000 sample replicates. The computational results were summarized in Table 2. It shows that the averaged model size produced by each setting is about the same, although it increases slowly as M and λ increase. The results of the first group of experiments show clearly that for BNN models there is a trade-off between BIAS and MSE or AVMSE by choosing the value of ν. The results of the second and third group of experiments show that BIAS, MSE, AVMSE and the coverage probability are rather stable to the variation of M and λ, although the latter three statistics have a slow tendency to increase as M and λ increase. The increasing trend of these statistics is due to the fact that the neural networks tend to be overfitted as M and λ increase.

Table 1

Bias, mean squared error, average of mean squared error estimates and empirical coverage of various estimation strategies for the population mean using $x_1,\ldots,x_4$ and x_{11} as auxiliary variables. Figures other than BNN are reproduced from Silva and Skinner (1997).

Estimation strategy	BIAS	MSE	AVMSE	Coverage[a] (%)
SM) Sample mean $(\bar{y}, V_s)$	0.25	620.09	619.05	91.8
CN) Cond. num. red. $(\bar{y}, V_s)$	0.34	507.33	483.63	89.8
RI) Ridge	2.12	304.95	257.07	82.5
Fs) Forward $(\bar{y}_r, V_s)$	0.40	233.78	239.62	82.7
Fd) Forward $(\bar{y}_r, V_d)$	**−1.25**	**188.08**	**196.88**	**82.0**
Fg) Forward $(\bar{y}_r, V_g)$	**−1.28**	**188.38**	**192.73**	**81.1**
Bs) Best $(\bar{y}_r, V_s)$	0.44	236.90	239.49	82.7
Bd) Best $(\bar{y}_r, V_d)$	**−1.22**	**190.52**	**196.84**	**82.0**
Bg) Best $(\bar{y}_r, V_g)$	**−1.24**	**190.83**	**192.71**	**81.1**
FI) Fixed $(\bar{y}_r, V_s)$	0.29	227.90	241.24	83.3
SS) Saturated $(\bar{y}_r, V_s)$	0.30	233.58	242.32	82.5
FR) Proc REG $(\bar{y}_r, V_s)$	0.38	235.86	240.26	82.5
BNN) $t(100)$	**−4.91**	**138.11**	**127.14**	**84.8**

[a] Nominal 95% coverage.

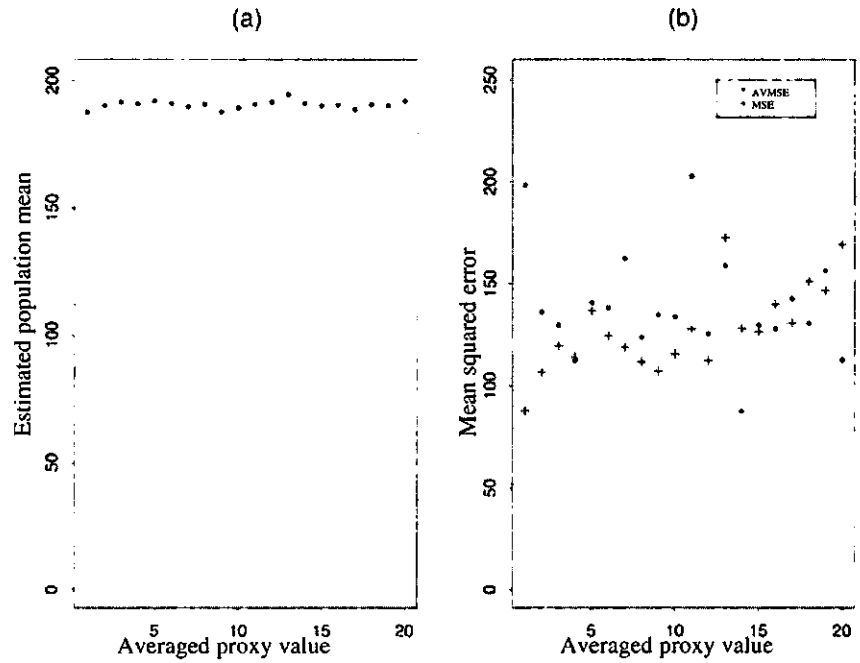


Figure 6. MEAN (panel (a)), MSE and AVMSE (Panel (b)) conditional on the averaged proxy values. The 1,000 sample replicates are ordered on $\bar{x}_{11}$ and divided into 20 groups of 50 samples.

Table 2
Assessment of the influence of ν, M and λ on BNN model size and prediction ability for the first explanatory set. For convenience of comparison, the results of the setting $\nu = 100$, $M = 8$ and $\lambda = 5$ were repeated in panels B and C.

Experiment	ν	M	λ	Size[a]	BIAS	MSE	AVMSE	Coverage[b] (%)
A	50	8	5	10.53	−6.78	131.78	90.08	82.0
	100	8	5	10.70	−4.91	138.11	127.14	84.8
	150	8	5	10.79	−3.81	156.55	160.28	85.5
B	100	6	5	9.52	−4.90	136.72	122.58	84.1
	100	8	5	10.70	−4.91	138.11	127.14	84.8
	100	10	5	11.83	−5.14	140.13	132.20	86.4
C	100	8	4	9.42	−4.94	138.04	125.99	85.2
	100	8	5	10.70	−4.91	138.11	127.14	84.8
	100	8	6	11.83	−4.92	139.62	128.64	85.7

[a] Size= $\sum_{k=1}^{1,000} \sum_{i=1}^{M} m(\Lambda_i)/M$ /1,000, where $m(\Lambda_i)$ is the number of connections of the neural network Λ_i.

[b] Nominal 95% coverage.

The above experiments also address the issue of model misspecification. Note the BNN model proposed in this paper is specified by the three parameters, ν, M and λ. Table 2 shows that the BNN model can still perform well even when the parameter setting has some departures from the optimal setting. In practice, the setting of ν, M and λ can be determined by a cross-validation experiment. This will be demonstrated in the second simulation study.

Finally, we consider the weaker set of auxiliary variables $x_1, \ldots, x_{10}$. For each sample replicate, we set $\nu = 100$, $\lambda = 5$ and $M = 8$ which produces 107 connections for the full BNN model. RJEMC was run as in section 3.3. The computational results were summarized in Table 3. It shows clearly that BNN models continue to provide a significant improvement over the linear regression based models in population mean estimation when the strongest predictor x_{11} is excluded. The BNN estimate has the smallest MSE value among all estimates in Table 3, and has the smallest bias and the highest nominal coverage probability among the estimates with smaller MSE values (the boldfaced rows).

To assess the influence of ν, M and λ on BNN model sizes and prediction abilities for the second explanatory set, we conducted the same three groups of experiments as for

the first explanatory set. The computational results were summarized in Table 4. Panel A shows again the trade-off between BIAS and MSE or AVMSE made for BNN models by the value of v. Panels B and C show that BIAS, MSE, AVMSE and the coverage probability have an even more stable performance across different choices of M and λ than that of the first explanatory set.

Table 3

Bias, mean squared error, average of mean squared error estimates and empirical coverage of various estimation strategies for the population mean using $x_1, \ldots, x_{10}$ as auxiliary variables. Figures other than BNN are reproduced from Silva and Skinner (1997).

Estimation strategy	BIAS	MSE	AVMSE	Coverage[a] (%)
SM) Sample mean $(\bar{y}, V_s)$	0.25	620.09	619.05	91.8
CN) Cond. num. red. $(\bar{y}, V_s)$	3.49	562.91	450.36	87.3
RI) Ridge	1.05	480.18	472.82	89.4
Fs) Forward $(\bar{y}_r, V_s)$	0.06	468.46	397.99	86.7
Fd) Forward $(\bar{y}_r, V_d)$	**−8.12**	**434.27**	**338.90**	**81.7**
Fg) Forward $(\bar{y}_r, V_g)$	**−7.90**	**433.71**	**328.46**	**81.6**
Bs) Best $(\bar{y}_r, V_s)$	-0.00	466.16	397.59	86.6
Bd) Best $(\bar{y}_r, V_d)$	**−7.90**	**434.54**	**336.88**	**81.5**
Bg) Best $(\bar{y}_r, V_g)$	**−7.60**	**433.26**	**326.05**	**81.6**
Fl) Fixed $(\bar{y}_r, V_s)$	0.45	490.49	461.86	89.0
SS) Saturated $(\bar{y}_r, V_s)$	−0.20	462.71	413.17	86.9
FR) Proc REG $(\bar{y}_r, V_s)$	−0.07	466.13	399.34	86.4
BNN) $t(100)$	**−5.78**	**395.25**	**323.12**	**86.5**

[a] Nominal 95% coverage.

4. SECOND SIMULATION STUDY

In the first simulation study, we show that the BNN model works well for the data sets with outliers. In this simulation study, we show that the BNN model works even better for the data sets without outliers. In this study, we also demonstrate how a cross-validation procedure can be applied to determine a setting for the parameters v, M and λ of the BNN model.

The simulation population comprises the records of the serious crimes of 141 large standard Metropolitan Statistical Areas (SMSAs) in the United States. A SMSA includes a city (or cities) of specified population size. The data generally pertains to the years 1976 and 1977, and is available in Neter, Kutner, Nachtsheim and Wasserman (1996). We consider the total number of serious crimes in 1977 as the survey variable (y) and the following 9 variables as potential auxiliary variables.

x_1	Land area (in square miles);
x_2	Estimated 1977 total population (in thousands);
x_3	Percent of 1976 SMSA population in central city or cities;
x_4	Percent of 1976 SMSA population 65 years old or older;
x_5	Number of professionally active nonfederal physicians as of December 31, 1977;
x_6	Total number of beds, cribs, and bassinets during 1977;
x_7	Percent of adult population (persons 25 years old or older) who completed 12 or more years of school, according to the 1970 Census of the Population;
x_8	Total number of persons in civilian labor force (persons 16 years old or older classified as employed or unemployed) in 1977 (in thousands);
x_9	Total current income received in 1976 by residents of the SMSA from all sources (in millions of dollars).

Table 4

Assessment of the influence of v, M and λ on BNN model size and prediction ability for the second explanatory set. For convenience of comparison, the results of the setting $v = 100$, $M = 8$ and $\lambda = 5$ were repeated in panels B and C of the table.

Experiment	v	M	λ	Size[a]	BIAS	MSE	AVMSE	Coverage[b] (%)
A	50	8	5	14.87	−9.30	394.11	270.09	82.5
	100	8	5	15.06	−5.78	395.25	323.12	86.5
	150	8	5	15.17	−4.38	412.56	346.75	87.1
B	100	6	5	13.90	−5.77	394.79	319.13	86.0
	100	8	5	15.06	−5.78	395.25	323.12	86.5
	100	10	5	16.05	−5.91	396.27	327.86	87.1
C	100	8	4	13.23	−5.62	397.65	323.68	86.4
	100	8	5	15.06	−5.78	395.25	323.12	86.5
	100	8	6	16.76	−5.78	396.45	321.98	86.6

[a] Size= $\sum_{k=1}^{1,000} \sum_{i=1}^{M} m(\Lambda_i)/M$ /1,000, where $m(\Lambda_i)$ is the number of connections of the neural network Λ_i.

[b] Nominal 95% coverage.

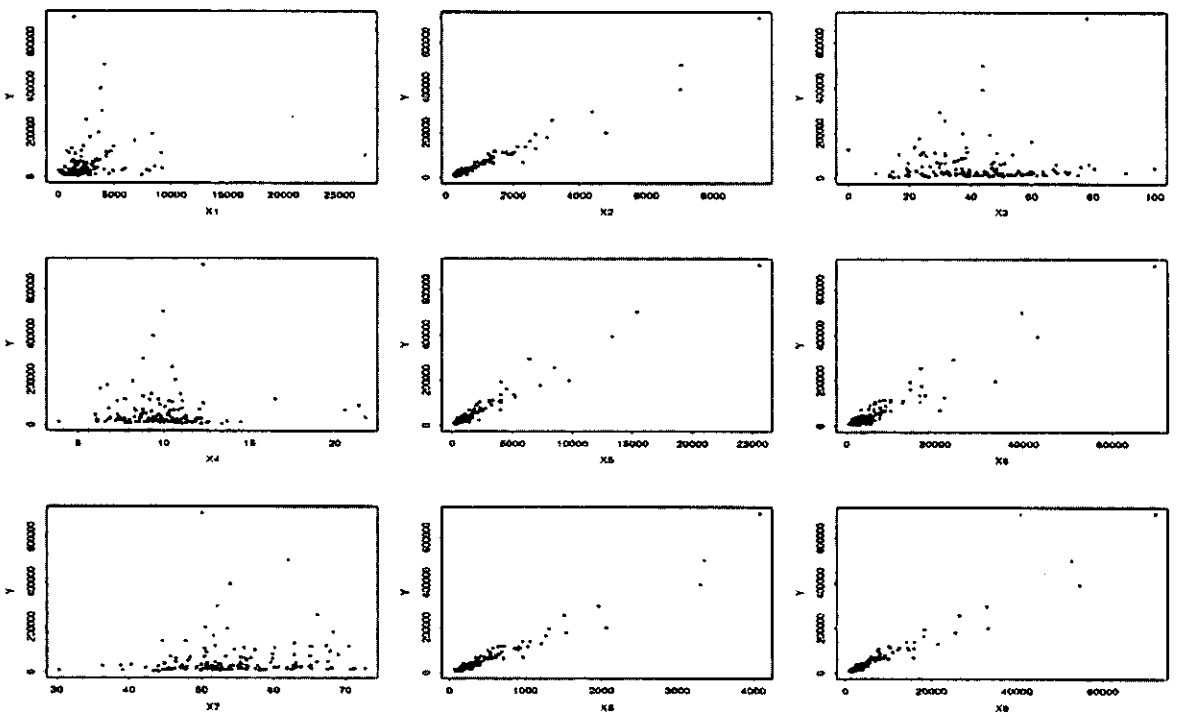


Figure 7: Scatter plots of the response variable y versus the auxiliary variables for the second simulation study.

Table 5
Cross-validation experiments for the SMSA example. For convenience of comparison, the results of the setting $v = 100$, $M = 3$ and $\lambda = 5$ were repeated in panels B and C.

Experiment	v	M	λ	Size	BIAS ($\times 10^3$)	MSE ($\times 10^6$)	AVMSE ($\times 10^6$)	Coverage[a] (%)
A	50	3	5	10.68	−0.472	4.78	4.19	91
	100	3	5	10.74	−0.527	5.04	4.24	92
	∞	3	5	10.74	−0.543	4.76	4.21	92
B	100	1	5	7.29	−0.466	4.63	3.66	89
	100	2	5	9.42	−0.500	4.61	3.91	90
	100	3	5	10.74	−0.527	5.04	4.24	92
	100	4	5	11.66	−0.480	4.74	4.47	91
C	100	3	4	9.56	−0.434	4.68	4.12	92
	100	3	5	10.74	−0.527	5.04	4.24	92
	100	3	6	11.82	−0.455	4.66	4.28	93

[a] Nominal 95% coverage.

Figure 7, the scatter plot of y versus the 9 auxiliary variables, suggests that a linear regression model may not be appropriate for the data set. There is a strong nonlinear relationship between y and x_1, x_3, x_4 and x_7. Also, the explanatory variables x_2, x_5, x_6, x_8 and x_9 are highly correlated. First, we demonstrate how a cross-validation procedure can be applied to determine the setting for the parameters v, M and λ of the BNN model. We treated the first 70 records as a small finite population, generated 100 sample replicates of size 50 from these 70 records by the method of simple random sampling without replacement, and then conducted the following experiments. In the first group of experiments, we fixed $M = 3$ and $\lambda = 5$, and varied the value of v, $v = 50$, 100 and ∞, where $v = \infty$ is just an indicator which indicates the normality assumption for the disturbance. Note $M = 3$ results in a full model of 43 connections, which has been large enough for the data set. In the second group of experiments, we fixed $v = 100$ and $\lambda = 5$, and varied the value of M, $M = 1, 2, 3, 4$. In the third group of experiments, we fixed $v = 100$ and $M = 3$, and varied the value of λ, $\lambda = 4, 5, 6$. For each setting, RJEMC was run as in the first simulation study. The computational results were summarized in Table 5. It shows that the performance of the BNN model is rather stable to the variation of the settings. It also suggests that the setting $v = 100$, $M = 3$ and $\lambda = 4$ probably be a good setting for this simulation population by a synthetical considerations on all values of BIAS, MSE, AVMSE and coverage probability.

In the further analysis, we generated 500 sample replicates of size 70 from all the 141 records by the method of simple random sampling without replacement. For each replicate, RJEMC was run as in the first simulation study. The computational results were summarized in Table 6. It shows that the BNN model also works well for this population. We also tried the other settings given in Table 5 for the 500 sample replicates. The computational results are all similar.

Table 6
Computational results for the second simulation study with $v = 100$, $M = 3$ and $\lambda = 4$

Size	BIAS ($\times 10^3$)	MSE ($\times 10^6$)	AVMSE ($\times 10^6$)	Coverage[a] (%)
9.20	−0.512	3.36	3.25	92.6

[a] Nominal 95% coverage.

5. DISCUSSION

In this article, we studied the use of Bayesian neural networks in finite population estimation. The numerical results show that it has made a significant improvement over the linear regression based methods. The improvement is not from Bayesian model averaging, but mainly from BNN models. We also applied the linear regression based Bayesian model averaging method (Liang, Truong and Wong 2001) to the same problem, and the improvement over Silva and Skinner (1997) is only marginal. Although our implementation for BNN models is not specific to finite populations, we do not think this is a shortcoming of our method. The generality of our method suggests its wide applications, for example, in nonlinear regression and nonlinear time series (the program is available by an request from the first author). Of course, a further research on how to use the known auxiliary variable information for a finite population in BNN training is also of interest.

APPENDIX

Before proving Theorem 2.1, we give one formula which will be used in the proof.

Formula 5.1 (*Laplace's method*)

$$\int b(\theta)\exp\{-nh(\theta)\}d\theta$$

$$= (2\pi/n)^{p/2}\left|\sum\right|^{1/2}\exp\{-nh(\hat{\theta})\}b(\hat{\theta})\{1+O(n^{-1})\}, \quad (24)$$

as $n \to \infty$, where $b(\cdot)$ is a general function which does not depend on n, $h(\theta)$ is a constant-order function of n as $n \to \infty$, p is the dimension of θ, $\hat{\theta}$ is the maximizer of $-h(\theta)$ and $\sum = (D^2 h(\hat{\theta}))^{-1}$ is the inverse of the negative Hessian matrix evaluated at $\hat{\theta}$.

For the general formulation of Laplace's method, see Kass and Vaidyanathan (1992).

Proof of Theorem 2.1

Proof: **Part (a)** By definition of expectation, $E_\pi |g(x_0, \theta_\Lambda)|^{2+\delta}$ can be written as

$$E_\pi \left| g(x_0, \theta_\Lambda)\right|^{2+\delta} = \sum_{k=0}^{K} P(\Lambda_k | D)$$

$$\int |g(x_0, \theta_\Lambda)|^{2+\delta} \pi(\theta_k | \Lambda_k, D) d\theta_k.$$

Following from the normality of the posterior distributions $\pi(\theta_k | \Lambda_k, D)$ (Walker 1969) and the fact that the activation function $\psi(\cdot)$ in (3) is bounded, we know (9) holds. Walker (1969) showed that the posterior distribution is Gaussian in the limit of infinite training data.

Part (b). For a given observation x_0, $E_\pi \hat{g}(x_0, \theta_\Lambda)$ can be written as

$$E_\pi = \hat{g}(x_0, \theta_\Lambda) = \frac{\sum_{\Lambda \in \Omega} P(\Lambda) \int \hat{g}(x_0, \theta_\Lambda) \exp\{-nh(\theta_\Lambda)\} \tilde{\pi}(\theta_\Lambda | \Lambda) d\theta_\Lambda}{\sum_{\Lambda \in \Omega} P(\Lambda) \int \exp\{-nh(\theta_\Lambda)\} \tilde{\pi}(\theta_\Lambda | \Lambda) d\theta_\Lambda} \quad (25)$$

where

$$\begin{aligned}\log \tilde{\pi}(\theta_\Lambda | \Lambda) = &-\log \sigma^2 - \frac{1}{2}\sum_{i=0}^{p} I_{\alpha_i}\left(\log \sigma_\alpha^2 + \frac{\alpha_i^2}{\sigma_\alpha^2}\right) \\ &- \frac{1}{2}\sum_{j=1}^{M} I_{\beta_j}\delta\left(\sum_{i=0}^{p} I_{\gamma_{ji}}\right)\left(\log \sigma_\beta^2 + \frac{\beta_j^2}{\sigma_\beta^2}\right) \\ &- \frac{1}{2}\sum_{j=1}^{M}\sum_{i=0}^{p} I_{\beta_j} I_{\gamma_{ji}}\left(\log \sigma_\gamma^2 + \frac{\gamma_{ji}^2}{\sigma_\gamma^2}\right) \\ &- \frac{m}{2}\log(2\pi) + m\log\lambda - \log(m!), \end{aligned} \quad (26)$$

and

$$\begin{aligned} h(\theta_\Lambda) &= \frac{1}{n}\left[\frac{n}{2}\log\sigma^2 + \frac{v+1}{2}\sum_{t=1}^{n}\log\left(1 + \frac{(y_t - \hat{f}(x_t))^2}{v\sigma^2}\right)\right] \\ &\approx \frac{1}{n}\left[\frac{n}{2}\log\sigma^2 + \frac{v+1}{2}\sum_{t=1}^{n}\frac{(y_t - \hat{f}(x_t))^2}{v\sigma^2}\right] \\ &\approx \frac{1}{2}\log\sigma^2 + \frac{v+1}{2v\sigma^2}E(y_t - \hat{g}(x_t, \theta_\Lambda))^2 \\ &= \frac{1}{2}\log\sigma^2 + \frac{v+1}{2v\sigma^2} \\ &\quad \left[E(y_t - g(x_t))^2 + (g(x_t) - \hat{g}(x_t, \theta_\Lambda))^2\right], \end{aligned} \quad (27)$$

where the first approximation follows from the Taylor expansion, $\log(1+z) \approx z$, when z lies in a neighbourhood of zero; and the second approximation follows from the weak law of large numbers by assuming that n is large. Note v is often set to a large number, say, a number greater than 30. In the first example of this paper, we set $v = 100$. The equation (27) implies that the minimum of $h(\boldsymbol{\theta}_\Lambda)$ is attained when $g(\boldsymbol{x}_i) = \hat{g}(\boldsymbol{x}_i, \boldsymbol{\theta}_\Lambda)$ holds, that is, $\hat{g}(\boldsymbol{x}_i, \boldsymbol{\theta}_\Lambda) = g(\boldsymbol{x}_i)$, where $\hat{\boldsymbol{\theta}}_\Lambda = \arg\min_{\boldsymbol{\theta}_\Lambda} h(\boldsymbol{\theta}_\Lambda)$.

By applying Laplace's method to the numerator of (25) with $b(\cdot) = \hat{g}(\boldsymbol{x}_0, \boldsymbol{\theta}_\Lambda)\tilde{\pi}(\boldsymbol{\theta}_\Lambda \mid D)$, we have

$$\begin{aligned}
&\sum_{\Lambda\in\Omega} P(\Lambda)\int \hat{g}(\boldsymbol{x}_0, \boldsymbol{\theta}_\Lambda)\exp\{-\tau H(\boldsymbol{\theta}_\Lambda)\}\tilde{\pi}(\boldsymbol{\theta}_\Lambda \mid \Lambda)d\boldsymbol{\theta}_\Lambda \\
&\approx \sum_{\Lambda\in\Omega} P(\Lambda)(2\pi/n)^{m/2}|\Sigma_\Lambda|^{1/2} \\
&\quad \exp\{-nh(\hat{\boldsymbol{\theta}}_\Lambda)\}\hat{g}(\boldsymbol{x}_0, \hat{\boldsymbol{\theta}}_\Lambda)\tilde{\pi}(\hat{\boldsymbol{\theta}}_\Lambda \mid D) \\
&\approx g(\boldsymbol{x}_0)\sum_{\Lambda\in\Omega} P(\Lambda)(2\pi/n)^{m/2}|\Sigma_\Lambda|^{1/2} \\
&\quad \exp\{-nh(\hat{\boldsymbol{\theta}}_\Lambda)\}\tilde{\pi}(\hat{\boldsymbol{\theta}}_\Lambda \mid D), \qquad (28)
\end{aligned}$$

where the first approximation follows from the Laplace formula (24), and the second approximation follows from the equality $\hat{g}(\boldsymbol{x}_i, \hat{\boldsymbol{\theta}}_\Lambda) = g(\boldsymbol{x}_i)$. Here we assume that the number of hidden units of each Λ is suffiently large such that $g(\cdot)$ can be approximated arbitrarily well by the network with properly adjusted weights. Otherwise, that term will take a small value and is negligible in the last approximation of (28).

Similarly, by applying the Laplace's method to the denominator of (25) with $b(\cdot) = \tilde{\pi}(\boldsymbol{\theta}_\Lambda \mid D)$, we have

$$\begin{aligned}
&\sum_{\Lambda\in\Omega} P(\Lambda)\int \exp\{-nh(\boldsymbol{\theta}_\Lambda)\}\tilde{\pi}(\boldsymbol{\theta}_\Lambda \mid \Lambda)d\boldsymbol{\theta}_\Lambda \\
&\approx \sum_{\Lambda\in\Omega} P(\Lambda)(2\pi/n)^{m/2}|\Sigma_\Lambda|^{1/2}\exp\{-nh(\hat{\boldsymbol{\theta}}_\Lambda)\}\tilde{\pi}(\hat{\boldsymbol{\theta}}_\Lambda \mid D). \qquad (29)
\end{aligned}$$

Following from (28), (29), and the approximation accuracy ($O(n^{-1})$) of Laplace's method, we have

$$E_\pi \hat{g}(\boldsymbol{x}_0, \boldsymbol{\theta}_{\Lambda_i}) \to g(\boldsymbol{x}_0), \qquad (30)$$

as $n \to \infty$. Following from (7), (9) and (30), we have

$$\frac{1}{M}\sum_{i=1}^{M} \hat{g}(\boldsymbol{x}_0, \boldsymbol{\theta}_{\Lambda_i}) \to g(x_0), \quad a.s.,$$

as $n \to \infty$ and $M \to \infty$.

Part (c). It follows from (8), (9), (30) and Slutsky's Theorem (Casella and Berger 2002). The proof is completed.

ACKNOWLEDGEMENTS

The authors would like to thank Chris Skinner for providing the test census data set, and thank the anonymous referees, the associate editor and editor Dr. M.P. Singh for their constructive comments which have led to a significant improvement of this paper.

REFERENCES

BANKIER, M.D. (1990). Two step generalized least squares estimation. Ottawa: Statistics Canada, Social Survey Methods Divison, Internal reports.

BARDSLEY, P., and CHAMBERS, R.L. (1984). Multipurpose estimation from unbalanced samples. *Applied Statistics*, 33, 290-299.

BILLINGSLEY, P. (1986). *Probability and Measure* (Second Edition). New York: John Wiley & Sons, Inc.

BREIDT, F.J., and OPSOMER, J.D. (2000). Local polynomial regression estimators in survey sampling. *Annals of Statistics*, 28, 1026-1053.

BUNTINE, W.L., and WEIGEND, A.S. (1991). Bayesian back-propagation. *Complex Systems*, 5, 603-643.

CASELLA, G., and BERGER, R.L. (2002). *Satistical Inference* (Second Edition). United States: Thompson Learning.

CHAMBERS, R.L., DORFMAN, A.H. and WEHRLY, T.E. (1993). Bias robust estimation in finite populations using nonparametric calibration. *Journal of the American Statistical Association*, 88, 268-277.

COCHRAN, W.G. (1977). *Sampling techniques* (3rd Ed.). New York: John Wiley & Sons, Inc.

CYBENKO, G. (1989). Approximations by superpositions of a sigmoidal function. *Mathematics of Control, Signals and Systems*, 2, 303-314.

DENG, L.Y., and WU, C.F.J. (1987). Estimation of variance of the regression estimator. *Journal of the American Statistical Association*, 82, 568-576.

DEVILLE, J.-C., and SÄRNDALL, C.-E. (1992). Calibration estimators in survey sampling. *Journal of the American Statistical Association*, 87, 376-382.

DORFMAN, A.H. (1992). Non-parametric regression for estimating totals in finite populations. In *Proceedings of the Section on Survey Research Methods*, American Statistical Association, Alexandria, VA. 622-625.

DUNSTAN, R., and CHAMBERS, R.L. (1986). Model-based confidence intervals in multipurpose surveys. *Applied Statistics*, 35, 276-280.

FIRTH, D., and BENNETT, K.E. (1998). Robust models in probability sampling. *Journal of the Royal Statistical Society B*, 60, 3-21.

FUNAHASHI, K. (1989). On the approximate realization of continuous mappings by neural networks. *Neural Networks*, 2, 183-192.

GELMAN, A., and RUBIN, D.B. (1992). Inference from iterative simulation using multiple sequences (with discussion). *Statistical Science*, 7, 457-472.

GREEN, P.J. (1995). Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika,* 82, 711-732.

HOETING, J.A., MADIGAN, D., RAFTERY, A.E. and VOLINSKY, C. (1999). Bayesian model averaging: a tutorial (with discussion). *Statistical Science,* 14, 382-417.

HOLMES, C.C., and MALLICK, B.K. (1998). Bayesian radial basis functions of variable dimension. *Neural Computation,* 10, 1217-1233.

HORNIK, K., STINCHCOMBE, M. and WHITE, H. (1989). Multilayer feedforward networks are universal approximators. *Neural Networks,* 2, 359-366.

KASS, R.E., and VAIDYANATHAN, S. (1992). Approximate Bayesian factor and orthogonal parameters, with applications to testing equality of two binomial proportions. *Journal of the Royal Statistical Society B,* 54, 129-144.

KUK, A.Y.C. (1993). A kernel method for estimating finite population distribution functions using auxiliary information. *Biometrika,* 80, 385-392.

KUK, A.Y.C., and WELSH, A.H. (2001). Robust estimation for finite populatons based on a working model. *Journal of the Royal Statistical Society B,* 63, 277-292.

LIANG, F., TRUONG, Y.K. and WONG, W.H. (2001). Automatic Bayesian model averaging for linear regression and applications in Bayesian curve fitting. *Statistica Sinica,* 11, 1005-1029.

LIANG, F., and WONG, W.H. (2001). Real parameter evolutionary Monte Carlo with applications in Bayesian mixture models. *Journal of the Amerian Statistical Association,* 96, 653-666.

MACKAY, D.J.C. (1992). A practical Bayesian framework for backprop networks. *Neural Computation,* 4, 448-472.

MADIGAN, D., and RAFTERY, A.E. (1994). Model selection and accounting for model uncertainty in graphical models using Occam's window. *Journal of the American Statistical Association,* 89, 1535-1546.

MARRS, A.D. (1998). An application of reversible-jump MCMC to multivariate spherical Gaussian mixtures. In *Advances in Neural Information Processing Systems 10.* San Mateo, CA: Morgan Kaufmann. 577-583.

MÜLLER, P., and INSUA, D.R. (1998). Issues in Bayesian analysis of neural network models. *Neural Computation,* 10, 749-770.

NEAL, R.M. (1996). *Bayesian Learning For Neural Networks.* New York: Springer-Verlag.

NETER, J., KUTNER, M.H., NACHTSHEIM, C.J. and WASSERMAN, W. (1996). *Applied Linear Statistical Models* (Fourth Edition). Chicago: Irwin.

ROBERTS, C.P., and CASELLA, G. (1999). *Monte Carlo Statistical Methods.* New York: Springer.

ROBERTS, G.O. (1996). Markov chain concepts related to sampling algorithms. In *Markov Chain Monte Carlo in Practice* (Eds. W.R. Gilks, S. Richardson and D.J. Spiegelhalter). London: Chapman & Hall/CRC. 45-57.

SÄRNDAL, C.-E., SWENSSON, B. and WRETMAN, J. (1989). The weighted residual technique for estimating the variance of the general regression estimator of the finite population total. *Biometrika,* 76, 527-537.

SILVA, P.L.D. (1996). Some asymptotic results on the mean squared error of the regression estimator under simple random sampling without replacement. Southampton: University of Southampton, Center for Survey Data Analysis Technical Report 6-2.

SILVA, P.L.D., and SKINNER, C. (1997). Variable selection for regression estimation in finite populations. *Survey Methodology,* 23, 23-32.

THEBERGE, A. (1999). Extensions of calibration estimators in survey sampling. *Journal American Statistical Association,* 94, 635-644.

TIERNEY, L. (1994). Markov chains for exploring posterior distributions (with discussion). *Annals of Statistics,* 22, 1701-1786.

VALLIANT, R., DORFMAN, A.H. and ROYALL, R.M. (2000). *Finite Population Sampling and Inference: A Prediction Approach.* New York: John Wiley & Sons, Inc.

WALKER, A.M. (1969). On the asymptotic behaviour of posterior distributions. *Journal Royal Statistics Society, B,* 31, 80-88.

WEIGEND, A.S., HUBERMAN, B.A. and RUMELHART, D.E. (1990). Predicting the future: A connectionist approach. *Int. J. Neural Syst. 1,* 193-209.

WU, C., and SITTER, R.R. (2001). A model-calibration approach to using complete auxiliary information from survey data. *Journal American Statistical Association,* 96, 185-193.

Survey Methodology, December 2004
Vol. 30, No. 2, pp. 235-242
Statistics Canada

Simultaneous Use of Multiple Imputation for Missing Data and Disclosure Limitation

JEROME P. REITER [1]

ABSTRACT

Several statistical agencies use, or are considering the use of, multiple imputation to limit the risk of disclosing respondents' identities or sensitive attributes in public use data files. For example, agencies can release partially synthetic datasets, comprising the units originally surveyed with some collected values, such as sensitive values at high risk of disclosure or values of key identifiers, replaced with multiple imputations. This article presents an approach for generating multiply-imputed, partially synthetic datasets that simultaneously handles disclosure limitation and missing data. The basic idea is to fill in the missing data first to generate m completed datasets, then replace sensitive or identifying values in each completed dataset with r imputed values. This article also develops methods for obtaining valid inferences from such multiply-imputed datasets. New rules for combining the multiple point and variance estimates are needed because the double duty of multiple imputation introduces two sources of variability into point estimates, which existing methods for obtaining inferences from multiply-imputed datasets do not measure accurately. A reference t-distribution appropriate for inferences when m and r are moderate is derived using moment matching and Taylor series approximations.

KEY WORDS: Confidentiality; Missing data; Public use data; Survey; Synthetic data.

1. INTRODUCTION

Many statistical agencies disseminate microdata, *i.e.*, data on individual units, in public use files. These agencies strive to release files that are (i) safe from attacks by ill-intentioned data users seeking to learn respondents' identities or attributes, (ii) informative for a wide range of statistical analyses, and (iii) easy for users to analyze with standard statistical methods. Doing this well is a difficult task. The proliferation of publicly available databases, and improvements in record linkage technologies, have made disclosures a serious threat, to the point where most statistical agencies alter microdata before release. For example, agencies globally recode variables, such as releasing ages in five year intervals or top-coding incomes above $100,000 as "$100,000 or more" (Willenborg and de Waal 2001); they swap data values for randomly selected units (Dalenius and Reiss 1982); or, they add random noise to continuous data values (Fuller 1993). Inevitably, these strategies reduce the utility of the released data, making some analyses impossible and distorting the results of others. They also complicate analyses for users. To analyze properly perturbed data, users should apply the likelihood-based methods described by Little (1993) or the measurement error models described by Fuller (1993). These are difficult to use for non-standard estimands and may require analysts to learn new statistical methods and specialized software programs.

An alternative approach to disseminating public use data was suggested by Rubin (1993): release multiply-imputed, synthetic datasets. Specifically, he proposed that agencies (i) randomly and independently sample units from the sampling frame to comprise each synthetic data set, (ii) impute unknown data values for units in the synthetic samples using models fit with the original survey data, and (iii) release multiple versions of these datasets to the public. These are called *fully synthetic* data sets. Releasing fully synthetic data can protect confidentiality, since identification of units and their sensitive data is nearly impossible when the values in the released data are not actual, collected values. Furthermore, with appropriate synthetic data generation and the inferential methods developed by Raghunathan, Reiter and Rubin (2003) and Reiter (2004b), it can allow data users to make valid inferences for a variety of estimands using standard, complete-data statistical methods and software. Other attractive features of fully synthetic data are described by Rubin (1993), Little (1993), Fienberg, Makov and Steele (1998), Raghunathan *et al.* (2003), and Reiter (2002, 2004a).

No statistical agencies have released fully synthetic datasets as of this writing, but some have adopted a variant of the multiple imputation approach suggested by Little (1993): release datasets comprising the units originally surveyed with some collected values, such as sensitive values at high risk of disclosure or values of key identifiers, replaced with multiple imputations. These are called *partially synthetic* datasets. For example, the U.S. Federal Reserve Board protects data in the U.S. Survey of Consumer Finances by replacing monetary values at high

[1] Jerome P. Reiter, Institute of Statistics and Decision Sciences, Duke University, Box 90251, Durham, NC 27708-0251. E-mail: jerry@stat.duke.edu.

disclosure risk with multiple imputations, releasing a mixture of these imputed values and the unreplaced, collected values (Kennickell 1997). The U.S. Bureau of the Census and Abowd and Woodcock (2001) protect data in longitudinal, linked data sets by replacing all values of some sensitive variables with multiple imputations and leaving other variables at their actual values. Liu and Little (2002) present a general algorithm, named SMIKe, for simulating multiple values of key identifiers for selected units.

All these partially synthetic approaches are appealing because they promise to maintain the primary benefits of fully synthetic data – protecting confidentiality while allowing users to make inferences without learning complicated statistical methods or software – with decreased sensitivity to the specification of imputation models (Reiter 2003). Valid inferences from partially synthetic datasets can be obtained using the methods developed by Reiter (2003, 2004b), whose rules for combining point and variance estimates again differ from those of Rubin (1987) and also from those of Raghunathan *et al.* (2003).

The existing theory and methods for partially synthetic data do not deal explicitly with an important practical complication: in most large surveys, there are units that fail to respond to some or all items of the survey. This article presents a multiple imputation approach that handles simultaneously missing data and disclosure limitation. The approach involves two steps. First, the agency uses multiple imputation to fill in the missing data, generating m multiply-imputed datasets. Second, the agency replaces the values at risk of disclosure in each imputed dataset with r multiple imputations, ultimately releasing mr multiply-imputed datasets. This double-duty of multiple imputation requires new methods for obtaining valid inferences from the multiply-imputed datasets, which are derived here.

The paper is organized as follows. Section 2 reviews multiple imputation for missing and partially synthetic data. Section 3 presents the new methods for generating partially synthetic data and obtaining valid inferences when some survey data are missing. Section 4 shows a derivation of these methods from a Bayesian perspective, and it discusses conditions under which the resulting inferences should be valid from a frequentist perspective. Section 5 concludes with a discussion of the challenges to implementing this multiple imputation approach on genuine data, with an aim towards stimulating future research.

2. REVIEW OF MULTIPLE IMPUTATION INFERENCES

To describe multiple imputation, we use the notation of Rubin (1987). For a finite population of size N, let $I_j = 1$ if unit j is selected in the survey, and $I_j = 0$ otherwise, where $j = 1, 2, ..., N$. Let $I = (I_1, .., I_N)$. Let R_j be a $p \times 1$ vector of response indicators, where $R_{jk} = 1$ if the response for unit j to survey item k is recorded, and $R_{jk} = 0$ otherwise. Let $R = (R_1, ..., R_N)$. Let Y be the $N \times p$ matrix of survey data for all units in the population. Let $Y_{\text{inc}} = (Y_{\text{obs}}, Y_{\text{mis}})$ be the $n \times p$ matrix of survey data for the n units with $I_j = 1$; Y_{obs} is the portion of Y_{inc} that is observed, and Y_{mis} is the portion of Y_{inc} that is missing due to nonresponse. Let X be the $N \times d$ matrix of design variables for all N units in the population, *e.g.*, stratum or cluster indicators or size measures. We assume that such design information is known approximately for all population units, for example from census records or the sampling frame(s). Finally, we write the observed data as $D = (X, Y_{\text{obs}}, I, R)$.

2.1 Multiple Imputation for Missing Data

The agency fills in values for Y_{mis} with draws from the Bayesian posterior predictive distribution of $(Y_{\text{mis}} \mid D)$, or approximations of that distribution such as those of Raghunathan, Lepkowski, Van Hoewyk and Solenberger (2001). These draws are repeated independently $l = 1, \ldots, m$ times to obtain m completed data sets, $D^{(l)} = (D, Y_{\text{mis}}^{(l)})$. Multiple rather than single imputations are used so that analysts can estimate the variability due to imputing missing data.

In each imputed data set $D^{(l)}$, the analyst estimates the population quantity of interest, Q, using some estimator q, and estimates the variance of q with some estimator u. We assume that the analyst specifies q and u by acting as if each $D^{(l)}$ was in fact collected data from a random sample of (X, Y) based on the original sampling design I, *i.e.*, q and u are complete-data estimators.

For $l = 1, \ldots, m$, let $q^{(l)}$ and $u^{(l)}$ be respectively the values of q and u in data set $D^{(l)}$. Under assumptions described in Rubin (1987), the analyst can obtain valid inferences for scalar Q by combining the $q^{(l)}$ and $u^{(l)}$. Specifically, the following quantities are needed for inferences:

$$\bar{q}_m = \sum_{l=1}^{m} q^{(l)}/m \tag{1}$$

$$b_m = \sum_{l=1}^{m} (q^{(l)} - \bar{q}_m)^2/(m-1) \tag{2}$$

$$\bar{u}_m = \sum_{l=1}^{m} u^{(l)}/m. \tag{3}$$

The analyst then can use $\bar{q}_m$ to estimate Q and $T_m = (1 + 1/m)b_m + \bar{u}_m$ to estimate the variance of $\bar{q}_m$.

Inferences can be based on t-distributions with degrees of freedom $v_m = (m-1)(1+\bar{u}_m/((1+1/m)b_m))^2$.

2.2 Multiple Imputation for Partially Synthetic Data when $Y_{inc} = Y_{obs}$

Assuming no missing data, *i.e.*, $Y_{inc} = Y_{obs}$, the agency constructs partially synthetic datasets by replacing selected values from the observed data with imputations. Let $Z_j = 1$ if unit j is selected to have any of its observed data replaced with synthetic values, and let $Z_j = 0$ for those units with all data left unchanged. Let $Z = (Z_1, ..., Z_n)$. Let $Y_{rep,i}$ be all the imputed (replaced) values in the i^{th} synthetic data set, and let Y_{nrep} be all unchanged (unreplaced) values of Y_{obs}. The $Y_{rep,i}$ are assumed to be generated from the posterior predictive distribution of $(Y_{rep,i} \mid D, Z)$, or a close approximation of it. The values in Y_{nrep} are the same in all synthetic data sets. Each synthetic data set, d_i, then comprises $(X, Y_{rep,i}, Y_{nrep}, I, Z)$. Imputations are made independently $i = 1, ..., r$ times to yield r different partially synthetic data sets, which are released to the public. Once again, multiple imputations enable analysts to account for variability due to imputation.

The values in Z can and frequently will depend on the values in D. For example, the agency may simulate sensitive variables or identifiers only for units in the sample with rare combinations of identifiers; or, the imputer may replace only incomes above $100,000 with imputed values. To avoid bias, the imputations should be drawn from the posterior predictive distribution of Y for those units with $Z_j = 1$. Reiter (2003) illustrates the problems that can arise when imputations are not conditional on Z.

Inferences from partially synthetic datasets are based on quantities defined in Equations (1)–(3). As shown by Reiter (2003), under certain conditions the analyst can use $\bar{q}_r$ to estimate Q and $T_p = b_r/r + \bar{u}_r$ to estimate the variance of $\bar{q}_r$. Inferences for scalar Q can be based on t-distributions with degrees of freedom $v_p = (r-1)(1+\bar{u}_r/(b_r/r))^2$.

3. PARTIALLY SYNTHETIC DATA WHEN $Y_{inc} \neq Y_{obs}$

When some data are missing, it seems logical to impute the missing and partially synthetic data simultaneously. However, imputing Y_{mis} and Y_{rep} from the same posterior predictive distribution can result in improper imputations. For an illustrative example, suppose univariate data from a normal distribution have some values missing completely at random (Rubin 1976). Further, suppose the agency seeks to replace all values larger than some threshold with imputations. The imputations for missing data can be based on a normal distribution fit using all of Y_{obs}. However, the imputations for replacements must be based on a posterior distribution that conditions on values being larger than the threshold. Drawing Y_{mis} and Y_{rep} from the same distribution will result in biased inferences.

Imputing the Y_{mis} and Y_{rep} separately generates two sources of variability, in addition to the sampling variability in D, that the user must account for to obtain valid inferences. Neither T_m nor T_p correctly estimate the total variation introduced by the dual use of multiple imputation. The bias of each can be illustrated with two simple examples. Suppose only one value needs replacement, but there are hundreds of missing values to be imputed. Intuitively, the variance of the point estimator of Q should be well approximated by T_m. and T_p should underestimate the variance, as it is missing a b_m. On the other hand, suppose only one value is missing, but there are hundreds of values to be replaced. The variance should be well approximated by T_p, and T_m should overestimate the variance, as it includes an extra b_m.

To allow users to estimate the total variability correctly, agencies can employ a three-step procedure for generating imputations. First, the agency fills in Y_{mis} with draws from the posterior distribution for $(Y_{mis} \mid D)$, resulting in m completed datasets, $D^{(1)}, ..., D^{(m)}$. Then, in each $D^{(l)}$, the agency selects the units whose values are to be replaced, *i.e.*, whose $Z_j^{(l)} = 1$. In many cases, the agency will impute values for the same units in all $D^{(l)}$ to avoid releasing any genuine, sensitive values for the selected units. We assume this is the case throughout and therefore drop the superscript l from Z. Third, in each $D^{(l)}$, the agency imputes values $Y_{rep,i}^{(l)}$ for those units with $Z_j = 1$, using the posterior distribution for $(Y_{rep} \mid D^{(l)}, Z)$. This is repeated independently $i = 1, ..., r$ times for $l = 1, ..., m$, so that a total of $M = mr$ datasets are generated. Each dataset, $d_i^{(l)} = (X, Y_{nrep}, Y_{mis}^{(l)}, Y_{rep,i}^{(l)}, I, R, Z)$, includes a label indicating the l of the $D^{(l)}$ from which it was drawn. These M datasets are released to the public. Releasing such nested, multiply-imputed datasets also has been proposed for handling missing data outside of the disclosure limitation context (Shen 2000; Rubin 2003).

Analysts can obtain valid inferences from these released datasets by combining inferences from the individual datasets. As before, let q be the analyst's estimator of Q, and let u be the analyst's estimator of the variance of q. We assume the analyst specifies q and u by acting as if each $d_i^{(l)}$ was in fact collected data from a random sample of (X, Y) based on the original sampling design I. For $l = 1, ..., m$ and $i = 1, ..., r$, let $q_i^{(l)}$ and $u_i^{(l)}$ be respectively the values of q and u in data set $d_i^{(l)}$. The following quantities are needed for inferences about scalar Q:

$$\bar{q}_M = \sum_{l=1}^{m}\sum_{i=1}^{r} q_i^{(l)}/(mr) = \sum_{l=1}^{m} \bar{q}^{(l)}/m \qquad (4)$$

$$\bar{b}_M = \sum_{l=1}^{m}\sum_{i=1}^{r}(q_i^{(l)} - \bar{q}^{(l)})^2/m(r-1)$$
$$= \sum_{l=1}^{m} b^{(l)}/m \quad (5)$$

$$B_M = \sum_{l=1}^{m}(\bar{q}^{(l)} - \bar{q}_M)^2/(m-1) \quad (6)$$

$$\bar{u}_M = \sum_{l=1}^{m}\sum_{i=1}^{r} u_i^{(l)}/(mr). \quad (7)$$

The $\bar{q}^{(l)}$ is the average of the point estimates in each group of datasets indexed by l, and the $\bar{q}_M$ is the average of these averages across l. The $b^{(l)}$ is the variance of the point estimates for each group of datasets indexed by l, and the $\bar{b}_M$ is average of these variances. The B_M is the variance of the $\bar{q}^{(l)}$ across synthetic datasets. The $\bar{u}_M$ is the average of the estimated variances of q across all synthetic datasets.

Under conditions described in section 4, the analyst can use $\bar{q}_M$ to estimate Q. An estimate of the variance of $\bar{q}_M$ is:

$$T_M = (1+1/m)B_M - \bar{b}_M/r + \bar{u}_M. \quad (8)$$

When n, m, and r are large, inferences can be based on the normal distribution, $(Q-\bar{q}_M) \sim N(0,T_M)$. When m and r are moderate, inferences can be based on the t-distribution, $(Q-\bar{q}_M) \sim t_{v_M}(0,T_M)$, with degrees of freedom

$$v_M = \left(\frac{((1+1/m)B_M)^2}{(m-1)T_M^2} + \frac{(\bar{b}_M/r)^2}{m(r-1)T_M^2}\right)^{-1}. \quad (9)$$

The behavior of T_M and v_M in special cases is instructive. When r is very large, $T_M \approx T_m$. This is because the $\bar{q}^{(l)} \approx q^{(l)}$, so that we obtain the results from analyzing the $D^{(l)}$. When the fraction of replaced values is small relative to the fraction of missing values, the $\bar{b}_M$ is small relative to B_M, so that once again $T_M \approx T_m$. In both these cases, the v_M approximately equals v_m, which is Rubin's (1987) degrees of freedom when imputing missing data only. When the fraction of missing values is small relative to the fraction of replaced values, the $B_M \approx \bar{b}_M/r$, so that T_M is approximately equal to T_p with M released datasets.

4. JUSTIFICATION OF NEW COMBINING RULES

This section presents a Bayesian derivation of the inferences described in section 3 and describes conditions under which these inferences are valid from a frequentist perspective. These results make use of the theory developed in Rubin (1987) and Reiter (2003). For the Bayesian derivation, we assume that the analyst and imputer use the same models.

Let $D^m = \{D^{(l)} : l = 1,...,m\}$ be the collection of all multiply-imputed datasets before any observed values are replaced. For each $D^{(l)}$, let $q^{(l)}$ and $u^{(l)}$ be the posterior mean and variance of Q. As in Rubin (1987, Chapter 3), let B_∞ be the variance of the $q^{(l)}$ obtained when $m = \infty$.

Let $d^M = \{d_i^{(l)} : i = 1,...,r; l = 1,...,m\}$ be the collection of all released synthetic datasets. For each $d_i^{(l)}$, let $q_i^{(l)}$ be the posterior mean of $q^{(l)}$. For each l, let $B^{(l)}$ be the variance of the $q_i^{(l)}$ obtained when $r = \infty$. Lastly, let B be the average of the $B^{(l)}$ obtained when $m = \infty$.

Using these quantities, the posterior distribution for $(Q \mid d^M)$ can be decomposed as

$$f(Q \mid d^M) = \int f(Q \mid d^M, D^m, B_\infty, B)$$
$$f(D^m, B_\infty \mid d^M, B)$$
$$f(B \mid d^M) dD^m dB_\infty dB. \quad (10)$$

The integration is over the distributions of the values in D that are missing and the values in each $D^{(l)}$ that are replaced with imputations; the observed, unaltered values remain fixed. We assume standard Bayesian asymptotics hold, so that complete-data inferences for Q can be based on normal distributions.

4.1 Evaluating $f(Q \mid d^M, D^m, B_\infty, B)$

Given D^m, the synthetic data are irrelevant, so that $f(Q \mid d^M, D^m, B_\infty, B) = f(Q \mid D^m, B_\infty)$. This is the posterior distribution of Q for multiple imputation for missing data, conditional on B_∞. As shown by Rubin (1987), this posterior distribution is approximately

$$(Q \mid D^m, B_\infty) \sim N(\bar{q}_m, (1+1/m)B_\infty + \bar{u}_m) \quad (11)$$

where $\bar{q}_m$ and $\bar{u}_m$ are defined as in (1) and (3). In multiple imputation for missing data, we integrate (11) over the posterior distribution of $(B_\infty \mid D^m)$. This is not done here, since we integrate over $(B_\infty \mid d^M)$.

4.2 Evaluating $f(D^m, B_\infty \mid d^M, B) f(B \mid d^M)$

Since the distribution for Q in (11) relies only on $\bar{q}_m$, $\bar{u}_m$, and B_∞, it is sufficient for $f(D^m, B_\infty \mid d^M, B)$ to determine

$$f(\bar{q}_m, \bar{u}_m, B_\infty \mid d^M, B) =$$
$$f(\bar{q}_m, \bar{u}_m \mid d^M, B_\infty, B) f(B_\infty \mid d^M, B).$$

Following Reiter (2003), we first assume replacement imputations are made so that, for all i, the sampling distributions of each $q_i^{(l)}$ and $u_i^{(l)}$ are,

$$(q_i^{(l)} \mid D^{(l)}, B^{(l)}) \sim \mathrm{N}(q^{(l)}, B^{(l)}) \quad (12)$$

$$(u_i^{(l)} \mid D^{(l)}, B^{(l)}) \sim (u^{(l)}, << B^{(l)}). \quad (13)$$

Here, the notation $F \sim (G, <<H)$ means that the random variable F has a distribution with expectation of G and variability much less than H. In actuality, $u_i^{(l)}$ is typically centered at a value larger than $u^{(l)}$, since synthetic data incorporate uncertainty due to drawing values of the parameters. For large sample sizes n, this bias should be minimal. The assumption that $E(q_i^{(l)} \mid D^{(l)}, B^{(l)}) = q^{(l)}$ and the normality assumption should be reasonable when the imputations are drawn from correct posterior predictive distributions, $f(Y_{\text{rep}} \mid D^{(l)}, Z)$, and the usual asymptotics hold.

Assuming flat priors for all $q^{(l)}$ and $v^{(l)}$, standard Bayesian theory implies that

$$\left(q^{(l)} \middle| d^M, B^{(l)}\right) \sim \mathrm{N}\left(\bar{q}^{(l)}, B^{(l)}/r\right) \quad (14)$$

$$\left(u^{(l)} \middle| d^M, B^{(l)}\right) \sim \left(\bar{u}^{(l)}, << B^{(l)}/r\right) \quad (15)$$

$$\left(\frac{(r-1)b^{(l)}}{B^{(l)}} \middle| d^M, B^{(l)}\right) \sim \chi^2_{r-1} \quad (16)$$

where $b^{(l)}$ is defined in (5). We next assume that $B^{(l)} = B$ for all l. This should be reasonable, since the variability in posterior variances tends to be of smaller order than the variability of posterior means. Averaging across l, we obtain

$$\left(\bar{q}_m \middle| d^M, B\right) \sim \mathrm{N}\left(\bar{q}_M, B/rm\right) \quad (17)$$

$$\left(\bar{u}_m \middle| d^M, B\right) \sim \left(\bar{u}_M, << B/rm\right) \quad (18)$$

where $\bar{q}_M$ is defined in (4) and $\bar{u}_M$ is defined in (7). The posterior distribution of $(B_\infty \mid d^M, B)$ is

$$\left(\frac{(m-1)B_M}{B_\infty + B/r} \middle| d^M, B\right) \sim \chi^2_{m-1} \quad (19)$$

where B_M is defined in (6).

Finally, the posterior distribution of $(B \mid d^M)$ is

$$\left(\frac{m(r-1)\bar{b}_M}{B} \middle| d^M\right) \sim \chi^2_{m(r-1)} \quad (20)$$

where $\bar{b}_M$ is defined in (5).

4.3 Evaluating $f(Q \mid d^M)$

We need to integrate the product of (11) and (17) with respect to the distributions in (19) and (20). This can be done by numerical integration, but it is desirable to have simpler approximations for users.

For large m and r, we can replace the terms in the variance with their approximate expectations: the $B_\infty \approx B_M - B/r$, and the $B \approx \bar{b}_M$. Hence, for large m and r, the posterior distribution of Q is approximately:

$$\begin{aligned}
&\left(Q \middle| d^M\right) \\
&\sim N(\bar{q}_M, (1+1/m)(B_M - \bar{b}_M/r) + \bar{b}_M/mr + \bar{u}_M) \\
&= N\left(\bar{q}_M, (1+1/m)B_M - \bar{b}_M/r + \bar{u}_M\right) \\
&= N(\bar{q}_M, T_M). \quad (21)
\end{aligned}$$

When m and r are moderately sized, the normal distribution may not be a good approximation. To derive an approximate reference t-distribution, we use the strategies of Rubin (1987) and Barnard and Rubin (1999). That is, we assume that for some degrees of freedom v_M to be estimated,

$$\left(\frac{v_M T_M}{\bar{u}_M + (1+1/m)B_\infty + B/mr} \middle| d^M\right) \sim \chi^2_{v_M} \quad (22)$$

so that we can use a t-distribution with v_M degrees of freedom for inferences about Q. We approximate v_M by matching the first two moments of (22) to those of a chi-squared distribution. The details showing that v_M is approximated by the expression in (9) are provided in the appendix.

The inferences based on (4) – (9) have valid frequentist properties under certain conditions. First, the analyst must use randomization-valid estimators, q and u. That is, when q and u are applied on D to get q_{obs} and u_{obs}, the $(q_{\text{obs}} \mid X, Y) \sim \mathrm{N}(Q, U)$ and $(u_{\text{obs}} \mid X, Y) \sim (U, << U)$, where the relevant distribution is that of I. Second, the imputations for missing data must be proper in the sense of Rubin (1987, Chapter 4). Essentially, this requires that inferences from the imputations for missing data be randomization-valid for q_{obs} and u_{obs}, under the posited non-response mechanism. Third, the imputations for partially synthetic data must be synthetically proper in the sense of Reiter (2003). This requires that the inferences from the replacement imputations associated with each $D^{(l)}$ be randomization valid for the $q^{(l)}$ and $u^{(l)}$.

In general, it is difficult to verify that imputations for missing data are proper in complex samples (Binder and Sun 1996). They may be proper for some analyses but not for others. As a result, some confidence intervals centered on unbiased estimators may not have nominal coverage rates; see Meng (1994) for a discussion of this issue. These difficulties exist for the multiple imputation approach used here, and indeed may be compounded because of the additional imputation of synthetic data.

5. CONCLUDING REMARKS

There are many challenges to using partially synthetic data approaches for disclosure limitation. Most important, agencies must decide which values to replace with imputations. General candidates for replacement include the values of identifying characteristics for units that are at high risk of identification, such as sample uniques and duplicates, and the values of sensitive variables in the tails of distributions. Confidentiality can be protected further by, in addition, replacing values at low disclosure risk (Liu and Little 2002). This increases the variation in the replacement imputations, and it obscures any information that can be gained just from knowing which data were replaced. As with any disclosure limitation method (Duncan, Keller-McNulty and Stokes 2001), these decisions should consider tradeoffs between disclosure risk and data utility. Guidance on selecting values for replacement is a high priority for research in this area.

There remain disclosure risks in partially synthetic data no matter which values are replaced. Users can utilize the released, unaltered values to facilitate disclosure attacks, for example via matching to external databases, or they may be able to estimate actual values of Y_{obs} from the synthetic data with reasonable accuracy. For instance, if all people in a certain demographic group have the same, or even nearly the same, value of an outcome variable, the imputation models likely will generate that value for imputations. Imputers may need to coarsen the imputations for such people. As another example, when users know that a certain record has the largest value of some Y_{obs}, that record can be identified when its value is not replaced.

On the data utility side, the main challenge is specifying imputation models, both for the missing and replaced data, that give valid results. For missing data, it is well known that implausible imputation models can produce invalid inferences, although this is less problematic when imputing relatively small fractions of missing data (Rubin 1987; Meng 1994). There is an analogous issue for partially synthetic data. When large fractions of data are replaced, for example entire variables, analyses involving the replaced values reflect primarily the distributional assumptions implicit in the imputation models. When these assumptions are implausible, the resulting analyses can be invalid. Again, this is less problematic when only small fractions of values are replaced, as might be expected in many applications of the partially synthetic approach.

Certain data characteristics can be especially challenging to handle with partially synthetic data. For example, it may be desirable to replace extreme values in skewed distributions, such as very large incomes. Information about the tails of these distributions may be limited, making it difficult to draw reasonable replacements while protecting confidentiality. As another example, randomly drawn imputations for highly structured data may be implausible, for instance unlikely combinations of family members' ages or marital statuses. These difficulties, coupled with the general limitations of inferences based on imputations, point to an important issue for research: developing and evaluating methods for generating partially synthetic data, including semi-parametric and non-parametric approaches.

We note that building the synthetic data models is generally an easier task than building the missing data models. Agencies can compare the distributions of the synthetic data to those of the observed data being replaced. When the synthetic distributions are too dissimilar from the observed ones, the imputation models can be adjusted. There usually is no such check for the missing data models.

It is, of course, impossible for agencies to anticipate every possible use of the released data, and hence impossible to generate models that provide valid results for every analysis. A more modest and attainable goal is to enable analysts to obtain valid inferences using standard methods and software for a wide range of standard analyses, such as some linear and logistic regressions. Agencies therefore should provide information that helps analysts decide what inferences can be supported by the released data. For example, agencies can include descriptions of the imputation models as attachments to public releases of data. Users whose analyses are not supported by the data may have to apply for special access to the observed data. Agencies also need to provide documentation for how to use the nested data sets. Rules for combining point estimates from the multiple data sets are simple enough to be added to standard statistical software packages, as has been done already for Rubin's (1987) rules in SAS, Stata, and S-Plus.

As constructed, the multiple imputation approach does not calibrate to published totals. This could make some users unhappy with or distrust the released data. It is not clear how to adapt the method – or, for that matter, many other disclosure limitation techniques that alter the original data – for calibration.

Missing data and disclosure risk are major issues confronting organizations releasing data to the public. The multiple imputation approach presented here is suited to handle both simultaneously, providing users with rectangular completed datasets that can be analyzed with standard statistical methods and software. There are challenges to implementing this approach in genuine applications, but, as noted by Rubin (1993) in his initial proposal, the potential payoffs of this use of multiple imputation are high. The next item on the research agenda is to investigate how well the theory works in practice, including comparisons of this approach with other disclosure limitation methods. These comparisons should

focus on measures of disclosure risks, obtained by simulating intruder behavior, and on measures of data utility for estimands of interest to users, including properties of point and interval estimates.

APPENDIX: DERIVATION OF APPROXIMATE DEGREES OF FREEDOM

Inferences from datasets with multiple imputations for both missing data and partially synthetic replacements are made using a t-distribution. A key step is to approximate the distribution of

$$\left(\frac{v_M T_M}{\bar{u}_M + (1+1/m) B_\infty + B/mr} \middle| \, d^M \right) \tag{23}$$

as a chi-squared distribution with v_M degrees of freedom. The v_M is determined by matching the mean and variance of the inverted χ^2 distribution to the mean and variance of (23).

Let $\alpha = (B_\infty + B/r)/B_M$, and let $\gamma = B/\bar{b}_M$. Then, $(\alpha^{-1} \mid d^M, B)$ and $(\gamma^{-1} \mid d^M)$ have mean square distributions with degrees of freedom $m-1$ and $m(r-1)$, respectively. Let $f = (1+1/m)B_M / \bar{u}_M$, and let $g = (1/r)\bar{b}_M / \bar{u}_M$. We can write (23) as

$$\frac{T_M}{\bar{u}_M + (1+1/m) B_\infty + B/mr} = \frac{\bar{u}_M (1+f-g)}{\bar{u}_M (1+\alpha f - \gamma g)}. \tag{24}$$

To match moments, we need to approximate the expectation and variance of (24).

For the expectation, we use the fact that

$$E\left(\frac{1+f-g}{1+\alpha f - \gamma g} \middle| d^M \right) = E\left(E\left(\frac{1+f-g}{1+\alpha f - \gamma g} \middle| d^M, B \right) \middle| \, d^M \right). \tag{25}$$

We approximate these expectations using first order Taylor series expansion in α^{-1} and γ^{-1} around their expectations, which equal one. As a result,

$$E\left(E\left(\frac{1+f-g}{1+\alpha f - \gamma g} \middle| d^M, B \right) \middle| \, d^M \right) \approx E\left(\frac{1+f-g}{1+f-\gamma g} \middle| \, d^m \right) \approx 1. \tag{26}$$

For the variance, we use the conditional variance representation

$$E\left(\mathrm{Var}\left(\frac{1+f-g}{1+\alpha f - \gamma g} \middle| d^M, B \right) \middle| d^M \right) + \mathrm{Var}\left(E\left(\frac{1+f-g}{1+\alpha f - \gamma g} \middle| d^M, B \right) \middle| d^M \right). \tag{27}$$

For the interior variance and expectation, we use a first order Taylor series expansion in α^{-1} around its expectation. Since $\mathrm{Var}(\alpha^{-1} \mid d^M, B) = 2/(m-1)$, the expression in (27) equals approximately

$$E\left(\frac{2(1+f-g)^2 f^2}{(m-1)(1+f-\gamma g)^4} \middle| \, d^M \right) + \mathrm{Var}\left(\frac{1+f-g}{1+f-\gamma g} \middle| \, d^M \right). \tag{28}$$

We now use first order Taylor series expansions in γ^{-1} around its expectation to determine the components of (28). The first term in (28) is,

$$E\left(\frac{2(1+f-g)f^2}{(m-1)(1+f-\gamma g)^4} \middle| d^M \right) \approx \frac{2f^2}{(m-1)(1+f-g)^2}. \tag{29}$$

Since $\mathrm{Var}(\gamma^{-1} \mid d^M) = 2/(m(r-1))$, the second term in (28) is

$$\mathrm{Var}\left(\frac{1+f-g}{1+f-\gamma g} \middle| d^M \right) \approx \frac{2g^2}{m(r-1)(1+f-g)^2}. \tag{30}$$

Combining (29) and (30), the variance of (23) equals approximately

$$\frac{2f^2}{(m-1)(1+f-g)^2} + \frac{2g^2}{m(r-1)(1+f-g)^2}. \tag{31}$$

Since a mean square random variable has variance equal to 2 divided by its degrees of freedom, we conclude that

$$v_M = \left(\frac{f^2}{(m-1)(1+f-g)^2} + \frac{g^2}{m(r-1)(1+f-g)^2} \right)^{-1}. \tag{32}$$

ACKNOWLEDGEMENTS

This research was supported by the U.S. Bureau of the Census under a contract through Datametrics Research. The author thanks Rod Little, Trivellore Raghunathan, Don Rubin, Laura Zayatz, and the referees for inspiration, guidance, and helpful comments on this topic.

REFERENCES

ABOWD, J.M., and WOODCOCK, S.D. (2001). Disclosure limitation in longitudinal linked data. In *Confidentiality, Disclosure, and Data Access: Theory and Practical Applications for Statistical Agencies*, (Eds. P. Doyle, J. Lane, L. Zayatz and J. Theeuwes), Amsterdam: North-Holland. 215–277.

BARNARD, J., and RUBIN, D.B. (1999). Small sample degrees of freedom with multiple imputation. *Biometrika*, 86, 948–955.

BINDER, D.A., and SUN, W. (1996). Frequency valid multiple imputation for surveys with a complex design. In *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 281–286.

DALENIUS, T., and REISS, S.P. (1982). Data-swapping: A technique for disclosure control. *Journal of Statistical Planning and Inference*, 6, 73–85.

DUNCAN, G.T., KELLER-MCNULTY, S.A. and STOKES, S.L. (2001). Disclosure risk *vs.* data utility: The R-U confidentiality map. Technical report, U.S. National Institute of Statistical Sciences.

FIENBERG, S.E., MAKOV, U.E. and STEELE, R.J. (1998). Disclosure limitation using perturbation and related methods for categorical data. *Journal of Official Statistics*, 14, 485–502.

FULLER, W.A. (1993). Masking procedures for microdata disclosure limitation. *Journal of Official Statistics*, 9, 383–406.

KENNICKELL, A.B. (1997). Multiple imputation and disclosure protection: The case of the 1995 Survey of Consumer Finances. In *Record Linkage Techniques*, (Eds. W. Alvey and B. Jamerson), Washington, D.C.: National Academy Press, 248–267.

LITTLE, R.J.A. (1993). Statistical analysis of masked data. *Journal of Official Statistics*, 9, 407–426.

LIU, F., and LITTLE, R.J.A. (2002). Selective multiple imputation of keys for statistical disclosure control in microdata. In *Proceedings of the Joint Statistical Meetings*, American Statistical Association, 2133–2138.

MENG, X.L. (1994). Multiple-imputation inferences with uncongenial sources of input (disc: P558-573). *Statistical Science*, 9, 538–558.

RAGHUNATHAN, T.E., LEPKOWSKI, J.M., VAN HOEWYK, J. and SOLENBERGER, P. (2001). A multivariate technique for multiply imputing missing values using a series of regression models. *Survey Methodology*, 27, 85–96.

RAGHUNATHAN, T.E., REITER, J.P. and RUBIN, D.B. (2003). Multiple imputation for statistical disclosure limitation. *Journal of Official Statistics*, 19, 1–16.

REITER, J.P. (2002). Satisfying disclosure restrictions with synthetic data sets. *Journal of Official Statistics*, 18, 531–544.

REITER, J.P. (2003). Inference for partially synthetic, public use microdata sets. *Survey Methodology*, 29, 181–189.

REITER, J.P. (2004a). Releasing multiply-imputed, synthetic public use microdata: An illustration and empirical study. *Journal of the Royal Statistical Society, Series A*. Forthcoming.

REITER, J.P. (2004b). Significance tests for multi-component estimands from multiply-imputed, synthetic microdata. *Journal of Statistical Planning and Inference*. Forthcoming.

RUBIN, D.B. (1976). Inference and missing data (with discussion). *Biometrika*, 63, 581–592.

RUBIN, D.B. (1987). *Multiple Imputation for Nonresponse in Surveys*. New York: John Wiley & Sons, Inc.

RUBIN, D.B. (1993). Discussion: Statistical disclosure limitation. *Journal of Official Statistics*, 9, 462–468.

RUBIN, D.B. (2003). Nested multiple imputation of NMES via partially incompatible MCMC. *Statistica Neerlandica*, 57, 3–18.

SHEN, Z. (2000). *Nested Multiple Imputation*. Ph. D. thesis, Harvard University, Dept. of Statistics.

WILLENBORG, L., and DE WAAL, T. (2001). *Elements of Statistical Disclosure Control*. New York: Springer-Verlag.

ACKNOWLEDGEMENTS

Survey Methodology wishes to thank the following people who have provided help or served as referees for one or more papers during 2004.

M. Axelson, *Statistics Sweden*
J.-F. Beaumont, *Statistics Canada*
D.R. Bellhouse, *University of Western Ontario*
P. Biemer, *Research Triangle Institute*
D.A. Binder, *Statistics Canada*
J.M. Brick, *Westat, Inc.*
J. Breidt, *Iowa State University*
D. Cantor, *Westat, Inc.*
P. Cantwell, *U.S. Bureau of the Census*
A. Chaudhuri, *Institute of Engineering & Technology Lucknow*
B.-C. Chen, *U.S. Census Bureau*
J. Chen, *University of Waterloo*
K.R. Copeland, *U.S. Bureau of Labor Statistics*
J.R. Chromy, *RTI, Inc.*
P. Dick, *Statistics Canada*
J. Droitcour, *United States General Accounting Office*
J.L. Eltinge, *U.S. Bureau of Labor Statistics*
D. Fogel, *Natural Selection, Inc.*
O. Frank, *Stockholm University*
W.A. Fuller, *Iowa State University*
J. Gambino, *Statistics Canada*
L. Granquist, *Statistics Sweden*
B. Graubard, *National Cancer Institute*
D. Haziza, *Statistics Canada*
D. Hedeker, *University of Illinois*
D. Hedin, *Statistics Sweden*
M.A. Hidiroglou, *Office for National Statistics*
M. Houbiers, *Statistics Denmark*
J. Jiang, *University of California at Davis*
D. Judkins, *Westat, Inc.*
G. Kalton, *Westat, Inc.*
M. Kalzoff, *National Centre for Health Statistics*
A. Kennickell, *Federal Research Board*
P. Kott, *USDA/NASS*
M. Kovačević, *Statistics Canada*
J. Kovar, *Statistics Canada*
M.D. Larsen, *Iowa State University*
P. Lahiri, *JPSM, University of Maryland*
P. Lavallée, *Statistics Canada*
R. Lehtonen, *University of Jyvaskyla*
S. Lohr, *Arizona State University*
D. Malec, *United States Bureau of the Census*
H. Mantel, *Statistics Canada*
D. Marker, *Westat, Inc.*
S.M. Miller, *U.S. Bureau of Labour Statistics*
J. Moore, *U.S. Bureau of the Census*
R. Mukerjee, *Indian Institute of Management*
G. Nathan, *Hebrew University*
D. Norris, *Statistics Canada*
J. Opsomer, *Iowa State University*
D. Paton, *Statistics Canada*
D. Pfeffermann, *Hebrew University*
J.N.K. Rao, *Carleton University*
T.J. Rao, *Indian Statistical Institute*
J. Reiter, *Duke University*
L.-P. Rivest, *Université Laval*
L. Rizzo, *Westat, Inc.*
R.A. Rottach, *U.S. Bureau of the Census*
K. Rust, *Westat, Inc.*
N. Schenker, *National Center for Health Statistics*
F.J. Scheuren, *National Opinion Research Center*
I. Şchiopu-Kratina, *Statistics Canada*
N. Shlomo, *Central Bureau of Statistics*
A.C. Singh, *RTI, Inc.*
B. Sinha, *ISI*
M.D. Sinclair, *Mathematica Policy research*
C.J. Skinner, *University of Southampton*
P. Smith, *CDC*
E. Stasny, *Ohio State University*
D. Steel, *University of Wollongong*
L. Stokes, *Southern Methodist University*
C. Swhwarz, *Simon Fraser University*
A. Tersine, Jr., *United States Bureau of Labor Statistics*
A. Thivierge, *Statistics Canada*
M. Thompson, *University of Waterloo*
Y. Tillé, *Université de Neuchâtel*
C. Tucker, *United States Bureau of Labor Statistics*
J. van der Brakel, *Central Bureau of Statistics*
R. Valliant, *JPSM, University of Michigan*
V. Vehovar, *University of Ljubljana*
J. Waksberg, *Westat, Inc.*
J. Wang, *Merck Research Labs, Merck & Co., Inc.*
F. Wein, *Federal Statistical Office*
K.M. Wolter, *Iowa State University*
P. Wong, *Statistics Canada*
C. Wu, *University of Waterloo*
Y. You, *Statistics Canda*
R. Yucel, *University of Massachusetts*
W. Yung, *Statistics Canada*
A. Zaslavsky, *Harvard University*

Acknowledgements are also due to those who assisted during the production of the 2004 issues: Anne-Marie Fleury, Francine Pilon-Renaud and Roberto Guido (Dissemination Division), Philippe Laroche (Marketing Division) and François Beaudin (Official Languages and Translation Division). Finally we wish to acknowledge Christine Cousineau, Céline Ethier, and Denis Lemire of Household Survey Methods Division, for their support with coordination, typing and copy editing.

Erratum:

In the June 2004 issue, we published a paper by D.N. Da Silva and Jean D. Opsomer on "Properties of the Weighting Cell Estimator Under a Nonparametric Response Mechanism" (pages 45-55). We would like to apologize for having incorrectly spelled out Dr. Da Silva's name. It should have read D. Nobrega Da Silva. Please note also that the corrected version appears on Statistics Canada's Web site.

JOURNAL OF OFFICIAL STATISTICS

An International Review Published by Statistics Sweden

JOS is a scholarly quarterly that specializes in statistical methodology and applications. Survey methodology and other issues pertinent to the production of statistics at national offices and other statistical organizations are emphasized. All manuscripts are rigorously reviewed by independent referees and members of the Editorial Board.

Contents
Volume 20, No. 2, 2004

All inquires about submissions and subscriptions should be directed to jos@scb.se

CONTENTS TABLE DES MATIÈRES

Volume 32, No. 1, March/mars 2004, 1-104

Volume 32, No. 2, June/juin 2004, 105-208

CONTENTS TABLE DES MATIÈRES

Volume 32, No. 3, September/septembre 2004, 209-334

GUIDELINES FOR MANUSCRIPTS

Before having a manuscript typed for submission, please examine a recent issue of *Survey Methodology* (Vol. 19, No. 1 and onward) as a guide and note particularly the points below. Articles must be submitted in machine-readable form, preferably in Word. A paper copy may be required for formulas and figures.

1. Layout

1.1 Manuscripts should be typed on white bond paper of standard size (8½ × 11 inch), one side only, entirely double spaced with margins of at least 1½ inches on all sides.
1.2 The manuscripts should be divided into numbered sections with suitable verbal titles.
1.3 The name and address of each author should be given as a footnote on the first page of the manuscript.
1.4 Acknowledgements should appear at the end of the text.
1.5 Any appendix should be placed after the acknowledgements but before the list of references.

2. Abstract

The manuscript should begin with an abstract consisting of one paragraph followed by three to six key words. Avoid mathematical expressions in the abstract.

3. Style

3.1 Avoid footnotes, abbreviations, and acronyms.
3.2 Mathematical symbols will be italicized unless specified otherwise except for functional symbols such as "exp(·)" and "log(·)", *etc.*
3.3 Short formulae should be left in the text but everything in the text should fit in single spacing. Long and important equations should be separated from the text and numbered consecutively with arabic numerals on the right if they are to be referred to later.
3.4 Write fractions in the text using a solidus.
3.5 Distinguish between ambiguous characters, (*e.g.*, w, ; 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.