SURVEY METHODOLOGY

Catalogue 12-001

A JOURNAL PUBLISHED BY STATISTICS CANADA

JUNE 1996
•
VOLUME 22
•
NUMBER 1

Statistics Canada Statistique Canada

Canada

SURVEY METHODOLOGY

A JOURNAL PUBLISHED BY STATISTICS CANADA

JUNE 1996 • VOLUME 22 • NUMBER 1

Published by authority of the Minister
responsible for Statistics Canada

June 1996

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

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

EDITORIAL POLICY

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

Submission of Manuscripts

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

Subscription Rates

The price of Survey Methodology (Catalogue no. 12-001-XPB) is $45 per year in Canada, US $50 in the United States, and US $55 per year for other countries. Subscription order should be sent to Statistics Canada, Operations and Integration Division, Circulation Management, 120 Parkdale Avenue, Ottawa, Ontario, Canada K1A 0T6. A reduced price is available to members of the American Statistical Association, the International Association of Survey Statisticians, the American Association for Public Opinion Research, and the Statistical Society of Canada.

SURVEY METHODOLOGY

A Journal Published by Statistics Canada

Volume 22, Number 1, June 1996

CONTENTS

Survey Methodology, June 1996
Vol. 22, No. 1, pp. 1–2
Statistics Canada

In This Issue

This issue of *Survey Methodology* contains articles dealing with a variety of subjects. In the first article, Steel, Holt and Tranmer examine the problem of using aggregated data in studies on relationships at the individual or household level. They propose a simple general model that seeks to take account of the geographical effects of aggregation. They then describe how this model effects both the estimation of population means and covariance matrices and analysis at the regional level. In addition, by introducing auxiliary variables for which certain external sources provide an estimate of the covariance matrix at the unit level, the authors propose methods that provide an unbiased estimate of the parameters at the individual level, so as to avoid the effect of geographical aggregation.

Binder gives a "cookbook" approach for deriving Taylor series approximations to the variances of a wide class of estimators from complex surveys. Several useful examples are presented, as well as new results on the application of this general technique to two-phase sampling. A justification of this method is given, showing the procedure to be consistent with the formulation given in earlier work by Binder and Patak.

Yung and Rao suggest a linear approximation to the jackknife variance estimator. This linearized jackknife inherits the good statistical properties of the usual jackknife variance estimator but is computationally much less intensive. The specific form of the proposed variance estimator is developed for the generalized regression estimator of a total and for the ratio of two generalized regression estimators. In a simulation study using data from the U.S. Current Population Survey, they found that the jackknife, the linearized jackknife, and the usual linearization variance estimators worked quite well for poststratified estimates of a total, while an incorrect form of the jackknife was badly biased.

Chaubey, Nebebe and Chen consider use of an Inverse Gaussian model for positively skewed data and develop a corresponding model assisted estimators for domain totals, which consist of Inverse Gaussian regression predictors together with an expansion estimators of the regression bias. A modified version of the estimator which gives reduced weight to the bias correction term, analogous to a modified regression estimator proposed by Särndal and Hidiroglou, is also proposed. In a simulation study using synthetic income data based on Statistics Canada's Survey of Household Income, Facilities and Finance the proposed estimators are found to work reasonably well.

Rizzo, Kalton and Brick investigate the use of auxiliary information in compensating for panel nonresponse through weight adjustment techniques. Using data from the Survey of Income and Program Participation (SIPP) to illustrate, they address two important issues, namely, the choice of auxiliary variables to be used in a nonresponse weight adjustment technique, and the choice of technique itself. A screening procedure in conjunction with logistic regression modelling are the means by which appropriate auxiliary variables are chosen. The nonresponse weighting adjustment methods considered are based on logistic regression models, categorical search algorithms and generalized raking. An empirical comparison of the various methods is discussed in detail.

Ding and Fienberg develop models of matching error which can be used in estimation of total population from a probabilistic match of two or more samples. They develop their models for the particular application of a multiple sample census, that is, a census supplemented by auxiliary samples. They illustrate the usefulness of their methods by applying them in an analysis of the 1988 St. Louis Dress Rehearsal Census data for which three samples were matched: the Census itself, the Post Enumeration Survey sample, and the Administrative List Supplement.

In a paper on optimal stratification, Slanta and Krenzke talk about the use of the Lavallée-Hidiroglou method. This iterative method minimizes the sample size while fixing the coefficient of variation. In a practical illustration, the authors present the difficulties with the Lavallée-Hidiroglou method and show how they were resolved.

Dagum proposes a new method for estimating underlying trends from seasonally adjusted data. The approach consists of two steps. The seasonally adjusted data are first extrapolated based on an ARIMA model. A 13-term Henderson filter is then applied to the extended series, using strict sigma limits for the identification and replacement of extreme values. The new method is compared to the standard method using data from several economic time series. It is found that the new method produces fewer unwanted ripples in the estimated trend, while identifying turning points as just quickly and requiring smaller revisions on average.

Tillé proposes an algorithm that generalizes the selection-rejection method used for constructing a simple random sample without replacement. A specific case of this algorithm, which is called the "mobile stratification algorithm", is discussed. It serves to obtain a smoothed stratification effect by using as a stratification variable the serial number of the units of observation. This algorithm gets around the thorny problem of a continuous variable in strata.

De Waal and Willenborg review recent research on statistical disclosure control for microdata files from the perspective of Statistics Netherlands. Models are developed for the probability that a particular record could be re-identified and for the probability that some record in a microdata file could be re-identified. Global recoding and local suppression are considered as methods to reduce disclosure risk. They conclude that there is still much need for further methodological research and development of efficient software.

Finally, it is with sadness that I note the recent passing away of Maria Gonzalez, who died of cardiac arrest while vacationing in Puerto Rico this past February. Among her many contributions to the statistical community, for the past several years Maria has been an Associate Editor for the *Survey Methodology* journal. Her contribution in this capacity to the quality and breadth of this journal was very much appreciated, and she will be sorely missed. An obituary, written by Elizabeth and Fritz Scheuren, appeared in the April issue of *Amstat News.*

The Editor

Survey Methodology, June 1996
Vol. 22, No. 1, pp. 3–15
Statistics Canada

Making Unit-Level Inferences From Aggregated Data

D.G. STEEL, D. HOLT and M. TRANMER[1]

ABSTRACT

Data are often available only as a set of group or area means. However, it is well known that statistical analysis based on such data will often produce results very different from those obtained from analysing the corresponding individual or household data. If the results of area level analyses are thought to apply to the individual level then we risk committing the ecological fallacy. Aggregation or ecological effects arise in part because geographic areas are not comprised of random groupings of people or households but exhibit strong socio-economic differences between areas. The population structure must be incorporated into the statistical model underpinning the analysis if aggregation effects are to be understood. A simple general model is proposed to achieve this and the consequences of the model and its implications for the estimation of population means and covariance matrices are obtained. Furthermore, methods are suggested which can provide unbiased estimates of individual level parameters from aggregated data and so avoid the ecological fallacy. These methods rely on identifying the "grouping variables" that characterise the process that led to the population structure, or at least characterise the area differences. An estimate of the unit level covariance matrix of the grouping variables is required from some source. Data from the 1991 Census of the United Kingdom have been analysed to identify the important grouping variables and evaluate the effectiveness of the proposed adjustment methods for the estimation of covariance matrices and correlation coefficients. These results lead to a suggested strategy for the analysis of aggregated data.

KEY WORDS: Aggregation; Ecological fallacy; Grouping; Selection; Variance components.

1. INTRODUCTION

Researchers are often faced with the problem of wishing to investigate individual level relationships but having to make use of aggregated data, such as the means or totals for geographic areas. Ideally unit level data collected in a sample survey or census would be used, but may not be accessible because of confidentiality restrictions, or because the variables have not been collected in a recent survey or census. Administrative systems provide information on a range of variables, for example on unemployment, health, morbidity, but because of confidentiality requirements these data are usually made available for aggregates, such as geographic areas. The census also provides data for geographic areas. For these reasons, analysis of group level data is still an option used widely in social and epidemiological research.

Consider a population in which each individual has associated a vector of variables of interest, whose distribution has mean μ_y and covariance matrix Σ_{yy}. We are interested in relationships among the variables of interest as reflected by correlations, regression coefficients and principal components, which may all be derived from the covariance matrix, Σ_{yy}, which is our basic target of inference. For example, the variables of interest might include a set of attainment tests in an educational study; the incidence of a particular disease and a set of explanatory variables in an epidemiological study; or a set of deprivation measures in a sociological study. We suppose that individual level data are unavailable. However, the region may be subdivided into a set of small areas such as Census Enumeration Districts (EDs), and for each small area, g, or for a sample of areas, we observe the vector of average values $\bar{y}_g$ for the variables of interest together with the sample size n_g on which this is based.

The objective of the analysis, Σ_{yy}, is a covariance matrix which spans the small areas. The target of inference is not conditional on small area membership but refers to the marginal distribution across small areas. This contrasts with situations, such as small area estimation, in which the target of inference is in the conditional distribution given the small area. This is a separate, legitimate objective with which we are not concerned. The same models may be applicable, but the targets of inference are different. However, our formulation does allow for group specific variables to be included as variables of interest if required. For example, if we associate with each individual a set of ED means for the area in which the individual is located, then these can be included within the vector, y, of interest. In particular, regression analyses which include small area means as explanatory variables in the regression model can be encompassed by the approach.

The literature associated with the analysis of aggregated data dates back to Gehlke and Biehl (1934) and includes significant contributions by Yule and Kendall (1950) and Robinson (1950), Blalock (1964), Openshaw and Taylor

[1] D.G. Steel, Department of Applied Statistics, University of Wollongong, NSW 2522, Australia; D. Holt and M. Tranmer, Department of Social Statistics, University of Southampton, S017 1BJ, United Kingdom.

(1979) and more recently Arbia (1989). There are also problems associated with the fact that the areal units used often have no special significance, being constructed for reasons of cost, operational or administrative convenience. Moreover, the results of the group level analysis will depend on the scale of the units, that is their average size and the particular set of boundaries chosen. Several empirical studies have demonstrated these effects, including Clark and Avery (1976), Perle (1977), Openshaw (1984), and Fotheringham and Wong (1991). However, these studies have not provided any generally applicable theory or practical methods of modifying the results of group level analyses to provide reliable unit level inferences.

Aggregation effects arise because geographic units are not comprised of random groupings of people. Individuals in the same area generally tend to be more alike because they choose to live in areas in a non-random way, or because they are subjected to common influences, or because they interact with one another. Thus there are socio-economic differences between areas which are confounded with the individual effects in any statistical analysis performed using aggregated data for the areas. A simple general model is proposed which seeks to incorporate these effects. The consequences of this model and its implications for area level analysis are obtained. Furthermore, methods are suggested which provide, under certain circumstances, unbiased estimates of individual level parameters from aggregated level data and so avoid the ecological fallacy. These methods involve auxiliary variables for which a unit level sample covariance matrix is available from some source. This approach has been applied to data from the 1991 Census of the United Kingdom and a strategy developed for the analysis of aggregated data.

2. MODELS FOR AREA EFFECTS

We consider a population of N individuals each having a vector y of characteristics of interest. The population is comprised of M groups and the random variable c_i indicates the area to which the i-th population unit belongs. The number of individuals in the g-th area is N_g.

We consider μ_y and Σ_{yy} to be superpopulation parameters and the following statistical theory is obtained in this framework. However, we consider some survey design issues at the end of section 2.

We assume that there exists a sample data set s of size n and that these individual data have been aggregated to provide a set of m area means which are available for analysis. The following area level statistics can be calculated:

the g-th area mean:

$$\bar{y}_g = \frac{1}{n_g} \sum_{i \in g,s} y_i \qquad (2.1)$$

the overall sample mean:

$$\bar{y} = \frac{1}{n} \sum_{g \in s} n_g \bar{y}_g = \frac{1}{n} \sum_{i \in s} y_i \qquad (2.2)$$

the area level sample covariance matrix:

$$\bar{S}_{yy} = \frac{1}{m-1} \sum_{g \in s} n_g (\bar{y}_g - \bar{y})(\bar{y}_g - \bar{y})'. \qquad (2.3)$$

Analogous unit level statistics may be defined but will be unavailable to the analyst. For example $S_{yy} = 1/(n-1) \sum_{i \in s} (y_i - \bar{y})(y_i - \bar{y})'$ is the unit level sample covariance matrix.

2.1 Random Grouping

While geographic groups are rarely formed randomly, such a situation is a useful starting point in considering ecological analysis. If groups are randomly formed then many group level analyses are valid, albeit with a reduced efficiency. Steel and Holt (1995) consider the properties of statistics such as means, variances, regression and correlation coefficients in this situation. When the groups are randomly formed *i.e.*, $y \perp c$ then

$$E[\bar{y}_g \mid s,c] = \mu_y \qquad (2.4)$$

$$V(\bar{y}_g \mid s,c) = \frac{1}{n_g} \Sigma_{yy}. \qquad (2.5)$$

The basic properties of the unit and group level statistics then follow readily

$$\mathrm{Cov}(\bar{y}_g, \bar{y}_h \mid s,c) = 0 \quad g \neq h \qquad (2.6)$$

$$E[\bar{y} \mid s,c] = \mu_y \qquad (2.7)$$

$$E[S_{yy} \mid s,c] = \Sigma_{yy} \qquad (2.8)$$

$$E[\bar{S}_{yy} \mid s,c] = \Sigma_{yy}. \qquad (2.9)$$

These properties apply if the sampling is ignorable given the group indicatives, which means the sample design can depend on the groups but not on y or any variable which is related to y conditional on c. For example a census or a simple random sample of groups and units within groups may be used.

Unweighted group level statistics may be used by setting $n_g = 1$ in equations (2.2) and (2.3). This leads to inefficient estimators. The degree of inefficiency will depend on the distribution of the group sample sizes. Weighting by the group sample sizes is important and when this is done inference can proceed as usual with appropriate adjustments to the degrees of freedom. Variability is determined by the number of areas rather than the number of individual observations and confidence intervals and tests are adjusted accordingly.

2.2 A Variance Component Model

A simple way to represent the positive intra-group correlation that is usually observed in grouped populations is through a variance components model, which in the multivariate case corresponds to

$$y_i = \mu_y + v_g + \epsilon_i \qquad i \in g$$

where v_g and ϵ_i are independent random components at the group and individual level respectively, both with zero expectation, $V(\epsilon_i \mid c) = \Sigma_{\epsilon\epsilon}$ and $V(v_g \mid c) = \Delta_{yy}$.

Model A:

$$E[y_i \mid c] = \mu_y \tag{2.10}$$

$$V(y_i \mid c) = \Sigma_{\epsilon\epsilon} + \Delta_{yy} = \Sigma_{yy} \tag{2.11}$$

$$\begin{aligned} \mathrm{Cov}(y_i, y_j \mid c) &= \Delta_{yy} \quad \text{if} \quad c_i = c_j \quad i \neq j \\ &= 0 \text{ otherwise.} \end{aligned} \tag{2.12}$$

The notation $V(\cdot \mid c)$ implies the covariance matrix conditional on the group labels c and hence determines common group membership. It is, however, taken to be unconditional over the group level random effects. Thus $V(y_i \mid c)$ contains the total variance from both the within group covariance matrix $\Sigma_{\epsilon\epsilon}$ and the group level covariance matrix Δ_{yy}.

The properties of the sample group level means follow readily from Model A, if the sampling is ignorable given c,

$$E[\bar{y}_g \mid s,c] = \mu_y \tag{2.13}$$

$$V(\bar{y}_g \mid s,c) = \frac{1}{n_g}(\Sigma_{yy} + (n_g - 1)\Delta_{yy}) \tag{2.14}$$

$$\mathrm{Cov}(\bar{y}_g, \bar{y}_h \mid s,c) = 0 \quad g \neq h. \tag{2.15}$$

The properties of the unit level and group level statistics are

$$E[\bar{y} \mid s,c] = \mu_y \tag{2.16}$$

$$E[S_{yy} \mid s,c] = \Sigma_{yy} - \frac{\bar{n}^0 - 1}{n - 1}\Delta_{yy} \tag{2.17}$$

$$E[\bar{S}_{yy} \mid s,c] = \Sigma_{yy} + (\bar{n}^* - 1)\Delta_{yy} \tag{2.18}$$

where $\bar{n} = n/m$, $\bar{n}^0 = 1/n \sum_{g \epsilon s} n_g^2 = \bar{n}(1 + C_n^2)$, $\bar{n}^* = \bar{n}(1 - C_n^2/(m-1))$ and $C_n^2 = 1/m \sum_{g \epsilon s}(n_g - \bar{n})^2/\bar{n}^2$ is the square of the coefficient of variation of the group sample sizes in the sample. We note that the coefficient of Δ_{yy} is $0(m^{-1})$ in (2.17) but is $0(\bar{n})$ in (2.18). This illustrates how a small bias in the unit level analysis can be magnified into a much larger bias in the aggregate level analysis. We will discuss these results further in section 2.4.

2.3 Grouping Models

In the discussion of ecological analysis, models have been proposed which take into account the group formation process. In this approach it is assumed that there is a grouping process which allocates individual units to groups according to a vector of grouping variables, z_i, either stochastically or deterministically. This approach is implicit in Blalock's (1964) analysis and used explicitly by Hannan and Burstein (1974), Litchman (1974), Langbein and Litchman (1978), Smith (1977) and Blalock (1979, 1985). Steel (1985) refers to these models as grouping models since it is assumed that groups are formed by some process involving the variables in the relationships under study. The grouping is seen as a distorting effect and the relationships of interest are defined before the grouping has occurred. It is often noted in the discussion of contextual models that apparent contextual effects may in fact be due to such factors. The multivariate version of this model is:

Model B:

$$E[y_i \mid z,c] = \mu_{y.z} + \beta_{yz}' z_i \tag{2.19}$$

$$V(y_i \mid z,c) = \Sigma_{yy.z} \tag{2.20}$$

$$\mathrm{Cov}(y_i, y_j \mid z,c) = 0 \quad i \neq j. \tag{2.21}$$

In this model the conditional expectation of y_i depends only on the value of the auxiliary variables for the i-th unit and is independent of the group to which the unit belongs or the values of the auxiliary variables of other units in the population. The conditional covariance between any two units is zero. This model covers grouping models in which the group formation process is characterised by the auxiliary variables z_i. The auxiliary variables can be thought of as those variables that determine to which group a unit belongs. More generally, the auxiliary variables can be regarded as the main individual level variables whose distributions are not random across groups because of the choice or migration processes to which the population has been subjected. Contextual variables can also be included in this model as auxiliary variables which take the same value for each unit in the group.

If the vector of auxiliary variables has a marginal distribution with mean μ_z and covariance matrix Σ_{zz}, then the marginal mean and covariance matrix of y are given by $\mu_y = \mu_{y.z} + \beta'_{yz}\,\mu_z$ and $\Sigma_{yy} = \Sigma_{yy.z} + \beta'_{yz}\,\Sigma_{zz}\,\beta_{yz}$ respectively. The properties of the sample group level means follow readily from Model B:

$$E[\bar{y}_g \mid s,z,c] = \mu_y + \beta'_{yz}(\bar{z}_g - \mu_z) \tag{2.22}$$

$$V(\bar{y}_g \mid s,z,c) = \frac{1}{n_g}\Sigma_{yy.z} \tag{2.23}$$

$$\mathrm{Cov}(\bar{y}_g,\bar{y}_h \mid s,z,c) = \mathbf{0} \qquad g \neq h. \tag{2.24}$$

The group level statistics then have the following properties

$$E[\bar{y} \mid s,z,c] = \mu_y + \beta'_{yz}(\bar{z} - \mu_z) \tag{2.25}$$

$$E[S_{yy} \mid s,z,c] = \Sigma_{yy} + \beta'_{yz}(S_{zz} - \Sigma_{zz})\beta_{yz} \tag{2.26}$$

$$E[\bar{S}_{yy} \mid s,z,c] = \Sigma_{yy} + \beta'_{yz}(\bar{S}_{zz} - \Sigma_{zz})\beta_{yz} \tag{2.27}$$

where S_{zz} and $\bar{S}_{zz}$ are defined analogously to S_{yy} and $\bar{S}_{yy}$ as given in equation (2.3) and the sentence that follows it.

2.4 A Combined Model

The two models considered so far can be thought of as competing explanations of the group effects, but they can be combined into a more realistic model which contains both grouping effects and residual variance components:

Model C:

$$E[y_i \mid z,c] = \mu_{y.z} + \beta'_{yz}\,z_i \tag{2.28}$$

$$V(y_i \mid z,c) = \Sigma_{yy.z} \tag{2.29}$$

$$\begin{aligned}\mathrm{Cov}(y_i,y_j \mid z,c) &= \Delta_{yy.z} \quad \text{if} \quad c_i = c_j \quad i \neq j\\ &= 0 \quad \text{otherwise.}\end{aligned} \tag{2.30}$$

This model allows for group formation processes which are characterised by the auxiliary variables z_i. It also includes residual within group correlations which reflect random effects which are interpreted as due to unobserved random group level variables after allowing for the grouping variables.

The properties of the sample group level means follow, if the sampling is ignorable given (z,c) from Model C,

$$E[\bar{y}_g \mid s,z,c] = \mu_y + \beta'_{yz}(\bar{z}_g - \mu_z) \tag{2.31}$$

and

$$V(\bar{y}_g \mid s,z,c) = \frac{1}{n_g}(\Sigma_{yy.z} + (n_g - 1)\Delta_{yy.z}) \tag{2.32}$$

$$\mathrm{Cov}(\bar{y}_g,\bar{y}_h \mid s,z,c) = \mathbf{0} \qquad g \neq h \tag{2.33}$$

$$E[\bar{y} \mid s,z,c] = \mu_y + \beta'_{yz}(\bar{z} - \mu_z) \tag{2.34}$$

$$E[S_{yy} \mid s,z,c] = \Sigma_{yy} + \beta'_{yz}(S_{zz} - \Sigma_{zz})\beta_{yz} - \frac{\bar{n}^0 - 1}{n - 1}\Delta_{yy.z} \tag{2.35}$$

$$E[\bar{S}_{yy} \mid s,z,c] = \Sigma_{yy} + \beta'_{yz}(\bar{S}_{zz} - \Sigma_{zz})\beta_{yz} + (\bar{n}^* - 1)\Delta_{yy.z}. \tag{2.36}$$

Equations (2.17) and (2.18) showed how the effect of aggregation in the variance components model, A, amplifies the contribution of the random group level effects. In equation (2.17) the coefficient of Δ_{yy} is $0(m^{-1})$ whereas in (2.18) it is $0(\bar{n})$. For the combined model, C, equations (2.35) and (2.36) show how inclusion of the grouping variables permit the partition of the bias into two additive terms: the first related to the grouping variables, their relationship to the variables of interest and their aggregation effect and the second term involving $\Delta_{yy.z}$, the residual components of variance after controlling for the grouping variables. Note that the coefficients of $\Delta_{yy.z}$ in equations (2.35) and (2.36) are still $0(m^{-1})$ and $0(\bar{n})$ respectively as they were in equations (2.17) and (2.18) but the residual components of variance should in general be smaller. The basic assumption in (2.29) is that the residual variance is constant across c.

The assumption that the sampling is ignorable given (z,c) means that the sample design can depend on the auxiliary variables and the group indicatives. This allows, for example, the use of stratification based on the values of z and cluster or multi-stage sampling based on the groups.

The weighted group level matrix $\bar{S}_{yy}$ is intended to estimate Σ_{yy}. The first bias term in (2.36) is due to the effect of the grouping variables and will be zero if $\beta_{yz} = \mathbf{0}$ or approximately so if $\bar{S}_{zz} \risingdotseq \Sigma_{zz}$. The condition $\beta_{yz} = \mathbf{0}$ is a strong condition and implies that the variables of interest are unrelated to the grouping variables. The effect of aggregation on the sample covariance of any two variables will depend on the relationships of the variables with the grouping variables z_i and we would expect the aggregation effects to be greater for variables more closely related to the grouping variables. The condition $\bar{S}_{zz} \doteqdot \Sigma_{zz}$ implies that there are no selection or aggregation effects for the z variables. These conditions are unlikely to apply in practice and hence bias will result for many variables. The bias due to the sampling and grouping involving the auxiliary variables is determined by $S_{zz} - \Sigma_{zz}$ for the unit level estimator and by $\bar{S}_{zz} - \Sigma_{zz}$ for the group level estimator. The term $\bar{S}_{zz} - \Sigma_{zz}$ reflects the net effect of the sampling and aggregation on the auxiliary variables.

The second bias term in (2.36) will be zero if $\Delta_{yy.z} = \mathbf{0}$ which implies that, conditional on the grouping variables, there is no residual intra-group correlation among the y variables. This is unlikely to occur in practice but it is desirable to identify grouping variables that account for as much of the aggregation effects as possible by making this residual term as small as possible.

The effects due to the grouping and sampling depending on z and the effect due to the residual within group correlation are additive; this will be the case for more complex forms of within group correlations provided the linearity of the model holds. If z follows a simple variance component model, like Model A then

$$E[\bar{S}_{zz} \mid s,c] = \Sigma_{zz} + (\bar{n}^* - 1)\Delta_{zz}$$

$$E[\bar{S}_{yy} \mid s,c] = \Sigma_{yy} + (\bar{n}^* - 1)\beta'_{yz}\Delta_{zz}\beta_{yz} + \Delta_{yy.z} \qquad (2.37)$$

and the intra-group covariances of the variables of interest are composed of a component due to the intra-group covariances of the auxiliary variables and the residual components. The right hand side of (2.37) represents a partition of (2.18) since if z follows a variance components model then so does y unconditionally. The motivation behind the basic model is to find auxiliary variables so that the residual or conditional within group covariances $\Delta_{yy.z}$ are small or, ideally, disappear.

2.5 Adjusting for Aggregation Effects

Few useful proposals have been made on how to adjust the area level analyses to produce reasonable estimates of the unit level relationship. Duncan and Davis (1953) considered the possible range of the correlation coefficient calculated from a 2 by 2 table with known margins. The resulting bounds are often too wide to be of practical use. Goodman (1959) identified specific conditions for a regression model under which ecological analysis could validly be used to draw inferences regarding relationships at the individual level. Langbein and Litchman (1978) consider some methods that can be applied when grouping is by the dependent variable and unit level variances are available for both the dependent and all the independent variables in the regression model. However, none of these approaches provide a general approach to the problem.

Examining the bias for $\bar{S}_{yy}$, given in (2.36) shows that if we add $\beta'_{yz}(\Sigma_{zz} - \bar{S}_{zz})\beta_{yz}$ to $\bar{S}_{yy}$, the bias term due to the grouping variables would be removed. Now (2.31) implies that

$$E[\bar{\boldsymbol{B}}_{yz} \mid s,z,c] = \beta_{yz} \qquad (2.38)$$

where $\bar{\boldsymbol{B}}_{yz} = \bar{S}_{zz}^{-1}\bar{S}_{zy}$.

If the covariance matrix of z, S_{zzs_0}, from a unit level sample s_0 drawn from m_0 groups was available then the adjusted estimator

$$\hat{\Sigma}_{yy}(z) = \bar{S}_{yy} + \bar{\boldsymbol{B}}'_{yz}(S_{zzs_0} - \bar{S}_{zz})\bar{\boldsymbol{B}}_{yz} \qquad (2.39)$$

should remove the aggregation bias due to the grouping variables z, provided S_{zzs_0} is close to Σ_{zz}. The source for S_{zzs_0} may be quite independent of the data used in $\bar{S}_{yy}$ and $\bar{\boldsymbol{B}}_{yz}$. Steel (1985) shows that the adjusted estimator (2.39) can be obtained as the MLE of Σ_{yy} (with the usual replacement of $m - 1$ by m *etc.*). If normality of the distribution of (y,z) applies, s_0 is a simple random sample from the population and $\Delta_{yy.z} = \mathbf{0}$. The adjusted estimator corresponds to the Pearson (1903) adjustment considered by Holt, Smith and Winter (1980) in the case of regression analysis and Smith and Holmes (1989) in the case of multivariate analysis. In these cases the adjustment is applied to statistics calculated from unit level data obtained from a sample whose design depends on the auxiliary variables. In our case the adjustment is applied to statistics calculated from area means and the auxiliary variables used in the adjustment include grouping variables as well as any design variables. The adjusted estimator of μ_y is

$$\hat{\mu}_y(z) = \bar{y} + \bar{\boldsymbol{B}}'_{yz}(\bar{z}_{s_0} - \bar{z}) \qquad (2.40)$$

where $\bar{z}_{s_0}$ is the mean calculated from s_0.

From (2.34) and (2.38) we see that

$$E[\hat{\mu}_y(z) \mid s,z,s_0,c] = \mu_y + \beta'_{yz}(\bar{z}_{s_0} - \mu_z). \qquad (2.41)$$

Moreover, Steel (1985) shows that (2.36) and (2.38) imply

$$E[\hat{\Sigma}_{yy}(z) \mid s,z,s_0,c] = \Sigma_{yy} + \beta'_{yz}(S_{zzs_0} - \Sigma_{zz})\beta_{yz} + (\bar{n}^* - 1)\Delta_{yy.z} + 0(m^{-1}) \qquad (2.42)$$

provided $\mathrm{tr}(\bar{S}_{zz}^{-1} S_{zzs_0})$ and $\bar{n}\,\mathrm{tr}\big((\bar{S}_{zz}^{-1} S_{zzs_0} - I)\bar{S}_{zz}^{-1} \bar{S}_{zz}^{(2)}\big)$ are bounded, where $\bar{S}_{zz}^{(2)}$ is defined similarly to $\bar{S}_{zz}$ with n_g replaced by $n_g^2/\bar{n}$.

Comparing (2.42) with (2.35) we see that the component of bias due to the grouping variables has been adjusted to that associated with the use of S_{yys_0}, if it had been available. The estimator adjusts for the aggregation effects that have acted through z. It also adjusts the effect of the sampling design from that associated with s to that associated with s_0.

Suppose that the sampling design used to generate s_0 and the values of the auxiliary variables are generated from a superpopulation such that

$$E[\bar{z}_{s_0} \mid s_0,c] = \mu_z + 0(m_0^{-1}) \qquad (2.43)$$

$$E[S_{zzs_0} \mid s_0,c] = \Sigma_{zz} + 0(m_0^{-1}) \qquad (2.44)$$

where m_0 is the number of groups in s_0.

In such cases

$$E[\hat{\mu}_y(z) \mid s,s_0,c] = \mu_y + 0(m_0^{-1}) \qquad (2.45)$$

$$E[\hat{\Sigma}_{yy}(z) \mid s,s_0,c] = \Sigma_{yy} + (\bar{n}^* - 1)\Delta_{yy.z} + 0(\tilde{m}^{-1}) \qquad (2.46)$$

where

$$\tilde{m} = \min(m,m_0).$$

Conditions (2.43) and (2.44) would apply if the population z values across groups arose from a variance component model similar to model A and the sampling design for s_0 depended only on the grouping but not any auxiliary variables. Sampling designs such as simple random sampling or equal probability cluster or multi stage sampling fulfil this condition. Use of census data, so that s_0 is the entire finite population is also applicable.

It is thus possible to adjust for the bias due to the grouping variables provided some unit level sample covariance matrix for z is available. The motivation for the approach is a situation where the predominant group effects can be attributed to selectivity or grouping effects acting through the grouping variables. The adjustment for the auxiliary variables removes the effect of the apparent intra-group correlation due to these variables. The adjusted estimator still has a component of bias due to $\Delta_{yy.z}$ and if z is not effective in significantly reducing the intra-group correlations then this term can still be important. This approach therefore relies on choice of appropriate auxiliary variables to reduce the intra-group correlations.

If the sampling design for s_0 and the superpopulation model for z are such that (2.43) and (2.44) do not apply then $\bar{z}_{s_0}$ and S_{zzs_0} can be replaced by estimators $\hat{\mu}_{zs_0}$ and $\hat{\Sigma}_{zzs_0}$ in the calculation of the adjusted estimators $\hat{\mu}_y(z)$ and $\hat{\Sigma}_{yy}(z)$. The resulting expectations of the adjusted estimators are given by (2.41) and (2.42) with $\bar{z}_{s_0}$ replaced by $\hat{\mu}_{zs_0}$ and S_{zzs_0} replaced by $\hat{\Sigma}_{zzs_0}$. There are a number of choices available for the estimators $\hat{\mu}_{zs_0}$ and $\hat{\Sigma}_{zzs_0}$ calculated from the sample s_0. Smith and Holmes (1989) consider a range of model based and design based estimators that can be used. For example suppose the sample design used to obtain s_0 involved stratification according to the values of the vector of size variables x. Denote the sample inclusion probability for population unit i as Π_i and the associated probability based weight is $w_i = (\Pi_i)^{-1}$. The probability weighted estimator of μ_z is $\bar{z}^*_{s_0} = \sum_{i\epsilon s_0} w_i z_i$, and of Σ_{zz} is $S_{zzs_0} = \sum_{i\epsilon s_0} w_i z_i z_i' - w_0^{-1} \bar{z}^*_{s_0} \bar{z}'^*_{s_0}$ where $w_0 = \sum_{i\epsilon s_0} w_i$.

The Pearson based adjusted estimators of μ_z and Σ_{zz} are $\bar{z}_{s_0} + B'_{zxs_0}(\bar{x}_u - \bar{x}_{s_0})$ and $S_{zzs_0} + B'_{zxs_0}(S_{xxu} - S_{xxs_0})$ B_{zxs_0} respectively. Here $\bar{x}_u$ and S_{xxu} are the mean vector and covariance matrix of the design variables in x in the finite population and $B_{zxs_0} = S_{xxs_0^{-1}} S_{xzs_0}$.

Pobability weighted Pearson based adjusted estimates may also be considered, *i.e.*, $\bar{z}^*_{s_0} + B'_{zxs_0}(\bar{x}_u - \bar{x}^*_{s_0})$ and $S^*_{zzs_0} + B'_{zxs_0}(S_{xxu} - S^*_{xxs_0})\, B^*_{zxs_0}$.

Here $\bar{x}^*_{s_0}$ and $S^*_{xxs_0}$ are defined analogously to $\bar{z}^*_{s_0}$ and $S^*_{zzs_0}$ respectively and $B^*_{zxs_0} = S^{*-1}_{xxs_0} S^*_{xzs_0}$. The approach taken so far is strongly model based and so model based estimators of μ_y and Σ_{zz} would be preferred. However, in practice the data available for use in the adjustment may comprise published p-weighted estimators of means and covariances obtained from the sample s_0, which is independent of s. Thus

$$E_{p_0}[\hat{\mu}_{zs_0} \mid z,c] = \bar{z}_u$$

$$E_{p_0}[\hat{\Sigma}_{zzs_0} \mid z,c] = S_{zzu}$$

where $\bar{z}_u$ and S_{zzu} are the mean vector and covariance matrix of the auxiliary variables in the finite population and E_{p_0} represents the expectation with respect to repeated sampling using the sampling design employed to obtain s_0, *i.e.*, the randomization distribution. Thus from (2.41) and (2.42)

$$E[\hat{\mu}_y(z) \mid s,z,c] = \mu_y + \beta_{yz'}(\bar{z}_u - \mu_z)$$

$$E[\hat{\Sigma}_{yy}(z) \mid s,z,c] = \Sigma_{yy} + \beta_{yz'}(S_{zzu} - \Sigma_{zz})\beta_{yz} + (\bar{n}^* - 1)\Delta_{yy.z} + 0(m^{-1}).$$

These expectations are taken over the statistical model generating the y values and the randomization distribution associated with s_0. In practice $\bar{z}_u$ and S_{zzu} will be very close to μ_z and Σ_{zz} respectively.

3. IDENTIFYING GROUPING VARIABLES

In the previous section we introduced a set of auxiliary variables, z, which characterised the area differences and which were used to adjust the aggregated analysis to reduce the aggregation bias. If the auxiliary variables were totally successful then $\Delta_{yy.z}$ would be reduced to zero and the adjustment method would remove the aggregation bias completely. In practice the auxiliary variables for which $\Delta_{yy.z} = \mathbf{0}$ are unknown. Also we will be restricted to sets of variables for which area level means are available as part of the data set under analysis and for which an estimate $\hat{\Sigma}_{zz}$ of the unit level covariance matrix is available. Basic demographic information and housing variables commonly available from the Census may be used. However these variables may not fully characterise the grouping process and so they may not explain as much of the between area difference as we might wish.

3.1 An Analysis Strategy

In practice the grouping variables will not be known. We need a strategy for identifying adjustment variables for which an estimate of the unit level covariance matrix is available and which account for group effects. One strategy involves the following steps:

1) Identify a set of variables that cover the same subject area as the variables of interest, but for which both area level and unit level data are available for some period in the past. Previous Census data may be suitable.

2) Add to this set, variables (such as demographic and housing variables) which are candidate z variables since they are known to be strongly associated with area differences. Estimates of both the area level and unit level covariance matrices must also be available for the same period in the past.

3) Carry out an analysis of these data to identify the variables which account most strongly for the area level effects among the variables of interest. This analysis, which we term a CGV analysis, will be described below.

4) Identify from (3) a set of adjustment variables which are available within the current data set and for which the current unit level covariance matrix is available from some source.

5) For some variables of interest it may be possible to obtain estimates of unit level variances or covariances, from published tables for example. From these calculate aggregation effects $\bar{Q}_{aa} = \bar{s}_{aa}/s_{aa}$ or $\bar{Q}_{ab} = \bar{s}_{ab}/s_{ab}$.

6) Use the variables identified in (4) to adjust the aggregate analysis for the variables of interest and check the adjusted aggregation effects corresponding to (5) to monitor the success of the adjustment.

3.2 The Ideal Grouping Variables

We first consider the ideal set of grouping variables that could be used for adjustment so as to identify the appropriate (CGV) analysis that could be followed for the analysis of aggregated data using the strategy outlined above.

Let us suppose that for the complete set of variables of interest we have the area level variance-covariance matrix $\bar{S}_{yy}$ and the unit level variance-covariance matrix S_{yys_1} based on a sample s_1. Of course if this occurred in practice the aggregation problem would disappear since we could discard $\bar{S}_{yy}$ and simply use S_{yys_1}, as an estimate of Σ_{yy}. However there are three reasons for considering this situation. Firstly it helps to throw light on the grouping structure which determines the relationship between $\bar{S}_{yy}$ and S_{yys_1}. Secondly it may be that $\bar{S}_{yy}$ and S_{yys_1} are available at some point in time such as census day but that further analysis of a new version of $\bar{S}_{yy}$ is to be based on inter-censal data when S_{yys_1} is unavailable. If the grouping structure persists over time, as we might expect, then the analysis of the census day versions of $\bar{S}_{yy}$ and S_{yys_1} might help the subsequent inter-censal analysis by identifying the key variables that explain a large proportion of the aggregation effects. These possibilities underpin the strategy outlined in section 3.1 above. Thirdly if the variables in y cover a large range of socio-economic and demographic variables, as occurs in the census, then the key variables that account for the grouping effects for the variables may also explain much of the grouping effects of other socio-economic and demographic variables. Note that the two samples s and s_1 may be identical but in general do not need to be. For example s may correspond to an administrative source which is effectively a census that provides aggregate data for geographic areas, and s_1 is a sample survey from which individual level data are made available without any geographic identifiers.

To help identify the important variables associated with the grouping Steel (1985) suggests that $\hat{\theta}_1, \ldots, \hat{\theta}_p$, the eigenvalues of $S_{yys_1}^{-1} \bar{S}_{yy}$, be calculated as well as the matrix $\hat{\boldsymbol{D}}_y = [\hat{d}_1, \ldots, \hat{d}_p]$ such that

$$\hat{\boldsymbol{D}}_y' \bar{S}_{yy} \hat{\boldsymbol{D}}_y = \operatorname{diag}(\hat{\theta}_k) \quad \text{and} \quad \hat{\boldsymbol{D}}_y' S_{yys_1} \hat{\boldsymbol{D}}_y = \boldsymbol{I}.$$

The variables defined by the transformation

$$\hat{\boldsymbol{u}}_i = \hat{\boldsymbol{D}}_y' \, \boldsymbol{y}_i$$

successively have maximum ratio of between group to sample total variance and have zero sample correlation at the unit and group level and unit level sample variance of 1. These variables are called the sample Canonical Grouping Variables (CGVs). The sample CGVs have the maximum intra-group correlation. Note that $\operatorname{tr}(S_{yys_1}^{-1} \bar{S}_{yy}) = \Sigma_k \hat{\theta}_k$ can be defined as the multivariate aggregation effect.

Note that the matrix $\hat{\boldsymbol{D}}_y$ will exist even if S_{yys_1} and $\bar{S}_{yy}$ are based on different samples so long as the former is positive definite and the latter is positive semi-definite. Furthermore the variances of the CGV's will be non-negative. However, when s and s_1 are distinct it is possible that the maximum variance of a CGV could exceed $(N-1)/(M-1)$ which is the maximum possible aggregation effect. In this case the CGV has an implied negative within group variance component. For our purposes this may not matter since we are interested in identifying important grouping variables but in principle the offending variance of the CGV could be set to its theoretical maximum. The sample CGVs are obtained from the eigenevectors of $A_{yy} = S_{yys_1}^{-1}\bar{S}_{yy}$. If s and s_1 are the same sample then $A_{y\bar{y}}$ is the sample regression coefficient for the regression of the group level means on the unit level values calculated over the unit level sample. In this case the sample CGVs are in fact the sample canonical variates relating the unit level and group level data and $\hat{\theta}_k$ are the sample canonical correlations.

Having calculated the CGVs the difference between the sample group level and unit level covariance matrix can be expressed as

$$\bar{S}_{yy} - S_{yys_1} = \sum_k (\hat{\theta}_k - 1)\hat{\phi}_k\hat{\phi}_k'$$

where $\hat{\phi}_k$ is the vector of sample covariances between the k-th CGV and the original variables. Hence the difference between the group level and unit level covariance matrix can be partitioned into k orthogonal elements, one for each CGV.

For the covariance between y_{ia} and y_{ib}, the difference between the sample group level covariance, $\bar{s}_{ab}$ and unit level covariance s_{ab} (where $\bar{s}_{ab}$ and s_{ab} elements of $\bar{S}_{yy}$ and S_{yys_1}, respectively) is

$$\bar{s}_{ab} = s_{ab} + (s_{aa}s_{bb})^{1/2}\sum_k (\hat{\theta}_k - 1)\hat{\rho}_{ak}\hat{\rho}_{bk}$$

where $\hat{\rho}_{ak} = \hat{\phi}_{ak}/s_{aa}^{1/2}$ is the sample correlation between the a-th variable and the k-th sample CGV.

If the first q sample CGVs are used to calculate an adjusted group level variance matrix, *i.e.*, $\hat{\boldsymbol{u}}_{qi} = \hat{\boldsymbol{D}}_q'\, y_i$ where $\hat{\boldsymbol{D}}_q = [\hat{d}_1, \ldots, \hat{d}_q]$, are used as the auxiliary variables

$$\hat{\Sigma}_{yy}(\hat{\boldsymbol{u}}_q) = \bar{S}_{yy} + \bar{\boldsymbol{B}}_{yu_q}'(S_{u_qu_qs_0} - \bar{S}_{u_qu_q})\bar{\boldsymbol{B}}_{yu_q}$$

then the first q terms of the decomposition are removed *i.e.*,

$$\hat{\Sigma}_{yy}(\hat{\boldsymbol{u}}_q) = S_{yys_1} + \sum_{k=q+1}^{p} (\hat{\theta}_k - 1)\hat{\phi}_k\hat{\phi}_k'$$

and $\mathrm{tr}(S_{yys_1}^{-1}\hat{\Sigma}_{yy}(\hat{\boldsymbol{u}}_q)) = \sum_{k=q+1}^{p}\hat{\theta}_k.$. In fact use of the first q CGVs provides the matrix of rank q that minimizes $\| S_{yys_1} - \hat{\Sigma}_{yy}(\hat{\boldsymbol{u}}_q) \|$. Hence by examining the quantities

$$\sum_{k=q+1}^{p}\hat{\theta}_k \quad \text{and} \quad 1 + \sum_{k=q+1}^{p}(\hat{\theta}_k - 1)\hat{\rho}_{ak}^2 \qquad \text{for} \quad q = 0, \ldots, p-1$$

it is possible to examine how the proportion of the overall aggregation effect and the aggregation effect for each variable can be explained by the first q sample CGVs.

The preceding analysis will suggest how many dimensions are required to effectively explain and hence remove a specified amount of the aggregation effects. Moreover by looking at the loadings of the original variables in the CGVs, it should be possible to identify which variables play the major role in "explaining" the aggregation effects of the other variables. It is these variables that researchers should concentrate on obtaining unit level data for, to use in the adjusted estimator.

These results have some important implications for the use of group level data supplemented by limited unit level data, since they open the way to combining sample survey data and group level data from one or more sources and suggest a strategy for the analysis of group effects and group level data.

4. SOME EMPIRICAL RESULTS

We illustrate the ideas of the previous sections with an analysis of the 1991 UK population census data for the Local Authority District (LAD) of Reigate, Banstead and Tandridge. The LAD population is 188,700 people contained in 371 EDs giving an average number of people per ED of $\bar{n} = 508.6$. Group level data are available on a complete count basis for each ED in the LAD from the Small Area Statistics (SAS) data file. Corresponding unit level data for the LAD are obtained from a 2 per cent Sample of Anonymized Records of individuals (SAR). The records in the SAR cannot be identified with any specific ED within the LAD thus in this situation we have $\bar{S}_{yy}$ based upon complete data for each ED from the SAS and we have an estimate of $\bar{S}_{yys_1}$ based on a 2 percent sample from the SAR. The following analysis is based upon 16 census variables for each person.

For each variable the group level data and the unit level data were used to calculate the aggregation effect, $\bar{Q}_a = \bar{s}_{aa}/s_{aas}$. The parameter $\delta_{aa} = \Delta_{aa}/\Sigma_{aa}$, defined on the appropriate diagonal elements of Δ_{yy} and Σ_{yy} is the intra-group correlation for the a-th variable. An estimate $\hat{\delta}_{aa}$ of the intra-group correlation can be obtained from (2.18) since $\bar{Q}_a = 1 + (\bar{n}^* - 1)\,\hat{\delta}_{aa}$. The results for the variables are given in Table 1. The intra-group

Table 1

Aggregation Effects and Intra-class Correlations for Census Variables in Reigate LAD

	Aggregation Effect	Intra-class Correlation
Persons aged 18-29	9.20	.016
Persons aged 30-44	4.56	.007
Persons aged 45-59*	5.97	.010
Persons aged 60 and over*	17.17	.032
Female	1.08	.000
Non-white*	8.29	.014
Married	6.24	.010
Limiting long term illness	7.24	.012
Persons employed full time	8.55	.015
Persons unemployed	2.27	.003
Other employment status	11.19	.020
Head of h'hold born UK	4.48	.007
Head of h'hold born New Commonwealth	3.59	.005
Migrant head of household	9.04	.016
$\leq$ 1.5 persons per room: density	27.96	.053
Persons in 0 car households	32.98	.063

* Selected for adjustment variables.
Source: Reigate and Banstead; Tandridge LAD 1991 census data.

correlations are generally small but the number of observations in each ED implies that the aggregation effects can be high (see the comment following equation (2.18)).

Figure 1a shows a plot of the group level correlation, $\bar{r}_{ab}$, against the individual level correlation, r_{ab}, for every pair of variables. Note the strong aggregation effects which are revealed through the characteristic S-shaped plot. Small correlations at the unit level are generally magnified so that for most cases $|\ \bar{r}_{ab}\ |$ is much larger than $|\ r_{ab}\ |$.

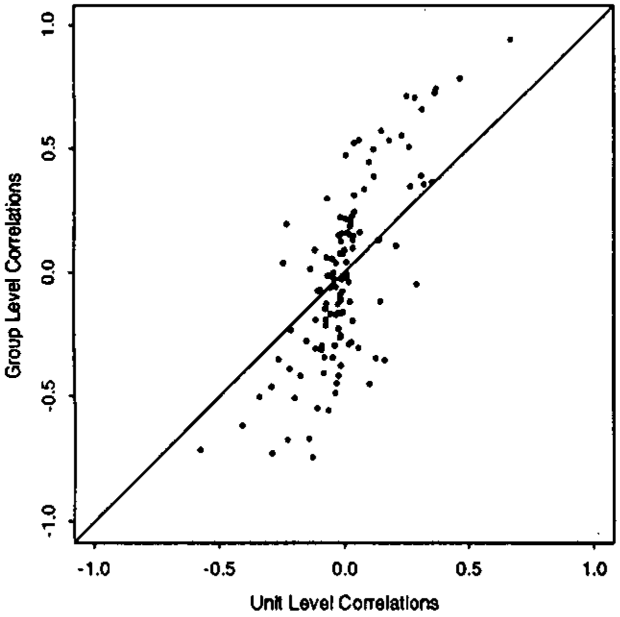

Figure 1a.

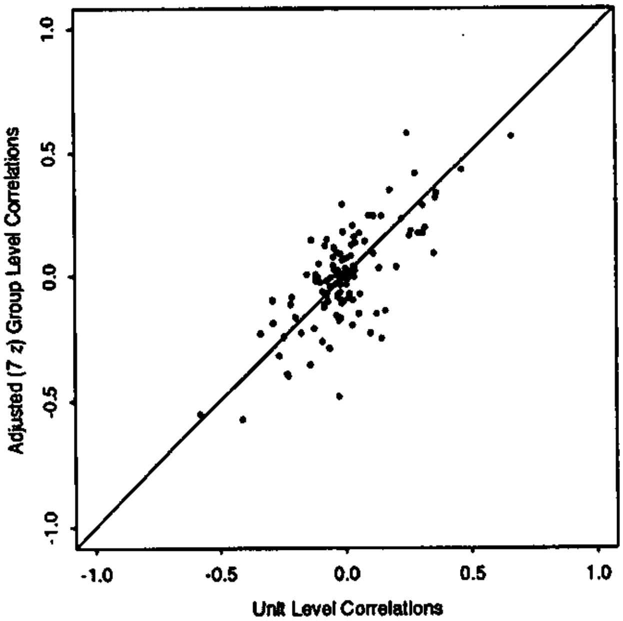

Figure 1b.

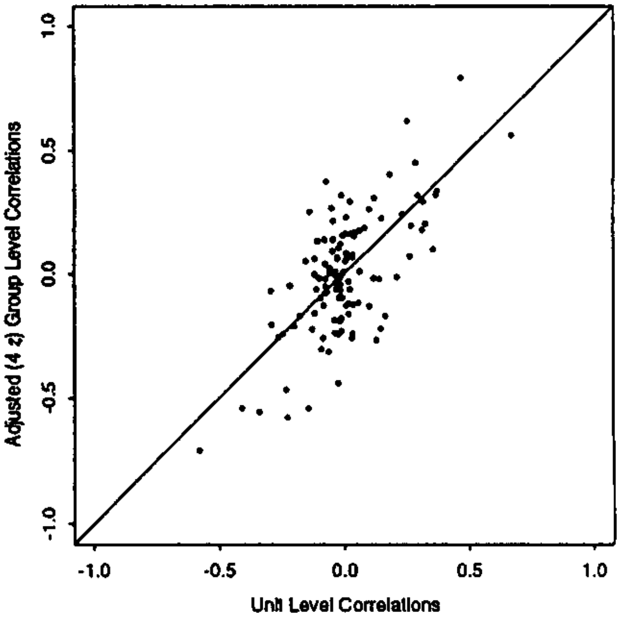

Figure 1c.

Since in this case we have $\bar{S}_{yy}$ and S_{yys_1} we may carry out a canonical grouping variable analysis so as to understand the more important features of the grouping structure. Table 2 shows the loadings on the 16 variables for the first five canonical grouping variables which together account for 89% of the multivariate aggregation effect.

The first CGV has high loadings on high density occupation and car (*i.e.*, auto) access and might be interpreted as a socio-economic factor. The second CGV has high loadings the variables indicating people in the two oldest age groups. It is noticeable, also, that the proportion of

Table 2

First Five CGV's for Variables in Table 1

	CGV1	CGV2	CGV3	CGV4	CGV5
Persons aged 18-29	0.4	0.3	0.9	1.1	0.1
Persons aged 30-44	0.1	0.5	0.36	1.0	0.2
Persons aged 45-59*	−0.1	1.2	−0.2	1.0	0.1
Persons aged 60 and over*	0.3	2.2	−0.5	2.6	0.9
Female	0.1	0.0	0.0	0.3	0.1
Non-white*	0.5	−0.4	1.4	−1.1	5.2
Married	−0.2	−0.5	−0.4	−0.8	−0.1
Limiting long term illness	0.3	0.1	−0.2	0.2	0.3
Persons employed full time	0.7	−0.3	0.2	1.2	0.4
Persons unemployed	0.7	0.0	−0.1	0.0	−0.4
Other employment status	0.1	0.1	0.0	−0.2	−0.1
Head of h'hold born UK	0.5	−0.1	−1.0	0.4	0.2
Head of h'hold born New Commonwealth	0.0	−0.1	−0.3	0.1	0.6
Migrant head of household	0.2	0.1	1.4	0.6	−1.3
≤ 0.5 persons per room	−1.4	0.3	1.2	−0.7	−0.2
Persons in 0 car households	2.2	0.6	0.8	−1.9	−0.7

* Selected for adjustment variables.
Source: Reigate and Banstead; Tandridge LAD 1991 census data.

non-white heads of household contributes to the later CGV's. As might be expected, variables such as proportion Female, that exhibit almost no intra-group correlation and hence no aggregation effect make virtually no contribution to the CGV's. Such variables do not vary across areas and hence generally have no explanatory power.

In usual practice a CGV analysis will not be possible since if S_{yy} was available there would usually be no need to carry out an aggregate analysis. However the CGV analysis suggests variables that may be important since they load highly on the first few CGVs.

It is well known in the UK context that housing tenure variables (which are not contained in the 16 variables of interest) have a powerful association with a wide variety of socio-economic, attitudinal and health variables. There are strong reasons for assuming that using these as auxiliary, z, variables for adjustment would account for a substantial proportion of the first socio-economic dimension and may act in place of the density of occupancy and car access variables that are seen to be important for the first CGV. The other reason for considering those variables is that if the present analysis is to act as an illustration of what might be achieved in other situations then basic tenure and housing variables are more likely to be available as adjustment variables than density of occupation and car access. In the light of the CGV analysis and in the spirit of identifying a small number of adjustment variables which could be expected to be available in many situations, we identify a set of seven potential adjustment variables. These are the three variables of interest identified in Table 1 identified by an asterisk (Age 45-59, Age 60+, non-white) and the four housing variables listed in Table 3 together with their aggregation effects and intra-cluster correlations.

Table 3

Aggregation Effects and Intra-class Correlations for Household Level Variables in Reigate LAD

Variable		Aggregation Effect	Intra-class Correlation
Tenure:	LA Rented	133.43	0.261
	Owner Occupier	90.83	0.177
Stock:	Det/semi/terrace	90.03	0.175
	Good Amenities	59.52	0.113

Source: Reigate and Banstead; Tandridge LAD 1991 census data.

In what follows the group level covariance matrix for the original 16 variables will be adjusted by the unit level covariance matrix for 7 z-variables (three of the basic demographic variables in the original set and four household variables).

Two overall measures of the effectiveness of the adjustment were calculated. The first is

$$1 - \frac{\operatorname{tr}\left(S_{yys_1}^{-1}\,\hat{\Sigma}_{yy}\,(z)\right) - 1}{\operatorname{tr}\left(S_{yys_1}^{-1}\,\bar{S}_{yy}\right) - 1}$$

which is the reduction in the multivariate aggregation effect and the second is

$$\frac{\| S_{yys_1} - \bar{S}_{yy} \| - \| S_{yys_1} - \hat{\Sigma}_{yy}(z) \|}{\| S_{yys_1} - \bar{S}_{yy} \|}$$

which shows the reduction in the generalised distance between the unit level and group level covariance matrices before and after adjustment.

Table 4

Z-variable Combination	No. of Variables	% reduction in	
		Multivariate Aggregation Effect	Generalised Distance
60+	1	16	24
45-59, 60+	2	38	53
Tenure	2	30	21
Stock	2	31	19
45-59, 6− +, NW	3	44	54
45-59, 60+, tenure	4	57	71
45-59, 60+, stock	4	57	69
45-59, 60+, tenure, NW	5	63	72
45-59, 60+, stock, NW	5	62	70
45-59, 60+, stock, tenure, NW	7	68	75

Table 4 shows the effect of using various combinations of variables for adjustment of the aggregated analysis. The two age variables are clearly important (accounting for 38% of the multivariate aggregation effect and 53% of the generalized distance) but the Tenure or Housing Stock variables are also important. When Tenure or Housing Stock are used in conjunction with age the percentage reduction in either measure is close to the sum of the effects of the variables separately showing that age and Tenure or Housing Stock are acting as distinct adjustment variables. Obviously the greatest success is achieved by including all 7 adjustment variables and accounts for 68% and 75% respectively of the two aggregation measures.

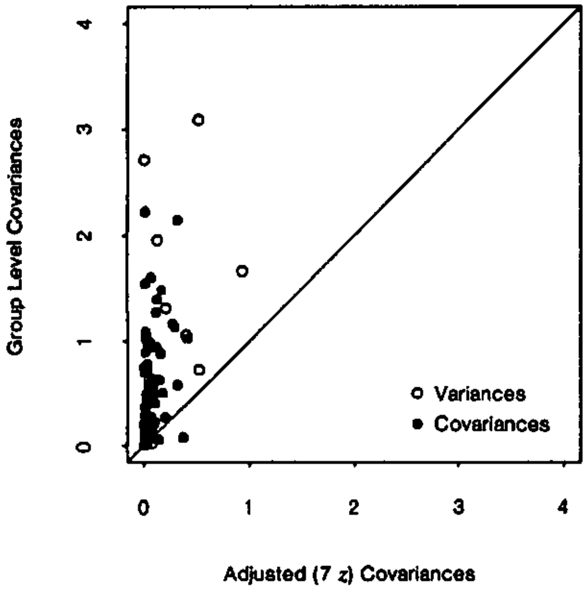

Figure 2a.

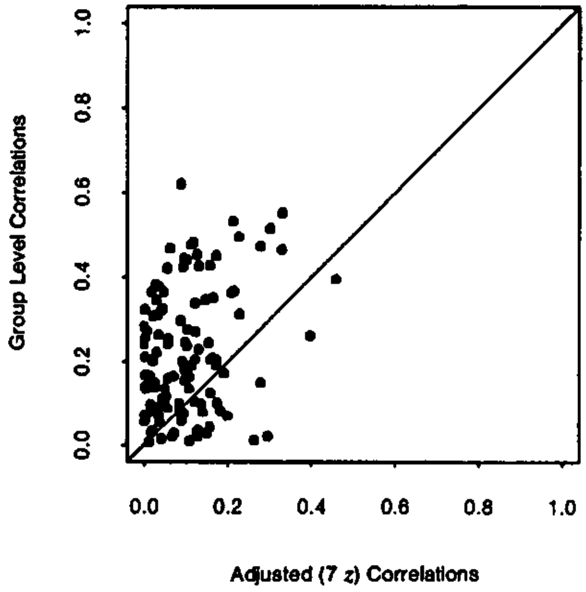

Figure 2b.

These results show that around 70% of the aggregation effects have been removed by the adjustment. Figures 2a and 2b show the effect of adjustment by these variables. In Figure 2a the vertical axis contains $|\bar{s}_{ab} - s_{abs_1}|$, the absolute bias for the group level covariance for each pair of variables. The horizontal axis contains $|\hat{\Sigma}_{ab}(z) - s_{abs_1}|$ the absolute bias of the adjusted estimator. The hollow symbol is used for variances of the y variables, and the solid symbol is used for covariances. Almost all of the plotted values show that the biases after adjustment are smaller (often much smaller) than the original bias. In almost all cases the adjustment has had a substantial improvement. Figure 2b shows the corresponding plot for correlations rather than covariances. (Correlations of y_a, y_a have obviously been omitted from this plot.) Again there is a strong improvement with the residual bias after adjustment being much smaller than the original bias for the group level analysis. The results are not as successful as for the covariances, since in some cases small biases for the group level analysis have been made worse. In this case the adjustments are applied to the covariance and the two variances used in each correlation coefficient. There is more potential for the relative changes in each component to lead to a correlation which is worse than the original. However, almost all of the large biases at the group level have been improved.

Figure 1b shows the plot of the adjusted group level correlations, $\bar{r}_{ab}(z)$, obtained from $\hat{\Sigma}_{yy(z)}$ against the unit level correlations and can be compared with the original unadjusted plot in Figure 1a. The characteristic S-shaped curve shown in Figure 1a has been replaced by a plot of points which lie about the line $\bar{r}_{ab}(z) = r_{ab}$ as we would want if aggregation bias is removed.

Figures 1b, 2a and 2b show that a substantial reduction to the aggregation effect can be achieved by using 4 housing variables and 3 of the original y variables. This implies adjusting the original 120 variances and covariances in the 16 × 16 matrix by 21 variances and covariances for the z variables. As an illustration of what might be achieved with minimal information we reduce the adjustment variables to the four involving age and Tenure. From Table 4 we see that these account for 57% and 71% of the two measures of aggregation. Figures 3a and 3b show the corresponding plots to Figures 2a and 2b for this case. Figure 1c shows the plot of the adjusted correlations using 4 variables against the individual level correlations. Obviously the adjustment is not as successful but it is encouraging to see what can be achieved with so few adjustment variables. As a further measure of the effect of the adjustment the median absolute difference between $\bar{r}_{ab}$ and r_{ab} was 0.186. After adjusting by 4 variables this was reduced to 0.126 and after adjusting 7 variables to 0.090. The corresponding median values for $|\bar{s}_{ab} - s_{ab}|$ were 0.173, 0.039 and 0.017 respectively.

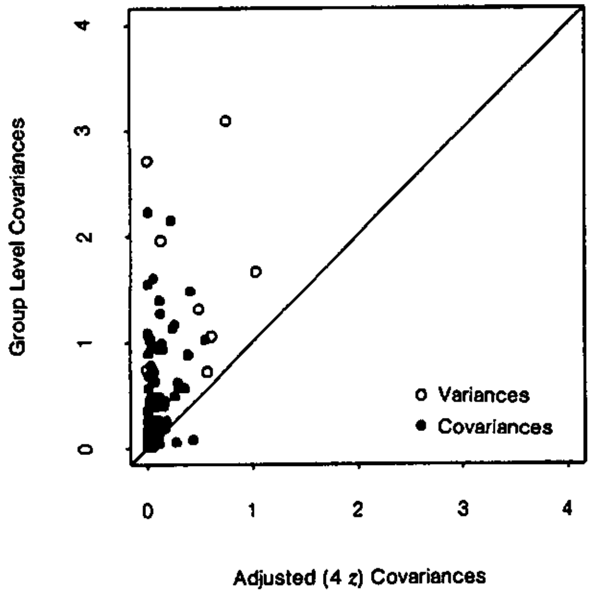

Figure 3a.

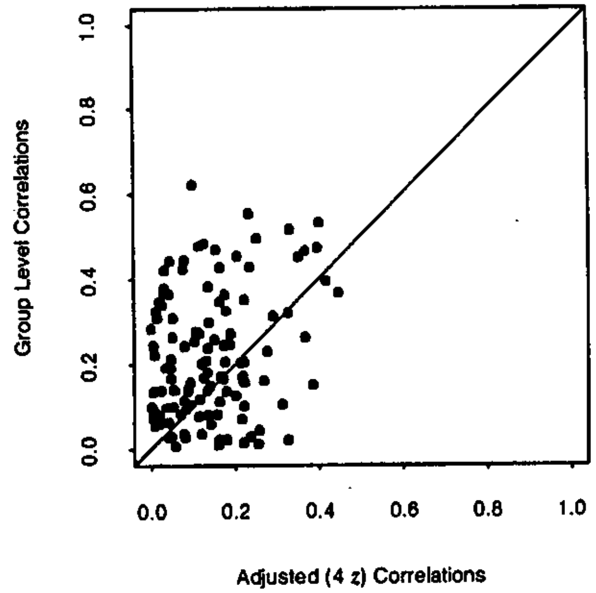

Figure 3b.

5. CONCLUSIONS AND DISCUSSION

A model for grouped populations has been proposed which leads to a decomposition of the bias observed in group level analysis based on covariance matrices into two components. The first component is due to the grouping variables and the second is due to the residual intra-group correlations between the y variables given the grouping variables z. This decomposition provides an understanding of the magnitude of aggregation effects. It also provides a way of removing the bias due to the grouping variables if additional information about the unit level covariance matrix of the grouping variables is available.

In many countries there are many group level data available at different levels of aggregation from the census and many other sources. The development of Geographic Information Systems will increase the availability of such data. It is important to analyse and decompose the group effects and the theory developed and the strategy proposed here provide a framework for achieving this. A proper understanding of which variables explain most of the group effects, and therefore should be used in adjusting ecological analyses, will open the way to making use of aggregated data.

ACKNOWLEDGEMENTS

This research was supported by Grant Number H507 26 5013 from the Economic and Social Research Council, United Kingdom. The authors also acknowledge gratefully helpful comments from the associate editor and referees.

REFERENCES

ARBIA, G. (1989). *Spatial Data Configuration in Statistical Analysis of Regional Economic and Related Problems.* Dordrecht: Kluman.

BLALOCK, H.M. (1964). *Causal Inference in Nonexperimental Research.* Chapel Hill NC: University of North Carolina Press.

BLALOCK, H.M. (1979). Measurement and conceptualization problems: The major obstacle to integrating theory and research. *American Sociological Review,* 44, 881-894.

BLALOCK, H.M. (1985). Cross level analysis. In *The Collection and Analysis of Community Data,* (Ed. J.B. Casterlin), ISI, World Fertility Survey.

CLARK, W.A.V., and AVERY, K.L. (1976). The effect of data aggregation in statistical analysis. *Geographical Analysis,* 8, 428-438.

DUNCAN, D.P., and DAVIS, B. (1953). An alternative to ecological correlation. *American Sociological Review,* 18, 665-666.

FOTHERINGHAM, A.S., and WONG, D.W.S. (1991). The modifiable areal unit problem in multivariate statistical analysis. *Environment and Planning,* A, 23, 1025-1044.

GEHLKE, C.E., and BIEHL, K. (1934). Certain effects of grouping upon the size of the correlation coefficient in census tract material. *Journal of the American Statistical Association,* 29, Supplement, 169-170.

GOODMAN, L.A. (1959). Some alternatives to ecological correlation. *American Journal of Sociology,* 64, 610-625.

HANNAN, M.T., and BUSTEIN, L. (1974). Estimation from grouped observations. *American Sociological Review,* 39, 374-392.

HOLT, D., SMITH, T.M.F., and WINTER, P.D. (1980). Regression analysis of data from complex surveys. *Journal of the Royal Statistical Society*, A, 143, 474-87.

HOLT, D., and SCOTT, A.J. (1981). Regression analysis using survey data. *The Statistician*, 30, 169-173.

LICHTMAN, A.J. (1974). Correlation, regression, and the ecological fallacy: A critique. *Journal of Interdisciplinary History*, 4, 417-433.

LANGBEIN, L.I., and LICHTMAN, A.J. (1978). *Ecological Inference*. Thousand Oaks, CA: Sage.

OPENSHAW, S. (1984). Ecological fallacies and the analysis of areal census data. *Environment and Planning*, A, 6, 17-31.

OPENSHAW, S., and TAYLOR, P.J. (1979). A million or so correlation coefficients: three experiments on the modifiable areal unit problem. In *Statistical Applications in the Spatial Sciences*, (Ed. N. Wrigley), 127-144.

PEARSON, K. (1903). On the influence of natural selection on the variability and correlation of organs. *Philosophical Transactions of the Royal Society*, A, 200, 1-66.

PERLE, E.D. (1977). Scale changes and impacts on factorial ecology structures. *Environment and Planning*, A, 9, 549-558.

RAO, C.R. (1973). *Linear Statistical Inference and its Applications*, (2nd Ed.). New York: Wiley.

ROBINSON, W.S. (1950). Ecological correlations and the behaviour of individuals. *American Sociological Review*, 15, 351-357.

SMITH, K.W. (1977). Another look at the clustering perspective on aggregation problems. *Sociological Methods and Research*, 5, 289-316.

SMITH, T.M.F., and HOLMES, D. (1989). Multivariate analysis. In *Analysis of Complex Surveys*, (Eds. C.J. Skinner, D. Holt and T.M.F. Smith), 165-187.

STEEL, D. (1985). Statistical Analysis of Populations with Group Structure. Unpublished PhD Thesis, Department of Social Statistics, University of Southampton.

STEEL, D., and HOLT, D. (1994). Analysing and Adjusting Aggregation Effects: The Ecological Fallacy Revisited. Department of Applied Statistics, University of Wollongong, Preprint 1/94.

STEEL, D., and HOLT, D. (1995). Rules for random aggregation. *Environment and Planning* (to appear).

YULE, U., and KENDALL, M.S. (1950). *An Introduction to the Theory of Statistics*. Glendale, CA: Griffin.

Survey Methodology, June 1996
Vol. 22, No. 1, pp. 17–22
Statistics Canada

Linearization Methods for Single Phase and Two-Phase Samples: A Cookbook Approach

DAVID A. BINDER[1]

ABSTRACT

There are a number of asymptotically equivalent procedures for deriving the Taylor series approximation of variances for complex statistics. In Binder and Patak (1994) the theoretical justification for one class of methods was derived. However, many of these methods can be derived for practical examples using straightforward techniques that are not clearly described in Binder and Patak. In this paper we give a "cookbook" approach that can be used for many examples, and that has been shown to have good finite sample properties. Normally the method of choice becomes clear through arguments such as model-assisted methods or linearizing the jackknife; however, using our approach yields the desired results more directly. As well, we present new results on the application of these techniques to two-phase samples.

KEY WORDS: Complex surveys; Variance estimation; Ratio estimator; Regression estimator; Wilcoxon rank sum test; Estimating equations.

1. THE METHOD

The derivation of the asymptotic variance for a wide class of estimators from complex survey samples is now well established in the literature, at least to a first order approximation. However, there are a number of competing estimators of the variance, all of which are asymptotically equivalent. In this paper, we discuss a simple derivation of one of the most favoured of these estimators in a general setting. This simple derivation is useful for practitioners, who may be baffled by the choices available, and need a quick solution to the problem.

We start with a simple example of the approach using the ratio estimator of a population total. Here the estimator is

$$\hat{Y}_R = \hat{R}X, \qquad (1)$$

for

$$\hat{R} = \hat{Y}/\hat{X}, \quad \text{and} \quad \hat{Y} = \sum_{k \in s} w_k y_k,$$

where, s is the set of indices corresponding to sampled units and w_k is the sampling weight, normalized so that $\sum w_k$ is an estimator of the population total; *e.g.*, $w_k = 1/\pi_k$, where π_k is the first order inclusion probability. The definition of $\hat{X}$ is analogous to that of $\hat{Y}$. Applying total differentials to both sides of (1), we obtain

$$(d\hat{Y}_R) = (d\hat{R})X, \qquad (2a)$$

where

$$(d\hat{R}) = \frac{(d\hat{Y})}{\hat{X}} - \frac{\hat{Y}}{\hat{X}^2}(d\hat{X}) \qquad (2b)$$

$$= \frac{1}{\hat{X}}[(d\hat{Y}) - \hat{R}(d\hat{X})].$$

We note that, in general, the total differential for $\hat{T} = g(\hat{Y}_1, \ldots, \hat{Y}_m)$ is given by

$$(d\hat{T}) = \sum \left[\frac{\partial g(\hat{Y})}{\partial \hat{Y}_i}\right](d\hat{Y}_i).$$

Although we could have avoided using $\hat{R}$ in (1) by simply defining

$$Y_{\hat{R}} = \frac{\hat{Y}}{\hat{X}}X,$$

thus removing the need for explicitly defining $(d\hat{R})$ in (2b), we did so to make the more complex examples, to be given in Section 1.2, clearer. We also note that (2a) does not include the total differential of X, the population total of the x-variable, since X is assumed to be fixed and known.

The next step is to replace all total differentials of estimated quantities by deviations from the their respective expected values. On the right hand side, we substitute for $(d\hat{Y})$ the expression $(\sum w_k y_k - Y)$, and so on. For the quantity of interest, $\hat{Y}_R$, we replace $d\hat{Y}_R$ by $\hat{Y}_R - Y$. From (2), performing this step, yields

$$\hat{Y}_R - Y \doteq \frac{X}{\hat{X}}\left[\left(\sum w_k y_k - Y\right) - \hat{R}\left(\sum w_k x_k - X\right)\right]. \qquad (3)$$

[1] David A. Binder, Director, Business Survey Methods Division, Statistics Canada, R.H. Coats Building, 11 "A", Ottawa, Ontario, Canada, K1A 0T6.

We see that this expression contains a number of weighted estimators - those that explicitly show their dependence on the w_k's, ($\sum w_k y_k$ and $\sum w_k x_k$) and those where the w_k's are implicit in the expression ($\hat{X}$ and $\hat{R}$).

For the last step, we isolate z_k, defined by rewriting (3) as

$$\hat{Y}_R - Y \doteq \sum w_k z_k + \text{other terms not depending } \textit{explicitly} \text{ on } w_k.$$

Here, we obtain

$$z_k = \frac{X}{\hat{X}}(y_k - \hat{R}x_k). \tag{4}$$

The justification for ignoring the terms not depending explicitly on w_k will be given in Section 4. Note that $\sum w_k z_k$ has the form of the estimate of the population total of the variable z.

Now to obtain the variance of $\hat{Y}_R$, we insert the new variable z_k into the k-th sample record, and use a standard procedure for estimating the variance of a total, applied to this variable. It is assumed that a variance estimator with good properties is available for the sample design under consideration.

A summary of the method in general is the following:

1. We let the estimator of T be $\hat{T}$ and take its total differential. We assume that $\hat{T}$ is asymptotically design consistent.
2. We replace total differential of $\hat{T}$, $d\hat{T}$, by $\hat{T} - T$. We replace all other total differentials of estimated quantities by the deviation from their respective expected values, where we substitute for $(d\hat{Y})$ the expression $(\sum w_k y_k - Y)$, and so on.
3. The last step is to isolate z_k, when we rewrite the result of Step 2 as

$$\hat{T} - T \doteq \sum w_k z_k + \text{other terms not depending } \textit{explicitly} \text{ on } w_k.$$

4. Finally, to obtain the estimated variance of $\hat{T}$, we insert the new variable z_k into each sampled record, and use the standard procedure (known to have good properties) for estimating the variance of a total, applied to this variable.

1.1 Simplest General Case

For one-phase samples, a simple general case is where the estimator can be expressed as a differentiable function of the estimated totals for certain survey variables, some of which may be derived variables at the final sampling unit level. In this case our approach gives:

$$\hat{T} = g(\hat{Y}_1, \ldots, \hat{Y}_m)$$

$$(d\hat{T}) = \sum \left[\frac{\partial g(\hat{Y})}{\partial \hat{Y}_i}\right](d\hat{Y}_i)$$

$$\begin{aligned}\hat{T} - T &\doteq \sum_i \left[\frac{\partial g(\hat{Y})}{\partial \hat{Y}_i}\right]\left(\sum_k w_k y_{ik} - Y_i\right) \\ &= \sum w_k z_k + \ldots,\end{aligned} \tag{5}$$

where

$$z_k = \sum_i \left[\frac{\partial g(\hat{Y})}{\partial \hat{Y}_i}\right] y_{ik} = \left[\frac{\partial g(\hat{Y})}{\partial \hat{Y}}\right]' y_k. \tag{6}$$

In what way is this formulation different from standard Taylor methods? The main difference is how expression (5) is treated. In standard methods, the partial derivatives are evaluated at their expected values before z_k is derived. Then, for those components of z_k that are unknown, an estimator is substituted. For the ratio estimator, (1), this would result in $X/\hat{X}$ disappearing from z_k in (4), since when $\hat{X}$ is replaced by its expected value, $X/\hat{X}$ becomes unity. The $\hat{R}$ remains in the expression, as it is used to estimate R, which is needed in the usual derivation of z_k.

Kott (1990) argues that the variance estimator for the ratio which we have derived has good conditional properties compared to the estimator which leaves out the factor $X/\hat{X}$. A number of others have come to similar conclusions. Rao (1995) showed that the method agrees with that obtained from the linearized jackknife. Our conjecture is that since the partial derivatives in expression (5) are evaluated at $\hat{Y}$ rather than Y, the linearization is "closer" to the original statistic, $\hat{T}$, so that the resulting variances have better properties. This is, of course, not a technical statement, but rather an intuitive justification of the method.

We note that in expression (6) for z_k, all the terms are directly observed from the sample, so that no substitution of estimators for unknown quantities is needed.

1.2 The Case with Extra Parameters

For many examples, the estimator is most easily defined in terms that include the use of parameters that are only used to simplify the definition of the parameter of interest. For the ratio estimator, $\hat{R}$ is an example of such an *extra parameter*. In this case, an explicit equation for the estimator of the extra parameter is available. The general method in the presence of extra parameters may be written as:

$$\hat{T} = g_1(\hat{Y}_1, \ldots, \hat{Y}_m, \hat{\lambda}), \quad \text{where} \quad \hat{\lambda} = g_2(\hat{Y}_1, \ldots, \hat{Y}_m),$$

$$(d\hat{T}) = \sum \left[\frac{\partial g_1(\hat{Y}, \hat{\lambda})}{\partial \hat{Y}_i}\right](d\hat{Y}_i) + \sum \left[\frac{\partial g_1(\hat{Y}, \hat{\lambda})}{\partial \hat{\lambda}_j}\right](d\hat{\lambda}_j),$$

where

$$(d\hat{\lambda}_j) = \sum_i \left[\frac{\partial g_{2j}(\hat{Y})}{\partial \hat{Y}_i}\right](d\hat{Y}_i),$$

$$\hat{T} - T \doteq \sum \frac{\partial g_1(\hat{Y},\hat{\lambda})}{\partial \hat{Y}_i}\left(\sum_k w_k y_{ik} - Y_i\right)$$

$$+ \sum \frac{\partial g_1(\hat{Y},\hat{\lambda})}{\partial \hat{\lambda}_j} \sum_i \frac{\partial g_{2j}(\hat{Y})}{\partial \hat{Y}_i}\left(\sum_k w_k y_{ik} - Y_i\right)$$

$$= \sum w_k z_k + \ldots,$$

where

$$z_k = \left[\frac{\partial g_1(\hat{Y},\hat{\lambda})}{\partial \hat{Y}}\right]' y_k + \left[\frac{\partial g_1(\hat{Y},\hat{\lambda})}{\partial \hat{\lambda}}\right]' \left[\frac{\partial g_2(\hat{Y})}{\partial \hat{Y}}\right] y_k. \quad (7)$$

For the case where the extra parameters are defined only implicitly through estimating equations, we have the following generalization:

$$\hat{T} = g(\hat{Y}_1, \ldots, \hat{Y}_m, \hat{\lambda}),$$

where

$$\hat{U}(\hat{Y}_1, \ldots, \hat{Y}_m, \hat{\lambda}) = 0. \quad (8)$$

$$(d\hat{T}) = \sum \left[\frac{\partial g(\hat{Y},\hat{\lambda})}{\partial \hat{Y}_i}\right](d\hat{Y}_i) + \left[\frac{\partial g(\hat{Y},\hat{\lambda})}{\partial \hat{\lambda}}\right]'(d\hat{\lambda}),$$

where by taking the total differential of (8) and isolating $(d\hat{\lambda})$, we have

$$(d\hat{\lambda}) = -\left[\frac{\partial \hat{U}(\hat{Y},\hat{\lambda})}{\partial \hat{\lambda}}\right]^{-1} \sum \left[\frac{\partial \hat{U}(\hat{Y},\hat{\lambda})}{\partial \hat{Y}_i}\right](d\hat{Y}_i). \quad (9)$$

$$\hat{T} - T \doteq \sum_i \left(\frac{\partial g}{\partial \hat{Y}_i}\right)\left(\sum_k w_k y_{ik} - Y_i\right)$$

$$- \left(\frac{\partial g}{\partial \hat{\lambda}}\right)' \left[\frac{\partial \hat{U}}{\partial \hat{\lambda}}\right]^{-1} \sum_i \left(\frac{\partial \hat{U}}{\partial \hat{Y}_i}\right)\left(\sum_k w_k y_{ik} - Y_i\right)$$

$$= \sum w_k z_k + \ldots,$$

where

$$z_k = \left[\frac{\partial g}{\partial \hat{Y}}\right]' y_k - \left[\frac{\partial g}{\partial \hat{\lambda}}\right]' \left[\frac{\partial \hat{U}}{\partial \hat{\lambda}}\right]^{-1} \left[\frac{\partial \hat{U}}{\partial \hat{Y}}\right]' y_k. \quad (10)$$

We see, of course, that (10) is a generalization of the previous forms for z_k given in (6) and (7).

2. OTHER EXAMPLES

Expressions (6), (7) and (10) above are displayed only for the purpose of giving the specific formulae for the various cases. However, in practice, we recommend using the basic steps from first principles. To demonstrate this, we give two examples: one is the familiar Generalized Regression Estimator (GREG); the other gives some new results for the Wilcoxon Rank Sum Test statistic for data from complex surveys.

2.1 Generalized Regression Estimator

The usual Generalized Regression Estimator, given, for example, in Särndal, Swensson and Wretman (1989), may be written as

$$\hat{Y}_{GREG} = \hat{Y} + \hat{\beta}'(X - \hat{X}), \quad (11)$$

where the extra parameter $\hat{\beta}$ is defined as the solution to

$$\sum_k w_k x_k (y_k - x_k'\hat{\beta})/c_k = 0,$$

where c_k is the factor to allow for heteroscedastic variance in the regression model. This is equivalent to

$$\hat{S}_{xx}\hat{\beta} - \hat{S}_{xy} = 0, \quad (12)$$

with obvious definitions for $\hat{S}_{xx}$ and $\hat{S}_{xy}$. Taking total differentials in (12) we get

$$(d\hat{S}_{xx})\hat{\beta} + \hat{S}_{xx}(d\hat{\beta}) - (d\hat{S}_{xy}) = 0,$$

so that

$$(d\hat{\beta}) = \hat{S}_{xx}^{-1}[(d\hat{S}_{xy}) - (d\hat{S}_{xx})\hat{\beta}].$$

Therefore, we have

$$\hat{\beta} - \beta \doteq \sum w_k \hat{S}_{xx}^{-1}[x_k(y_k - x_k'\hat{\beta})]/c_k + \ldots .$$

Now, taking total differentials of (11), we have

$$(d\hat{Y}_{GREG}) = (d\hat{Y}) - \hat{\beta}'(d\hat{X}) + (d\hat{\beta})'(X - \hat{X})$$

$$= (d\hat{Y}) - \hat{\beta}'(d\hat{X}) +$$

$$[(d\hat{S}_{xy}') - \hat{\beta}'(d\hat{S}_{xx})]S_{xx}^{-1}(X - \hat{X}).$$

After some algebraic manipulation, we obtain

$$\hat{Y}_{GREG} - Y = \sum w_k e_k [1 + x_k' \hat{S}_{xx}^{-1}(X - \hat{X})/c_k] + \ldots,$$

where $e_k = y_k - x_k' \hat{\beta}$. We, therefore, define

$$z_k = e_k[1 + x_k' \hat{S}_{xx}^{-1}(X - \hat{X})/c_k].$$

Taking the variance of the estimated total of this z-variable is identical to the variance proposed in Särndal, Swensson and Wretman (1989). There, it is argued on the basis of the validity of the regression model, that this variance is preferred to other Taylor expansion estimators for the variance. We see that the derivation of this z-variable is natural in our approach.

2.2 Wilcoxon Rank Sum Statistic

We now show how our method works in the case of a more difficult non-standard case. We assume that our sampled units belong to one of two subpopulations which we name Population 1 and Population 2. We define

$$I\{x \le y\} = \begin{cases} 1 \text{ if } x \le y, \\ 0 \text{ otherwise,} \end{cases} \text{ and } \delta_k = \begin{cases} 1 \text{ if } k \in \text{Pop. } 1 \\ 0 \text{ otherwise.} \end{cases}$$

We let

$$\hat{N}_1(t) = \sum_{k \in s} w_k \delta_k I\{x_k \le t\},$$

which corresponds to the estimated number of Population 1 units that have values less than or equal to t. We define $\hat{N}_2(t)$ analogously. We denote $\hat{N}_j = \hat{N}_j(\infty)$, the estimated number of units in Population j. Now a weighted version of the Wilcoxon Rank Sum Test statistic is

$$\hat{T}_W = \int_0^\infty [\hat{N}_1(t) + \hat{N}_2(t)] d\hat{N}_1(t). \qquad (13)$$

This corresponds to the weighted sum of the ranks from Population 1 among the weighted ranks of the combined sample. To derive the asymptotic expected value of $\hat{T}_W$ in (13), we let $N_i(t) = E[\hat{N}_i(t)]$ for $i = 1, 2$, and substitute $N_i(t)$ for $\hat{N}_i(t)$ in (13). We then define $F_i(t) = N_i(t)/N_i$, where $N_i = E(\hat{N}_i)$ and we give the null hypothesis as $F_1(t) = F_2(t) = F(t)$, say. This results in the asymptotic expectation being

$$\int_0^1 (N_1 + N_2) F(t) N_1 dF(t) = N_1(N_1 + N_2)/2.$$

Note that in the case of independent samples of size N_1 and N_2 from Population 1 and Population 2, respectively, where each population is assumed to have a continuous distribution function and the samples are taken using simple random sampling, the exact expected value for $\hat{T}_W$ in (13) is $N_1(N_1 + N_2 + 1)/2$.

We consider the statistic

$$\hat{T}_W^* = \int_0^\infty [\hat{N}_1(t) + \hat{N}_2(t)] d\hat{N}_1(t) - \frac{\hat{N}_1(\hat{N}_1 + \hat{N}_2)}{2}.$$

We use Δ rather than d to denote the total differential, since d is used under the integral. Therefore, we have

$$(\Delta \hat{T}_W^*) = \int_0^\infty [\Delta\hat{N}_1(t) + \Delta\hat{N}_2(t)] d\hat{N}_1(t)$$
$$+ \int_0^\infty [\hat{N}_1(t) + \hat{N}_2(t)] d\Delta\hat{N}_1(t)$$
$$- \frac{(\Delta\hat{N}_1)(\hat{N}_1 + \hat{N}_2) + \hat{N}_1(\Delta\hat{N}_1 + \Delta\hat{N}_2)}{2}.$$

Continuing with our usual approach, we have

$$\hat{T}_W^* - T_W^* \doteq \int_0^\infty \left(\sum w_k I\{x_k \le t\} \right) d\hat{N}_1(t)$$
$$+ \sum w_k \delta_k [\hat{N}_1(x_k) + \hat{N}_2(x_k)]$$
$$- \frac{\sum w_k \delta_k (\hat{N}_1 + \hat{N}_2) + \hat{N}_1 \sum w_k}{2} + \dots,$$

so that

$$z_k = \sum_j w_j \delta_j I\{x_k \le x_j\} + \delta_k[\hat{N}_1(x_k) + \hat{N}_2(x_k)]$$
$$- \frac{\delta_k(\hat{N}_1 + \hat{N}_2) + \hat{N}_1}{2}. \qquad (14)$$

We are not aware of this result previously being documented. It can be shown that when the null hypothesis is true and we select independently from two populations using simple random sampling, where the populations have continuous distribution functions, the variance we obtain from the z-variables in (14) is asymptotically equivalent to the usual classical formula.

3. TWO-PHASE SAMPLES

The method described above extends quite easily to the case of two-phase samples. For example, consider the two-phase ratio estimator of the population total, given by

$$\hat{Y}_{R(2)} = \frac{\hat{Y}}{\hat{X}} \hat{X}^{(1)} = \hat{R}\hat{X}^{(1)}, \qquad (15)$$

where $\hat{X}^{(1)} = \sum w_k x_k$ is the first phase estimate of X based on first phase weights $\{w_k\}$, and $\hat{Y}$ and $\hat{X}$ are the estimates of Y and X, respectively, based the second phase sample units with weights $\{w_k w_{2k}\}$, where w_{2k} is the weight assigned to the selected second phase unit, conditional on being in the first phase sample. In particular, letting

$$a_k = \begin{cases} 1 & \text{if the } k\text{-th unit is in the second phase sample,} \\ 0 & \text{otherwise,} \end{cases}$$

we have

$$\hat{Y} = \sum_{k \in s} w_k w_{2k} a_k y_k,$$

where s is the set of indices corresponding to units in the first phase sample.

Taking total differentials of (15), we have

$$(d\hat{Y}_{R(2)}) = \left(\frac{\hat{X}^{(1)}}{\hat{X}}\right)[(d\hat{Y}) - \hat{R}(d\hat{X})] + \hat{R}(d\hat{X}^{(1)}).$$

We now replace the total differentials by weighted sums over first phase units:

$$\hat{Y}_{R(2)} - \hat{Y} \doteq \sum_{k \in s} w_k \left[a_k w_{2k} \left(\frac{\hat{X}^{(1)}}{\hat{X}}\right)(y_k - \hat{R}x_k) + \hat{R}x_k \right] + \ldots,$$

so that

$$z_k = a_k w_{2k}\left(\frac{\hat{X}^{(1)}}{\hat{X}}\right)(y_k - \hat{R}x_k) + \hat{R}x_k. \qquad (16)$$

We see that the steps we have taken are essentially the same as in the one phase sample case. However, it is important to note that now z_k contains the random variable, a_k, that is used to indicate whether or not the sample unit is in the second phase sample. This is needed to compute the two phase variance estimator.

Variances obtained from the z-variable in (16) are identical to those given in Rao and Sitter (1995), who used a linearization of the jackknife to obtain their results.

Extensions to other estimation problems in two phase samples are straightforward. Suppose, for example, that $(\hat{Y}_1, \ldots, \hat{Y}_m)$ are estimates of $(Y_1, \ldots, Y_m)$ from the second phase samples, and that $(\hat{X}_1^{(1)}, \ldots, \hat{X}_p^{(1)})$ are estimates of variables available only for first phase sample units. We suppose that a set of extra parameters, λ, are defined only in terms of the units in the second phase, and that the variable of interest is defined in terms of these extra parameters and the $\hat{X}_j^{(1)}$'s. Formally, then, we have

$$U(\hat{\lambda}, \hat{Y}) = 0,$$

and

$$\hat{T} = g(\hat{X}^{(1)}, \hat{\lambda}).$$

Taking total differentials, we have as in (9),

$$(d\hat{\lambda}) = -\left[\frac{\partial \hat{U}}{\partial \hat{\lambda}}\right]^{-1}\left[\frac{\partial \hat{U}}{\partial \hat{Y}}\right](d\hat{Y}),$$

so that

$$\hat{T} - T \doteq \left[\frac{\partial g}{\partial \hat{X}^{(1)}}\right]'\left(\sum_k w_k x_k - X\right) - \left[\frac{\partial g}{\partial \hat{\lambda}}\right]'\left[\frac{\partial \hat{U}}{\partial \hat{\lambda}}\right]^{-1}\left[\frac{\partial \hat{U}}{\partial \hat{Y}}\right]\left(\sum_k a_k w_k w_{2k} y_k - Y\right).$$

Therefore, the general expression for z_k is

$$z_k = \left[\frac{\partial g}{\partial \hat{X}^{(1)}}\right]' x_k - \left[\frac{\partial g}{\partial \hat{\lambda}}\right]'\left[\frac{\partial \hat{U}}{\partial \hat{\lambda}}\right]^{-1}\left[\frac{\partial \hat{U}}{\partial \hat{Y}}\right] a_k w_{2k} y_k.$$

It then becomes necessary to put the z-variable into the algorithm that estimates the variance of the estimator of a total from a two phase sample.

4. JUSTIFICATION

The technique we have described can be considered as a direct result of the formulation given in Binder and Patak (1994). We will summarize one of the main results in that paper. Suppose we are interested in parameter θ, defined as the solution to

$$\hat{U}_1(\theta, \hat{\lambda}_\theta) = \sum_{k \in s} w_k u_1(y_k, \theta, \hat{\lambda}_\theta) = 0,$$

where $\hat{\lambda}_\theta$ is the estimate of an extra parameter, defined as the solution to

$$\hat{U}_2(\theta, \hat{\lambda}_\theta) = \sum_{k \in s} w_k u_2(y_k, \theta, \hat{\lambda}_\theta) = 0,$$

for a given θ. Through an argument based on removing extra parameters for problems of testing hypotheses on θ, Binder and Patak recommend basing inferences about θ on the variable

$$u^* = u_1(y, \theta, \hat{\lambda}_\theta) - \left[\frac{\partial \hat{U}_1}{\partial \hat{\lambda}_\theta}\right]\left[\frac{\partial \hat{U}_2}{\partial \hat{\lambda}_\theta}\right]^{-1} u_2(y, \theta, \hat{\lambda}_\theta). \qquad (17)$$

In particular, two-sided confidence intervals for θ are to be based on

$$\left\{\theta \;\middle|\; \frac{\hat{U}_1^2(\theta,\hat{\lambda}_\theta)}{\hat{W}} \leq \chi^2_{1-\alpha}(1)\right\},$$

where $\hat{W}$ is the estimated variance of the estimator of a total when the variable being estimated is u^*.

We let $u_1 = g(\lambda_1,\lambda_2) - \theta$. The kernel of the estimating equations for the y-totals will be given by $u_{21} = y - \lambda_1$ and the kernel of the estimating equations for λ_2 is given by $u_{22}(\lambda_1,\lambda_2)$. We let

$$\hat{U}_2 = \sum w_k \begin{bmatrix} u_{21} \\ u_{22} \end{bmatrix} = \begin{bmatrix} \hat{Y} - \hat{N}\hat{\lambda}_1 \\ \hat{N}u_{22} \end{bmatrix}, \text{ where } \hat{N} = \sum w_k.$$

After some algebra, from (17) the variance of interest is the variance of the estimated total based on the variable u^*, given by,

$$\left[\frac{\partial g(\hat{\lambda}_1,\hat{\lambda}_2)}{\partial\hat{\lambda}_1}\right]' y$$

$$- \left[\frac{\partial g(\hat{\lambda}_1,\hat{\lambda}_2)}{\partial\hat{\lambda}_2}\right]' \left[\frac{\partial u_{22}(\hat{\lambda}_1,\hat{\lambda}_2)}{\partial\hat{\lambda}_2}\right]^{-1} \left[\frac{\partial u_{22}(\hat{\lambda}_1,\hat{\lambda}_2)}{\partial\hat{\lambda}_1}\right] y$$

$$+ \text{ constant terms.}$$

This is equivalent to expression (10), thus showing that the methods here are consistent with those in Binder and Patak (1994).

ACKNOWLEDGEMENTS

I wish to thank Georgia Roberts and Alain Théberge who had many useful suggestions on improving the readability of an earlier draft. I am also grateful to J.N.K. Rao for many useful discussions on this topic, and to an anonymous referee for some constructive remarks.

REFERENCES

BINDER, D.A., and PATAK, Z. (1994). Use of estimating functions for interval estimation from complex surveys. *Journal of the American Statistical Association*, 89, 1035-1043.

KOTT, P.S. (1990). Estimating the conditional variance of a design consistent regression estimator. *Journal of Statistical Planning and Inference*, 24, 287-296.

RAO, J.N.K. (1995). Private communication.

RAO, J.N.K., and SITTER, R.R. (1995). Variance estimation under two-phase sampling with application to imputation for missing data. *Biometrika*, 82, 453-460.

SÄRNDAL, C.-E., SWENSSON, B., and WRETMAN, J.H. (1989). The weighted residual technique for estimating the variance of the general regression estimator of the finite population total. *Biometrika*, 76, 527-537.

Survey Methodology, June 1996
Vol. 22, No. 1, pp. 23–31
Statistics Canada

Jackknife Linearization Variance Estimators Under Stratified Multi-Stage Sampling

W. YUNG and J.N.K. RAO[1]

ABSTRACT

Variance estimation for the poststratified estimator and the generalized regression estimator of a total under stratified multi-stage sampling is considered. By linearizing the jackknife variance estimator, a jackknife linearization variance estimator is obtained which is different from the standard linearization variance estimator. This variance estimator is computationally simpler than the jackknife variance estimator and yet leads to values close to the jackknife. Properties of the jackknife linearization variance estimator, the standard linearized variance estimator, and the jackknife variance estimator are studied through a simulation study. All of the variance estimators performed well both unconditionally and conditionally given a measure of how far away the estimated totals of auxiliary variables are from the known population totals. A jackknife variance estimator based on incorrect reweighting performed poorly, indicating the importance of correct reweighting when using the jackknife method.

KEY WORDS: Generalized regression estimator; Jackknife variance estimator; Linearized variance estimator; Poststratified estimator.

1. INTRODUCTION

Large-scale sample surveys often use stratified multi-stage designs with large numbers of strata, L, and relatively few primary sampling units (clusters), $n_h(\geq 2)$, sampled within each stratum. Within each cluster, some elements (ultimate units) are sampled according to some sampling method. We do not specify the number of stages or the sampling methods used after the first-stage sampling, but we assume that subsampling within sampled clusters is performed to ensure unbiased estimation of cluster totals, Y_{hi}, $i = 1, \ldots, n_h$; $h = 1, \ldots, L$.

From the specification of the survey design, basic weights $w_{hik}(> 0)$, attached to the (hik)-th element, are obtained. Often these basic weights w_{hik} are subjected to poststratification adjustment to ensure consistency with known totals of poststratification variables. In the case of a single poststratifier, the weights are ratio-adjusted to the known population counts (*e.g.*, age-sex counts). To handle two or more poststratifiers with known marginal population counts, the weights w_{hik} can be calibrated through generalized regression (see section 4), as in the Canadian Labour Force Survey(CLFS).

The CLFS uses the jackknife method for estimating the variance of the generalized regression estimator. The jackknife method is computer intensive but it is readily applicable to general smooth statistics, unlike the linearization method. Moreover, it possesses good conditional properties. For example, in the context of simple random sampling and the ratio estimator, Royall and Cumberland (1981) showed that the jackknife variance estimator tracks the conditional variance given the sample mean of the auxiliary variable x.

The main purpose of this paper is to study variance estimation for the ratio-adjusted poststratified estimator and the generalized regression estimator under stratified sampling. By linearizing the jackknife variance estimator, a jackknife linearization variance estimator is obtained which is different from the standard linearization variance estimator. In the case of the poststratified estimator, this variance estimator is identical to Rao's (1985) variance estimator. The proposed variance estimator is computationally simpler than the jackknife variance estimator and yet leads to values close to the jackknife.

Section 2 introduces the jackknife variance estimator for the basic expansion estimator of the total, Y. Section 3 presents the jackknife and the jackknife linearization variance estimators for the poststratified estimator. These results are extended in section 4 to the generalized regression estimator in the context of multiple poststratification variables. Section 5 deals with variance estimation for a ratio of two totals, both of which are estimated using a generalized regression estimator. Results of a simulation study on the relative performances of the usual linearization variance estimator, the jackknife and the jackknife linearization variance estimators are reported in section 6.

2. BASIC ESTIMATOR

Using the basic weights w_{hik}, an unbiased estimator of the population total Y is of the form

$$\hat{Y} = \sum_{(hik)\in s} w_{hik} y_{hik}, \tag{2.1}$$

[1] W. Yung, Statistics Canada, Household Survey Methods Division, R.H. Coats Building, Tunney's Pasture, Ottawa, Ontario, K1A 0T6; and J.N.K. Rao, Department of Mathematics and Statistics, Carleton University, Ottawa, Ontario, K1S 5B6.

where s denotes the sample of elements and y_{hik} is the value of the characteristic of interest associated with the sample element $(hik) \epsilon s$. For simplicity, we assume complete response in this paper.

It is common practice to sample clusters without replacement. However, at the stage of variance estimation, the calculations are greatly simplified by treating the sample as if the clusters are sampled with replacement. This approximation generally leads to overestimation of the variance of $\hat{Y}$, but the relative bias is likely to be small if the first-stage sampling fractions are small.

An estimator of the variance of $\hat{Y}$ is given by

$$v(\hat{Y}) = \sum_{h=1}^{L} \frac{1}{n_h(n_h-1)} \sum_{i=1}^{n_h} (y_{hi} - \bar{y}_h)^2 = v(y_{hi}), \quad (2.2)$$

where $y_{hi} = \sum_k (n_h w_{hik}) y_{hik}$, and $\bar{y}_h = (1/n_h) \sum_i y_{hi}$. The operator notation $v(y_{hi})$ denotes that $v(\hat{Y})$ depends only on the y_{hi}'s.

To introduce the jackknife method, we need the estimator $\hat{Y}_{(gj)}$ for each (gj) obtained from the sample after omitting the data from the j-th sampled cluster in the g-th stratum $(j = 1, \ldots, n_g; g = 1, \ldots, L)$. It is simply obtained from (2.1) by letting $w_{gjk} = 0$, changing $w_{gik} (i \neq j)$ to $n_g w_{gik}/(n_g - 1)$ and retaining the original weights w_{hik} for $h \neq g$, *i.e.*,

$$w_{hik(gj)} = \begin{cases} 0 & \text{if } (hi) = (gj) \\ \dfrac{n_g}{(n_g-1)} w_{gik} & \text{if } h = g \text{ and } i \neq j \\ w_{hik} & \text{if } h \neq g. \end{cases}$$

These jackknife weights, $w_{hik(gj)}$, are calculated for each cluster (gj). The resulting estimator of Y is

$$\hat{Y}_{(gj)} = \sum_{(hik)\epsilon s} w_{hik(gj)} y_{hik}.$$

The jackknife variance estimator is then given by

$$v_J(\hat{Y}) = \sum_{g=1}^{L} \frac{n_g - 1}{n_g} \sum_{j=1}^{n_g} (\hat{Y}_{(gj)} - \hat{Y})^2. \quad (2.3)$$

The variance estimator (2.3) is applicable to general smooth statistics, say $\hat{\theta} = g(\hat{Y})$, by simply replacing $\hat{Y}_{(gj)}$ and $\hat{Y}$ with $\hat{\theta}_{(gj)} = g(\hat{Y}_{(gj)})$ and $\hat{\theta}$ respectively. In the linear case, $\hat{\theta} = \hat{Y}$, the jackknife variance estimator is identical to the customary variance estimator (2.2).

3. POSTSTRATIFIED ESTIMATOR

Suppose the population is partitioned into C poststrata with known population counts ${}_cM$, $c = 1, \ldots, C$. We will use the prescript c to denote poststrata. An estimator of ${}_cM$ is given by

$${}_c\hat{M} = \sum_{(hik)\epsilon_c s} w_{hik}, \quad (3.1)$$

where ${}_cs$ is the sample of elements belonging to the c-th poststratum. Similarly, an estimator of the poststratum total ${}_cY$ is

$${}_c\hat{Y} = \sum_{(hik)\epsilon_c s} w_{hik} y_{hik}.$$

Using the estimators ${}_c\hat{Y}$ and ${}_c\hat{M}$, we obtain a poststratified estimator of the total Y as

$$\hat{Y}_{ps} = \sum_c \frac{{}_cM}{{}_c\hat{M}} {}_c\hat{Y}. \quad (3.2)$$

We can rewrite (3.2) as

$$\hat{Y}_{ps} = \sum_c \sum_{(hik)\epsilon_c s} {}_cw_{hik} y_{hik}$$

where ${}_cw_{hik} = w_{hik}({}_cM/{}_c\hat{M})$ is the ratio-adjusted weight for $(hik)\epsilon_c s$. If y_{hik} is the indicator variable for a poststratum, say c, then $\hat{Y}_{ps} = {}_cM$, thus ensuring consistency with known totals, ${}_cM$.

The standard linearization variance estimator is given by (2.2) with y_{hi} changed to

$$\tilde{e}_{hi} = \sum_c \sum_{k\epsilon_c s} (n_h w_{hik}) {}_ce_{hik},$$

where ${}_ce_{hik} = y_{hik} - {}_c\hat{Y}/{}_c\hat{M}$ for the k-th element in the (hi)-th cluster belonging to ${}_cs$, *i.e.*,

$$v_L(\hat{Y}_{ps}) = v(\tilde{e}_{hi}). \quad (3.3)$$

Rao (1985) proposed an alternative linearization variance estimator using the ratio-adjusted weights ${}_cw_{hik}$:

$$v_R(\hat{Y}_{ps}) = v(e^*_{hi}) \quad (3.4)$$

where

$$e^*_{hi} = \sum_c \sum_{k\epsilon_c s} (n_h \, {}_cw_{hik}) {}_ce_{hik}.$$

Turning to the jackknife method, we need to recalculate the poststratification weights ${}_cw_{hik}$ each time a cluster (gj) is deleted. This is done by using the jackknife weights $w_{hik(gj)}$ in (3.1) to get ${}_c\hat{M}_{(gj)}$ and then using ${}_cw_{hik(gj)} = ({}_cM/{}_c\hat{M}_{(gj)}) w_{hik(gj)}$ to get

$$\hat{Y}_{ps(gj)} = \sum_{c} \sum_{(hik)\in_c s} {}_c w_{hik(gj)} y_{hik}.$$

The jackknife variance estimator is then obtained as

$$v_J(\hat{Y}_{ps}) = \sum_{g=1}^{L} \frac{n_g - 1}{n_g} \sum_{j=1}^{n_g} (\hat{Y}_{ps(gj)} - \hat{Y}_{ps})^2. \qquad (3.5)$$

By linearizing (3.5), we obtain a jackknife linearization variance estimator, $v_{JL}(\hat{Y}_{ps})$, which is identical to Rao's variance estimator (3.4); see also Valliant (1993). In the important special case of $n_h = 2$ clusters per stratum, (3.4) and (3.5) are in fact asymptotically equal to higher order terms, as the number of strata L increases (Yung 1996).

Rao (1985) justified (3.4) on heuristic grounds by noting that for simple random sampling it reduces to a conditionally valid variance estimator given the poststrata sample sizes, unlike the standard linearization variance estimator (3.3). Särndal, Swensson and Wretman (1989) obtained a variance estimator of the form (3.4) in the context of unistage sampling under a model-assisted framework. Since $v_{JL}(\hat{Y}_{ps})$ and $v_J(\hat{Y}_{ps})$ are approximately equal, the foregoing results suggest that both variance estimators should be "robust" in the sense of possessing good conditional properties given the estimated poststrata counts. Valliant (1993) conducted a simulation study to demonstrate the "robustness" of $v_J(\hat{Y}_{ps})$ and $v_{JL}(\hat{Y}_{ps})$.

4. GENERALIZED REGRESSION ESTIMATOR

In practice, it is common to form poststrata according to two or more auxiliary variables. If the resulting cell level population counts are available, the ratio-adjusted poststratified estimator can be used to increase the efficiency of the estimates. However, these cell counts may not be known in practice. For instance, marginal counts may be known only for age groups and race groups but not cell counts for the individual age-race groups. This means that in terms of a two-way table, the marginal counts are known but not the cell level counts. To handle several poststratifiers with known marginal population counts, we can use a generalized regression estimator of Y by using indicator auxiliary variables to denote the categories of the poststratifiers (Huang and Fuller 1978; Deville and Särndal 1992).

Let x_{hik} be a vector of auxiliary variables with known population totals X. The generalized regression estimator of Y is then given by

$$\hat{Y}_r = \hat{Y} + (X - \hat{X})^T \hat{B}, \qquad (4.1)$$

where

$$\hat{X} = \sum_{(hik)\in s} w_{hik} x_{hik},$$

and $\hat{B}$ is the vector of estimated regression coefficients

$$\hat{B} = \hat{A}^{-1}\hat{b},$$

where

$$\hat{A} = \sum_{(hik)\in s} w_{hik} x_{hik} x_{hik}^T,$$

and

$$\hat{b} = \sum_{(hik)\in s} w_{hik} x_{hik} y_{hik}.$$

The poststratified estimator, $\hat{Y}_{ps}$, is a special case of (4.1) by letting x_{hik} denote the vector of indicator variables for the poststrata. In this case, $\hat{X} = ({}_1\hat{M}, \ldots, {}_C\hat{M})^T$, $X = ({}_1M, \ldots, {}_CM)^T$, and $\hat{B} = ({}_1\hat{R}, \ldots, {}_C\hat{R})^T$ with ${}_c\hat{R} = {}_c\hat{Y}/{}_c\hat{M}$. Thus,

$$\hat{Y}_r = \hat{Y} + \sum_c {}_c\hat{R}({}_cM - {}_c\hat{M}) = \hat{Y}_{ps}.$$

In the case of two or more poststratifiers, X corresponds to the vector of marginal population counts.

The generalized regression estimator may be rewritten as

$$\hat{Y}_r = \sum_{(hik)\in s} w_{hik}^* y_{hik},$$

where

$$w_{hik}^* = w_{hik} a_{hik} \qquad (4.2)$$

is the "final" or "calibration" weight with

$$a_{hik} = 1 + x_{hik}^T \hat{A}^{-1}(X - \hat{X}).$$

In the special case of $\hat{Y}_{ps}$, we have $a_{hik} = {}_cM/{}_c\hat{M}$ for $(hik)\in_c s$. Writing $\hat{Y}_r$ in the operator notation as $\hat{Y}_r(y_{hik})$, it is readily verified that the generalized regression estimator $\hat{X}_r = \hat{Y}_r(x_{hik}) = X$, thus ensuring consistency with known totals X.

Turning to variance estimation, the standard linearization variance estimator is again given by (2.2) with y_{hi} changed to

$$\tilde{e}_{hi} = \sum_k (n_h w_{hik}) e_{hik},$$

where

$$e_{hik} = y_{hik} - x_{hik}^T \hat{B} \qquad (4.3)$$

are the estimated residuals, *i.e.*,

$$v_L(\hat{Y}_r) = v(\tilde{e}_{hi}). \qquad (4.4)$$

For the jackknife method we need to recalculate the calibration weights w^*_{hik} each time a cluster (gj) is deleted. These weights are given by

$$w^*_{hik(gj)} = w_{hik(gj)} a_{hik(gj)},$$

where

$$a_{hik(gj)} = 1 + x^T_{hik} \hat{A}^{-1}_{(gj)} (X - \hat{X}_{(gj)}),$$

$$\hat{A}_{(gj)} = \sum_{(hik)\in s} w_{hik(gj)} x_{hik} x^T_{hik},$$

and

$$\hat{X}_{(gj)} = \sum_{(hik)\in s} w_{hik(gj)} x_{hik}.$$

Denote the resulting generalized regression estimator as

$$\hat{Y}_{r(gj)} = \sum_{(hik)\in s} w^*_{hik(gj)} y_{hik}$$

$$= \hat{Y}_{(gj)} + (X - \hat{X}_{(gj)})^T \hat{\boldsymbol{B}}_{(gj)}$$

where $\hat{\boldsymbol{B}}_{(gj)}$ is the vector of estimated regression coefficients when the (gj)-th cluster is deleted:

$$\hat{\boldsymbol{B}}_{(gj)} = \hat{A}^{-1}_{(gj)} \hat{\boldsymbol{b}}_{(gj)}$$

with

$$\hat{\boldsymbol{b}}_{(gj)} = \sum_{(hik)\in s} w_{hik(gj)} x_{hik} y_{hik}.$$

The jackknife variance estimator of $\hat{Y}_r$ is then given by

$$v_J(\hat{Y}_r) = \sum_{g=1}^{L} \frac{n_g - 1}{n_g} \sum_{j=1}^{n_g} (\hat{Y}_{r(gj)} - \hat{Y}_r)^2. \qquad (4.5)$$

It is shown in the Appendix that by linearizing the jackknife variance estimator (4.5), one obtains

$$v_{JL}(\hat{Y}_r) = v(e^*_{hi}) \qquad (4.6)$$

with

$$e^*_{hi} = \sum_{k} (n_h w^*_{hik}) e_{hik}$$

where w^*_{hik} is defined in (4.2) and e_{hik} is defined in (4.3). It is interesting to note that the jackknife linearization variance estimator (4.6) is similar to the model-assisted variance estimator proposed by Särndal, Swensson and Wretman (1989) in the context of unistage sampling. Yung (1996) established the asymptotic equivalence of $v_J(\hat{Y}_r)$ and $v_{JL}(\hat{Y}_r)$ to higher order terms in the important special case of $n_h = 2$ clusters per stratum. Note that the above results are also applicable to general auxiliary variables, x_{hik}.

Binder (1996) proposed a new linearization method which also leads to $v_{JL}(\hat{Y}_r)$. In this method, the partial derivatives are evaluated at the estimates $\hat{Y}, \hat{X}$ and $\hat{\boldsymbol{B}}$, rather than the population values Y, X and $\boldsymbol{B}$ as in the traditional linearization method. Given that v_J and v_{JL} are design-consistent (Yung 1996) and possess good conditional properties, our results provide theoretical justification for Binder's method which was proposed as a "cookbook approach".

The computation of the jackknife variance estimator involves the inversion of the matrix $\hat{A}_{(gj)}$ for each (gj). However, the jackknife variance estimator can be approximated by retaining the inverse for the full sample, $\hat{A}^{-1}$, and then using modified weights

$$\tilde{w}_{hik(gj)} = w_{hik(gj)} \tilde{a}_{hik(gj)}$$

with

$$\tilde{a}_{hik(gj)} = 1 + (w_{hik}/w_{hik(gj)}) x'_{hik} \hat{A}^{-1} (X - \hat{X}_{(gj)}).$$

The resulting estimator of Y, when the (gj)-th cluster is deleted, is given by

$$\tilde{Y}_{r(gj)} = \sum_{(hik)\in s} \tilde{w}_{hik(gj)} y_{hik}$$

and the corresponding jackknife variance estimator is

$$v_{J1}(\hat{Y}_r) = \sum_{g=1}^{L} \frac{n_g - 1}{n_g} \sum_{j=1}^{n_g} (\tilde{Y}_{r(gj)} - \hat{Y}_r)^2. \qquad (4.7)$$

It is readily seen that (4.7) is exactly equal to the standard linearization variance estimator (4.4).

5. ESTIMATION OF A RATIO

Often a ratio of two estimated totals is required. For example, in a family expenditure survey, one may be interested in the proportion of income spent on clothing. Let

$$\hat{Y}_r = \hat{Y} + (X - \hat{X})^T \hat{\boldsymbol{B}}_1$$

be a generalized regression estimator of the total amount spent on clothing, Y. Similarly, let

$$\hat{Z}_r = \hat{Z} + (X - \hat{X})^T \hat{\boldsymbol{B}}_2$$

be a generalized regression estimator of the total income, Z. The proportion of interest is $\theta = Y/Z$, and can be estimated by

$$\hat{\theta} = \hat{Y}_r / \hat{Z}_r.$$

The jackknife variance estimator is given by

$$v_J(\hat{\theta}) = \sum_g \frac{n_g - 1}{n_g} \sum_j (\hat{\theta}_{(gj)} - \hat{\theta})^2 \qquad (5.1)$$

where

$$\hat{\theta}_{(gj)} = \hat{Y}_{r(gj)}/\hat{Z}_{r(gj)}.$$

Linearizing the jackknife variance estimator, (5.1), we obtain a jackknife linearization variance estimator

$$v_{JL}(\hat{\theta}) = v(r_{hi}^{**}) \qquad (5.2)$$

where

$$r_{hi}^{**} = \frac{1}{\hat{Z}_r} \sum_k (n_h w_{hik}^*) e_{hik}^*$$

with

$$e_{hik}^* = e_{hik} - \frac{\hat{Y}_r}{\hat{Z}_r} \tilde{e}_{hik},$$

and

$$e_{hik} = y_{hik} - \boldsymbol{x}_{hik}^T \hat{\boldsymbol{B}}_1, \quad \tilde{e}_{hik} = z_{hik} - \boldsymbol{x}_{hik}^T \hat{\boldsymbol{B}}_2.$$

Proof of (5.2) is omitted for simplicity.

6. SIMULATION STUDY

We performed a simulation study to investigate the unconditional and conditional finite sample properties of the variance estimators in the case of a single poststratifier as well as two poststratification variables. For this purpose, we used a fixed finite population, considered by Valliant (1993), consisting of 10,841 persons included in the September 1988 Current Population Survey (CPS) of the United States. The variable of interest, y, is the weekly wages for each person. The single poststratifier was defined on the basis of age, race and sex, while the two poststratifiers were based on the variables age, with five levels, and race, with two levels (see Tables 1 and 2 for details).

Table 1

Assignment of Age/Race/Sex Categories to Poststrata: Single Poststratifier

Age	Nonblack		Black	
	Male	Female	Male	Female
19 and under	1	1	1	1
20-24	2	3	3	3
25-34	5	6	4	4
35-64	7	8	4	4
65 and over	2	3	3	1

Note: Cell numbers (1-8) are poststratum identification numbers.

Table 2

Assignment of Age/Race Categories to Poststrata: Two Poststratifiers

Age	Nonblack	Black	
19 and under	(1,1)	(1,2)	PS1(1)
20-24	(2,1)	(2,2)	PS1(2)
25-34	(3,1)	(3,2)	PS1(3)
35-64	(4,1)	(4,2)	PS1(4)
65 and over	(5,1)	(5,2)	PS1(5)
	PS2(1)	PS2(2)	

Note: Number in margins are poststratum identification numbers. Cells (i,j) denote poststrata $(i = 1, \ldots, 5; j = 1, 2)$.

The study population contained 2,826 geographical segments, each composed of about four neighbouring households. One hundred design strata ($L = 100$) were created with each stratum having about the same total number of households. We used a stratified two-stage sampling design with segments as clusters and persons as the second-stage units. In each stratum $n_h = 2$ segments were selected with probability proportional to the number of persons in each segment, and a simple random sample of $m_{hi} = 4$ persons was selected without replacement if the sample segment contained more than four persons. In sample segments with four or fewer persons, all persons in the segment were selected. Using this design, we selected two sets of 10,000 independent samples, one set for the one-way poststratification case and the other set for the two-way poststratification case.

From each sample, we computed the basic estimator, the relevant poststratified estimator, $\hat{Y}_{ps}$ or $\hat{Y}_r$, and four variance estimators: the standard linearization variance estimator v_L, the jackknife linearization variance estimator v_{JL}, the jackknife v_J, and an incorrect jackknife variance estimator v_J^*. In applying the jackknife procedure, it is questioned whether or not the "final" or "calibrated" weights need to be recalculated each time a cluster is deleted. The correct jackknife variance estimator does recalculate the "final" weight whenever a cluster is deleted while the incorrect jackknife variance estimator fails to do this. For the one-way poststratification case, $v_J^*(\hat{Y}_{ps})$ uses the full adjustment ${}_cM/{}_c\hat{M}$ instead of ${}_cM/{}_c\hat{M}_{(gj)}$ when the (gj)-th cluster is deleted, *i.e.*, $\hat{Y}_{ps(gj)}$ uses the weights $({}_cM/{}_c\hat{M}) w_{hik(gj)}$ instead of $({}_cM/{}_c\hat{M}_{(gj)}) w_{hik(gj)}$. Similarly, for the two-way poststratification case, $v_J^*(\hat{Y}_r)$ uses the full adjustment a_{hik} instead of $a_{hik(gj)}$ when the (gj)-th cluster is deleted, *i.e.*, $\hat{Y}_r$ uses the weights $w_{hik(gj)} a_{hik}$ instead of $w_{hik(gj)} a_{hik(gj)}$. The linearized version of v_J^* is the same as the variance estimator v_R (equation 3.4) with ${}_ce_{hik}$ replaced by y_{hik} in the case of $\hat{Y}_{ps}$, and v_{JL} (equation 4.6) with e_{hik} replaced by y_{hik} in the case of the generalized regression estimator $\hat{Y}_r$. That is,

$$v_J^*(\hat{Y}_{ps}) = v(y_{hi}^*)$$

with

$$y^*_{hi} = \sum_c \sum_{k\epsilon_c s} (n_h\, {}_c w_{hik}) y_{hik}$$

and

$$v^*_J(\hat{Y}_r) = v(y^*_{hi})$$

with

$$y^*_{hi} = \sum_{k\epsilon s} (n_h w^*_{hik}) y_{hik}.$$

Since v^*_J uses the y's instead of the residuals e's, it is clear that v^*_J should overestimate the true variance of the estimator, although it is computationally simpler than v_J.

(i) Unconditional Results

To compare the unconditional performances of the variance estimators we computed the empirical relative bias (RB) for each variance estimator: RB of a variance estimator v is

$$\mathrm{RB} = \frac{1}{\mathrm{MSE}}\left[\frac{1}{10{,}000}\sum_i v_1\right] - 1$$

where v_i is the value of v for the i-th simulated sample ($i = 1, \ldots, 10{,}000$) and MSE is the empirical MSE of the estimator, say $\tilde{Y}$:

$$\mathrm{MSE} = \frac{1}{10{,}000}\sum_i (\tilde{Y}_i - Y)^2$$

where $\tilde{Y}_i$ is the value of $\tilde{Y}$ in the i-th simulated sample.

Error rates for normal theory confidence intervals on the total Y were also calculated for each variance estimator, using a nominal error rate of 5%:

error rate =

$$1 - \frac{1}{10{,}000}\,(\text{number of samples with } L_i \leq Y \leq U_i),$$

where $L_i \leq Y \leq U_i$ is a confidence interval on Y for the i-th simulated sample. Lower and upper error rates were calculated as:

lower error rate =

$$\frac{1}{10{,}000}\,(\text{number of samples with } Y < L_i)$$

upper error rate =

$$\frac{1}{10{,}000}\,(\text{number of samples with } Y > U_i).$$

We also calculated the average lengths of the confidence intervals as

$$\text{average length} = \frac{1}{10{,}000}\sum_i (U_i - L_i).$$

Table 3 reports the unconditional results for the poststratified estimator $\hat{Y}_{ps}$ using the above performance measures. With respect to relative bias, v_{JL} and v_J both perform well with RB < 1% while the incorrect jackknife v^*_{JL} severely overestimates the MSE (RB = 37%). We note that v_L is also estimating the MSE of $\hat{Y}_{ps}$ well unconditionally (RB < 1%), contrary to Valliant's (1993) claim. Valliant (1993) reported RB of 35% for v_L using the same data set. In view of the design-consistency of v_L supplemented by our simulation results on v_L, we conjecture that Valliant's calculations on v_L might be incorrect.

Table 3

Unconditional Results for the Poststratified Estimator

Performance Measure	$v_L(\hat{Y}_{ps})$	$v_{JL}(\hat{Y}_{ps})$	$v_J(\hat{Y}_{ps})$	$v^*_J(\hat{Y}_{ps})$
Relative bias (%)	−0.44	0.12	0.26	37.16
Error rate (%)	5.20	5.09	5.06	2.41
Lower error rate (%)	2.41	2.35	2.33	0.99
Upper error rate (%)	2.79	2.74	2.73	1.42
Average length	3.81	3.82	3.83	4.48

Turning to confidence interval performance, Table 3 shows that the error rates associated with v_J, v_{JL} and v_L are close to the nominal 5% while the error rate for v^*_J is considerably lower than 5% (about 2.5%). Performances with respect to lower and upper error rates are also similar. The variance estimators, v_J, v_{JL} and v_L, perform similarly in terms of average length of confidence intervals while the average length associated with v^*_J is significantly larger due to overestimation bias. Finally, we note that the performance measures for v_J and v_{JL} are very close, supporting the asymptotic equivalence of v_J and v_{JL}.

Table 4

Unconditional Results for the Generalized Regression Estimator

Performance Measure	$v_L(\hat{Y}_r)$	$v_{JL}(\hat{Y}_r)$	$v_J(\hat{Y}_r)$	$v^*_J(\hat{Y}_r)$
Relative bias (%)	−0.96	0.76	0.57	25.87
Error rate (%)	5.30	5.27	5.23	3.07
Lower error rate (%)	2.24	2.21	2.19	1.08
Upper error rate (%)	3.06	3.06	3.04	1.99
Average length	3.94	3.95	3.95	4.44

Unconditional results for the generalized regression estimator $\hat{Y}_r$ are reported in Table 4. As in the case of $\hat{Y}_{ps}$, the variance estimators v_J, v_{JL} and v_L perform well both in terms of relative bias and error rates of confidence intervals. On the other hand, the incorrect jackknife v_J^* leads to severe overestimation which in turn is reflected in the lower than nominal error rates and larger average length of confidence intervals.

(ii) Conditional Results

We have also studied conditional properties of the variance estimators, following Valliant (1993). For the poststratified estimator, we divided the 10,000 simulated samples into 10 groups each containing 1,000 samples using the measure (Valliant 1993)

$$D_{ps} = \sum_c \left(\frac{{}_c\hat{M}}{{}_cM} - 1 \right).$$

The measure D_{ps} was calculated for each sample and the 10,000 samples were sorted in ascending order according to the D_{ps}-values and then divided into groups. We may interpret D_{ps} as a measure of how "balanced" the sample is with respect to the distribution of the poststrata counts.

For the generalized regression estimator, we used the following natural extension of D_{ps}:

$$D_r = \sum_a \left(\frac{{}_a\hat{M}}{{}_aM} - 1 \right) + \sum_b \left(\frac{{}_b\hat{M}}{{}_bM} - 1 \right),$$

where a and b index the levels of the two poststratification variables and $({}_a\hat{M}, {}_aM)$ and $({}_b\hat{M}, {}_bM)$ are the corresponding marginal counts. We may interpret D_r as a measure of how "balanced" the sample is with respect to the distribution of the marginal poststrata counts.

Table 5

Conditional Relative Biases (%) for the Poststratified Estimator

Group	$v_L(\hat{Y}_{ps})$	$v_{JL}(\hat{Y}_{ps})$	$v_J(\hat{Y}_{ps})$	$v_J^*(\hat{Y}_{ps})$
1	−5.00	−8.05	−7.88	17.83
2	0.55	−1.18	−1.01	28.06
3	8.33	7.03	7.19	41.29
4	−1.10	−1.56	−1.42	31.82
5	−0.76	−0.69	−0.55	34.77
6	2.50	3.39	3.53	41.69
7	6.10	7.51	7.66	48.86
8	6.60	8.82	8.96	53.54
9	−4.46	−1.43	−1.31	41.11
10	−13.56	−9.17	−9.07	36.63

Table 6

Conditional Error Rates (%) for the Poststratified Estimator

Group	$v_L(\hat{Y}_{ps})$	$v_{JL}(\hat{Y}_{ps})$	$v_J(\hat{Y}_{ps})$	$v_J^*(\hat{Y}_{ps})$
1	5.5	5.9	5.9	3.4
2	4.6	4.8	4.8	2.9
3	3.7	3.8	3.8	1.9
4	5.7	5.8	5.8	2.9
5	4.9	4.8	4.7	2.6
6	5.1	5.0	4.8	2.2
7	5.2	4.8	4.8	2.1
8	4.5	4.3	4.3	1.3
9	5.8	5.4	5.4	2.4
10	7.0	6.3	6.3	2.4

The results for the poststratified estimator are given in Tables 5 and 6: conditional relative biases in Table 5 and conditional error rates (nominal 5%) in Table 6. These performance measures were computed in the same manner as the unconditional case but from each group separately. It is clear from Tables 5 and 6 that v_J, v_{JL} and v_L all perform well, although v_L is somewhat worse in the extreme groups 1 and 10, while v_J^* performed poorly as before. It is somewhat surprising to see v_L performing so well conditionally. A possible explanation is that with our particular sampling design we have $\hat{M} = \sum_{(hik) \in s} w_{hik} = M$ so that

$$\sum_c {}_c\hat{M} = \hat{M} = M.$$

Because of this, we do not obtain samples which are poorly balanced since if some poststrata counts ${}_c\hat{M}$ are gross overestimates, say, then the other counts correct for the overestimation in order to satisfy the above constraint. Thus, we see mostly well balanced samples in which case v_L is expected to perform well.

Table 7

Conditional Relative Biases (%) for the Generalized Regression Estimator

Group	$v_L(\hat{Y}_r)$	$v_{JL}(\hat{Y}_r)$	$v_J(\hat{Y}_r)$	$v_J^*(\hat{Y}_r)$
1	9.25	4.95	5.13	26.51
2	3.99	1.50	1.67	24.96
3	−3.24	−4.76	−4.59	17.53
4	−2.66	−3.43	−3.26	20.53
5	7.90	7.61	7.80	35.46
6	−3.60	−3.12	−2.94	23.38
7	−9.24	−8.27	−8.08	17.41
8	3.34	5.30	5.50	35.84
9	−3.75	−0.85	−0.62	30.84
10	−8.68	−4.15	−3.92	28.50

Table 8

Conditional Error Rates (%) for the Generalized Regression Estimator

Group	$v_L(\hat{Y}_r)$	$v_{JL}(\hat{Y}_r)$	$v_J(\hat{Y}_r)$	$v_J^*(\hat{Y}_r)$
1	4.3	4.5	4.4	3.0
2	4.9	5.0	5.0	3.3
3	5.0	5.1	5.1	3.8
4	5.7	5.9	5.9	3.3
5	3.9	4.0	4.0	2.3
6	5.7	5.8	5.7	3.0
7	5.9	5.8	5.8	2.9
8	5.8	5.7	5.7	2.8
9	5.5	5.1	4.9	3.0
10	6.3	5.8	5.8	3.3

The results for the generalized regression estimator are given in Tables 7 and 8: conditional relative biases in Table 7 and conditional error rates (nominal 5%) in Table 8. The results are very similar to those for the one stratifier case. In both cases we again note that the performance measures for v_J and v_{JL} are very close, supporting the asymptotic equivalence of v_J and v_{JL}.

In summary, the three variance estimators v_J, v_{JL} and v_L performed similarly. The incorrect jackknife v_J^* performed poorly indicating that reweighting must be done each time a cluster is deleted.

7. CONCLUDING REMARKS

Beebakhee (1995) applied the three variance estimators, v_J, v_{JL} and v_L, to a number of household surveys conducted by Statistics Canada. Her empirical results showed that the jackknife linearization variance estimator, v_{JL}, consistently consumed less time and money for all study surveys than the jackknife variance estimator, v_J, and yet approximated v_J very well. These results are practically important because the users wanted a computationally simpler variance estimator which can approximate the currently used v_J very well. The standard linearization variance estimator v_L performed similar to v_{JL} in terms of cost and time, but it did not approximate v_J as well as v_{JL}.

If the primary interest is the estimation of totals or ratios, then the jackknife linearization variance estimator, v_{JL}, is attractive because it is computationally simpler than the jackknife variance estimator, v_J, and yet leads to values close to the jackknife. But for general smooth statistics v_{JL} suffers from the same disadvantage as the standard linearization variance estimator, v_L, in the sense that both require the derivation of a separate formula for each statistic, unlike v_J. In terms of statistical properties, our simulation study suggests that the three variance estimators, v_J, v_{JL}, and v_L, perform similarly. On the other hand, the incorrect jackknife v_J^*, which uses the same adjustment whenever a cluster is deleted, performs poorly indicating that reweighting must be done each time a cluster is deleted.

ACKNOWLEDGEMENT

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

APPENDIX

Proof of the Result $v_J(\hat{Y}_r) \approx v_{JL}(\hat{Y}_r)$

To establish the desired result, we first approximate the difference $\hat{A}_{(gj)}^{-1} - \hat{A}^{-1}$. Using the matrix identity,

$$(I + PQ)^{-1} = I - P(I + QP)^{-1}Q$$

we get

$$\begin{aligned}\hat{A}_{(gj)}^{-1} - \hat{A}^{-1} &= \hat{A}^{-1}[I + (\hat{A}_{(gj)} - \hat{A})\hat{A}^{-1}]^{-1} - \hat{A}^{-1}\\ &= \hat{A}^{-1}[I - (\hat{A}_{(gj)} - \hat{A})\\ &\qquad (I + \hat{A}^{-1}(\hat{A}_{(gj)} - \hat{A}))^{-1}\hat{A}^{-1}] - \hat{A}^{-1}\\ &\approx -\hat{A}^{-1}(\hat{A}_{(gj)} - \hat{A})\hat{A}^{-1}. \qquad \text{(A.1)}\end{aligned}$$

The approximation (A.1) follows by noting that (i) $\hat{A}_{(gj)} - \hat{A}$ is of lower order than $\hat{A}$ under the assumption that no cluster contribution is of disproportionate size as the number of strata L increases (see Yung (1996) for details on regularity conditions) and (ii) $[I + \hat{A}^{-1}(\hat{A}_{(gj)} - \hat{A})]^{-1} \approx I - \hat{A}^{-1}(\hat{A}_{(gj)} - \hat{A})$.

Using (A.1), we obtain

$$\begin{aligned}\hat{\boldsymbol{B}}_{(gj)} - \hat{\boldsymbol{B}} &= (\hat{A}_{(gj)}^{-1} - \hat{A}^{-1} + \hat{A}^{-1})(\hat{\boldsymbol{b}}_{(gj)} - \hat{\boldsymbol{b}} + \hat{\boldsymbol{b}})\\ &\qquad - \hat{A}^{-1}\hat{\boldsymbol{b}}\\ &\approx (\hat{A}_{(gj)}^{-1} - \hat{A}^{-1})\hat{\boldsymbol{b}} + \hat{A}^{-1}(\hat{\boldsymbol{b}}_{(gj)} - \hat{\boldsymbol{b}})\\ &\approx -\hat{A}^{-1}(\hat{A}_{(gj)} - \hat{A})\hat{\boldsymbol{B}} + \hat{A}^{-1}(\hat{\boldsymbol{b}}_{(gj)} - \hat{\boldsymbol{b}}). \qquad \text{(A.2)}\end{aligned}$$

It now follows from (A.2) that

$$\begin{aligned}\hat{Y}_{r(gj)} - \hat{Y}_r &\approx (\hat{Y}_{(gj)} - \hat{Y}) - (\hat{X}_{(gj)} - \hat{X})^T\hat{\boldsymbol{B}}\\ &\qquad - (\hat{X} - X)^T(\hat{\boldsymbol{B}}_{(gj)} - \hat{\boldsymbol{B}})\\ &\approx \frac{1}{n_g - 1}(\bar{e}_g^* - e_{gj}^*), \qquad \text{(A.3)}\end{aligned}$$

where $e^*_{gj} = \sum_k (n_g w^*_{gjk}) e_{gjk}$ and $\bar{e}^*_g = (1/n_g) \sum_j e^*_{gj}$. We used the following results in arriving at (A.3):

$$(\hat{Y}_{(gj)} - \hat{Y}) - (\hat{X}_{(gj)} - \hat{X})^T \hat{B} = \frac{1}{n_g - 1} (\bar{e}_g - e_{gj})$$

and

$$(\hat{X} - X)^T (\hat{B}_{(gj)} - \hat{B}) \approx$$

$$(X - \hat{X})^T \hat{A}^{-1} \left[\frac{1}{n_g - 1} (\bar{u}_g - u_{gj}) \right],$$

where $e_{gj} = \sum_k (n_g w_{gjk}) e_{gjk}$ and $u_{gj} = \sum_k (n_g w_{gjk}) x_{gjk} e_{gjk}$.

It now follows from (A.3) that

$$v_J(\hat{Y}_r) \approx \sum_{h=1}^{L} \frac{1}{n_h(n_h - 1)} \sum_{i=1}^{n_h} (e^*_{hi} - \bar{e}^*_h)^2$$

$$= v(e^*_{hi}) = v_{JL}(\hat{Y}_r).$$

REFERENCES

BEEBAKHEE, R. (1995). A comparison of two variance estimation methods: The Jackknife and the linearized Jackknife. Methodology Branch Working Paper, HSMD-95-005E. Statistics Canada.

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

CASADY, R.J., and VALLIANT, R. (1993). Conditional properties of poststratified estimators under normal theory. *Survey Methodology*, 19, 183-192.

DEVILLE, J., and SÄRNDAL, C.E. (1992). Calibration estimators in survey sampling. *Journal of the American Statistical Association*, 87, 376-382.

HUANG, E.T., and FULLER, W.A. (1978). Nonnegative regression estimation for sample survey data. *Proceedings of the Social Statistics Section, American Statistical Association*, 300-305.

RAO, J.N.K. (1985). Conditional inference in survey sampling. *Survey Methodology*, 11, 15-31.

ROYALL, R.M., and CUMBERLAND, W.G. (1981). An empirical study of the ratio estimator and estimator of its variance. *Journal of the American Statistical Association*, 76, 66-88.

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.

STATISTICS CANADA (1990). *Methodology of the Canadian Labour Force Survey*. Catalogue No. 71-526.

VALLIANT, R. (1993). Poststratification and conditional variance estimation. *Journal of the American Statistical Association*, 88, 89-96.

YUNG, W. (1996). Contributions to poststratification in stratified multi-stage samples. Unpublished Ph.D. thesis, Carleton University, Ottawa, Canada.

Survey Methodology, June 1996
Vol. 22, No. 1, pp. 33–41
Statistics Canada

Small Area Estimation Under an Inverse Gaussian Model

Y.P. CHAUBEY, F. NEBEBE and P.S. CHEN[1]

ABSTRACT

In this paper, we consider analysis of variance methodology for inverse Gaussian distribution and adapt it for estimation of small area parameters in finite populations. It is demonstrated, through a Monte Carlo study, that these estimators offer a competitive choice for positively skewed survey data such as income or yield of a particular sector.

KEY WORDS: Interactions; Inverse Gaussian; Monte Carlo; Regression estimates; Synthetic estimates; Särndal-Hidiroglou estimator; Unbalanced model.

1. INTRODUCTION

Recently, a large number of methods appeared in the literature for the problem of small area estimation; for example Prasad and Rao (1990), Särndal and Hidiroglou (1989), Choudhry and Rao (1988), and Särndal (1984) and the references cited there, especially Särndal and Råbäck (1983), Fay and Herriot (1979), Schaible (1979), Holt, Smith and Tomberlin (1979), and Gonzalez and Hoza (1978), to name a few. The need for small area estimates of several characteristics of a given population has generated various useful procedures that produced realistic and sufficiently accurate estimates for local areas and other special subgroups. Several of the techniques suggested by the authors mentioned above were implicitly and/or explicitly model-based and utilized the standard normal theory. Others have tackled the provision of estimates for local areas from Bayesian and empirical Bayes perspectives by finding a compromise between the sample mean of an area (that is assumed to be normal) and an estimator based on regression on one or more covariates (see *e.g.*, Stroud 1987; MacGibbon and Tomberlin 1989). For an extensive review of recent developments in small area estimation, the reader may refer to Ghosh and Rao (1994).

The standard normal theory analysis of factorial experiments may be inappropriate to apply in situations where data are generated from markedly positively skewed distributions. While most of the inference procedures are analytically tractable, the accuracy and reliability of the results may be questionable in many practical applications. Thus, such an analysis based on positively skewed distributions is called for.

The objective of this paper is to consider inference procedures for unbalanced as well as balanced two-factor experiments under inverse Gaussian model that may be used to produce estimates for small regions. Hidiroglou and Särndal (1985) reported on a Monte Carlo study where a modified regression estimator is preferred as a compromise between the synthetic estimator and the generalized regression estimator. Särndal and Hidiroglou (1989) also presented further comparisons of estimators on the basis of conditional inference. The generalized regression estimator is basically derived from a super population regression model without any distributional assumptions. Chaubey (1991) considered super population models of Durbin (1959) with gamma auxiliary and inverse Gaussian auxiliary in which case the generalized regression estimator has the property of being the best linear unbiased predictor (see Prasad and Rao 1990). In fact, the best linear unbiased predictor for the population total does not depend on the form of the distribution of the characteristic variable, hence this technique is preferable given that maximum likelihood estimates (MLE) may be hard to obtain. As we have seen that the super population distributions (as transfused in the populations) may resemble closely to inverse Gaussian distributions for variety of populations we would like to exploit this aspect of the population.

The use of inverse Gaussian distribution is not merely a superficial one but it has been used successfully in many situations (see Folks and Chhikara 1978) and resembles closely to gamma, log normal and Weibull populations which are common in modeling positively skewed non negative random variables. In this paper, we study the use of inverse Gaussian model in applying to the small area estimation. The approach of Fries and Bhattacharyya (1983) which discusses the analysis of two factor experiments under an inverse Gaussian model is of major importance. The above paper gives estimation in balanced, no-interaction model. We have extended this approach to unbalanced case, which is essential for estimation of domain totals or means. In this respect the general multiple regression approach of Bhattacharyya and Fries (1986), and Whitmore (1983) may be adapted, but we have chosen to take the direct approach. In Section 2 we specify the

[1] Y.P. Chaubey, Professor, Department of Mathematics and Statistics; F. Nebebe, Associate Professor, Department of Decision Sciences & M.I.S.; and P.S. Chen, Research Assistant, Department of Finance, Concordia University, Montreal, Canada.

model and present our proposed estimators under the inverse Gaussian model. In Section 3, a numerical study is carried out for evaluation of the performance of the proposed estimator through Monte Carlo simulation. Finally, Section 4 presents summary and conslusions.

2. THE INVERSE GAUSSIAN REGRESSION MODEL FOR SMALL AREA ESTIMATION

Suppose that a finite population $\mathcal{U}$ is divided into D non-overlapping domains $U_{d.}$, $d = 1(1)D$, with $N_{d.}$ as the size of $U_{d.}$. The population is further divided along a second dimension, into G non-overlapping groups $U_{.g}$, $g = 1(1)G$, with the size of $U_{.g}$ denoted by $N_{.g}$. The cross-classification of domains and groups give rise to DG population cells U_{dg}, $d = 1(1)D, g = 1(1)G$, with N_{dg} as the size of U_{dg}. The population size N can then be expressed as $N = \sum_d N_{d.} = \sum_g N_{.g} = \sum_{dg} N_{dg}$. Our interest lies in estimating domain totals $t_d = \sum_{U_{d.}} y_k$, where y represents the characteristic variable and y_k is the observation on k-th unit. A sample s of size n is selected from $\mathcal{U}$ by a simple random sampling. Denote by $s_{d.}$, $s_{.g}$ and s_{dg} the parts of s that happen to fall in $U_{d.}$, $U_{.g}$ and U_{dg}. The corresponding sample sizes are denoted by $n_{d.}$, $n_{.g}$ and n_{dg}, respectively.

2.1 Regression Method for Inverse Gaussian Data

We refer readers to two recent comprehensive reviews about the developments in the inverse Gaussian distribution, namely, Chhikara and Folks (1989), and Iyengar and Patwardhan (1988). The probability density function of an inverse Gaussian variate with parameters (θ, σ), $IG(\theta, \sigma)$, is given by

$$f(y;\theta,\sigma) = (2\pi\sigma)^{-1/2}y^{-3/2}\exp[-(2\sigma y)^{-1}(y\theta^{-1} - 1)^2]; \tag{2.1}$$

with $y > 0, \theta > 0, \sigma > 0$. The mean and variance of this distribution are θ and $\theta^3\sigma$, respectively. Bhattacharyya and Fries (1982) proposed a reciprocal linear model for θ. Specifically, they assume a model of the form $\theta_k^{-1} = x_k'\eta$. An estimator of η, similar to the estimator of the regression parameter in the usual linear model (see Särndal 1984) in this situation is given by

$$\hat{\eta} = \left(\sum_{k \epsilon S_{d.}} \frac{x_k x_k' y_k}{\pi_k}\right)^{-1} \sum_{k \epsilon S_{d.}} \frac{x_k}{\pi_k}. \tag{2.2}$$

This is called pseudo Maximum Likelihood estimator, because it is obtained by unconditional maximization of the likelihood function and therefore $x_k'\hat{\eta} > 0$ may not be satisfied for all k. Then an estimator of the total t_d of the d-th domain in the spirit of Särndal's (1984) modified regression estimator may be constructed as

$$\hat{t}_{dIG} = \sum_{k \epsilon U_{d.}} \hat{y}_k + \sum_{k \epsilon S_{d.}} \frac{e_k}{\pi_k} \tag{2.3}$$

where $\hat{y}_k = x_k'\hat{\eta}$ and $e_k = y_k - \hat{y}_k$. In what follows, we denote the mean of the (d,g) cell by θ_{dg}, and consider the case of simple random sampling in which case π_k's are constant. We first discuss the prediction of observations for the use of (2.3) based on an additive effects model given by,

$$\theta_{dg}^{-1} = \mu + \alpha_d + \beta_g, \quad \sum \alpha_d = \sum \beta_g = 0, \tag{2.4}$$

where μ, α_d's and β_g's represent the overall effect, the domain or row effects, and the group or column effects, respectively. For the inverse Gaussian distribution we must also have $\theta_{dg} > 0$ for all (d,g) and $\sigma > 0$. Thus the parameters μ, $\alpha = (\alpha_1, \alpha_2, \ldots, \alpha_D)$, $\beta = (\beta_1, \beta_2, \ldots, \beta_G)$, and σ lie in the set $\Omega = \{(\mu, \alpha, \beta, \sigma) : \sum_d \alpha_d = 0, \sum_g \beta_g = 0;\ \mu + \alpha_d + \beta_g > 0,\ \forall(d,g);\ \sigma > 0\}$. Under this setup estimation of parameters for prediction can be accomplished through unconditional maximization of the likelihood function. Conditional on the population and the sample sizes n_{dg} and referring to (2.1) and (2.3), the log-likelihood function of the parameters is given by

$$\ell = -\frac{1}{2}\log \sigma \sum_d \sum_g n_{dg}$$
$$-(2\sigma)^{-1}\sum_d \sum_g \sum_k y_{dgk}^{-1}[y_{dgk}(\mu + \alpha_d + \beta_g) - 1]^2. \tag{2.5}$$

We first note that the parameters are effectively given by $(\mu, \alpha_d, \beta_g, d = 1, 2, \ldots, D - 1; g = 1, 2, \ldots, G - 1)$. Thus, differentiating the above with respect to $(\mu, \alpha_d, \beta_g, d = 1, 2, \ldots, D - 1; g = 1, 2, \ldots, G - 1)$ and equating the resulting partial derivatives to zero gives the following equations for the estimators $(\hat{\mu}, \hat{\alpha}_d, \hat{\beta}_g, d = 1, 2, \ldots, D - 1; g = 1, 2, \ldots, G - 1)$,

$$\hat{\mu}y_{..} + \sum_{d=1}^{D-1} \hat{\alpha}_d(y_{d.} - y_{D.}) + \sum_{g=1}^{G-1} \hat{\beta}_g(y_{.g} - y_G) = n_{..},$$

$$\hat{\mu}(y_{d.} - y_{D.}) + \hat{\alpha}_d y_{d.} + \sum_{j=1}^{D-1} \alpha_j y_{D.}$$
$$+ \sum_{g=1}^{G-1} \hat{\beta}_g\{(y_{dg} - y_{Dg}) - (y_{dG} - y_{DG})\} = n_{d.} - n_{D.},$$

$$\hat{\mu}(y_{.g} - y_{.G}) + \sum_{d=1}^{D-1} \hat{\alpha}_d\{(y_{dg} - y_{dG}) - (y_{Dg} - y_{DG})\}$$
$$+ \hat{\beta}_g y_{.g} + \sum_{j=1}^{G-1} \hat{\beta}_j y_{.G} = n_{.g} - n_{.G}, \tag{2.6}$$

where the totals and means are represented by the notations

$$y_{dg} = \sum_k y_{dgk},\; y_{d.} = \sum_g y_{dg},\; y_{.g} = \sum_d y_{dg}, \qquad (2.7a)$$

$$n_{d.} = \sum_g n_{dg},\; n_{.g} = \sum_d n_{dg},\; n_{..} = \sum_d \sum_g n_{dg}. \qquad (2.7b)$$

The solutions $(\hat{\mu}, \hat{\alpha}_d, \hat{\beta}_g)$, $d = 1(1)D$, $g = 1(1)G$, provide the pseudo Maximum Likelihood estimator and may not yield nonnegative response estimates but will coincide with proper MLE as $n_{dg} \rightarrow \infty$ (see Fries and Bhattacharyya 1983) with probability one. Negative values of the response estimates may thus be truncated to zero.

In the case of the $IG(\theta, \sigma)$ model with interaction, the usual parameterization of the interaction effects suggests the model

$$\theta_{dg}^{-1} = \mu + \alpha_d + \beta_g + \gamma_{dg},$$

$$\sum \alpha_d = \sum \beta_g = \sum_d \gamma_{dg} = \sum_g \gamma_{dg} = 0, \qquad (2.8)$$

where now γ_{dg} is the interaction effect when domain is at the d-th level and group is at the g-th level. The estimators of parameters may be obtained in this case following the method outlined above. However, noting that the maximum likelihood estimator (MLE) of θ_{dg} is $\bar{y}_{dg}$ and there is one to one relation between the parameters in the reparametrized model in terms of $(\mu, \alpha_d, \beta_g, \gamma_{dg})$ and the original parameters θ_{dg}, explicit formulae for the MLE of different parameters are not needed. Corresponding to equation (2.3), therefore, for a two-factor model with interaction, our estimator is

$$\hat{t}_{dWI} = \sum_g N_{dg} \bar{y}_{dg}, \qquad (2.9)$$

which is the post stratified estimator and is not of further interest in small area estimation. For the model without interaction, the estimator is given as

$$\hat{t}_{dWOI} = \sum_g N_{dg} \hat{\theta}_{dg} + \sum_g \hat{N}_{dg} (\bar{y}_{dg} - \hat{\theta}_{dg}), \qquad (2.10)$$

where $\hat{\theta}_{dg}^{-1} = \hat{\mu} + \hat{\alpha}_d + \hat{\beta}_g$, the estimators being obtained from (2.6) and $\hat{N}_{dg} = n_{dg} N / n_{..}$.

In order to judge the effectiveness of this estimator a numerical study has been performed and is reported in the following section.

3. A NUMERICAL STUDY OF THE INVERSE GAUSSIAN REGRESSION ESTIMATOR

In this section we provide the results of a simulation study which evaluates the performance of the estimators developed in the previous section. The modified regression estimator due to Särndal and Hidiroglou (1989) given below will be used as the bench mark for the above purpose;

$$\hat{t}_{dS-H} = \sum_g N_{dg} \bar{y}_{.g} + \sum_g F_d \hat{N}_{dg} (\bar{y}_{dg} - \bar{y}_{.g}), \qquad (3.1)$$

where $F_d = N_{d.}/\hat{N}_{d.}$ if $\hat{N}_{d.} \geq N_{d.}$, otherwise $F_d = \hat{N}_{d.}/N_{d.}$. Here, $\hat{N}_{d.} = n_{d.} N / n_{..}$. An alternative form of this estimator which takes into account both group and domain effects can be obtained by replacing $\bar{y}_{.g}$ by $\bar{y}_{.g} + \bar{y}_{d.} - \bar{y}_{..}$ but this has not been pursued here. It should be noted that the above estimators cannot be computed when n_{dg} is zero. When this happens the estimators are simply taken to be the sample means of the respective domains. We also include the following modified version of $\hat{t}_{dWOI}$,

$$\hat{t}_{dWOIM} = \sum_g N_{dg} \hat{\theta}_{dg} + \sum_g F_d \hat{N}_{dg} (\bar{y}_{dg} - \hat{\theta}_{dg}), \qquad (3.2)$$

for comparison.

3.1 Design of the Simulation Study

We consider Household Income data for Canadians in 1986, obtained from Household Income, Facilities and Equipment microdata tape of Statistics Canada (1987), for generating the values of parameters to be used for simulation. Using Household incomes, from these data, dividing them into 10 provinces and 6 educational groups, we first fit an inverse Gaussian model given by equation (2.4). The estimates of parameters are then used in forming the true parameters of the inverse Gaussian super population model which are summarized in appendix A. The values of D, G, N_{dg} are chosen from this population (see appendix B), where D represents the number of provinces (*i.e.*, $D = 10$) and G represents the number of education groups (*i.e.*, $G = 6$). Further sets of values of θ_{dg} and σ are obtained by considering various combinations of (c_1, c_2); $c_1 = 0(1)4$ and $c_2 = 1, .25, .1, .01$ where c_1 is used to transform θ_{dg} to $10^{-c_1}\theta_{dg}$ and c_2 is used to transform σ to $c_2 \sigma$. Note that $c_1 = 0$ and $c_2 = 1$ gives the parameter values for the original population. Also, the higher values of c_1 indicate smaller values of the means and those of c_2 indicate higher value of the dispersion parameter.

For the simulation study, first we generate for a given set of θ_{dg} and σ values an inverse Gaussian random sample using the algorithm in Michael *et al.* (1976) with number of observations according to the values given in the appendix B. This random sample is then used as a finite population from which we select 1000 random samples for each of the sample fractions, 1%, and 5% with replacement. We had actually selected several random samples and obtained similar results as reported here. From each sample we computed the estimators of totals for the 10 domains using estimators $\hat{t}_{dS-H}$, $\hat{t}_{dWOI}$ and $\hat{t}_{dWOIM}$. The criteria for evaluating the performance of the estimators are the mean absolute relative error (MARE) and the absolute relative bias (ARB) defined as follows:

$$\text{MARE}(\hat{t}_d) = \frac{1}{1000} \sum_{i=1}^{1000} \mid \hat{t}_{di} - t_d \mid / t_d \qquad (3.3)$$

$$\text{ARB}(\hat{t}_d) = \left| \frac{1}{1000} \sum_{i=1}^{1000} \hat{t}_{di} - t_d \right| / t_d. \qquad (3.4)$$

Here $\hat{t}_d$ denotes a typical estimator of t_d and $\hat{t}_{di}$ denotes the value of the i-th Monte Carlo sample ($i = 1, \ldots, 1000$).

Table 1

Mean Absolute Relative Error (%) of Different Estimators

Domain	1% Sample			5% Sample			1% Sample			5% Sample		
	SH	WOI	WOIM	SH	WOI	WOIM	SH	WOI	WOIM	SH	WOI	WOIM
	$c_1 = 0, c_2 = 1$						$c_1 = 0, c_2 = .01$					
1	13.27	13.05	13.19	6.60	6.48	6.47	3.72	2.46	2.45	1.80	0.89	0.89
2	14.57	13.61	14.20	7.53	7.61	7.69	3.79	3.56	3.48	2.10	0.59	0.60
3	25.27	27.86	26.88	19.07	20.74	20.80	2.52	1.51	1.52	1.19	0.77	0.77
4	11.83	11.70	11.74	5.29	5.61	5.59	1.83	1.08	1.09	0.93	0.58	0.58
5	10.57	11.72	11.68	6.80	7.10	7.11	0.92	0.90	0.91	0.42	0.40	0.40
6	7.12	7.45	7.52	3.85	3.95	3.97	1.94	1.22	1.22	0.93	0.64	0.64
7	11.78	13.91	14.23	7.39	8.01	8.05	1.22	1.13	1.14	0.86	0.64	0.64
8	11.48	12.56	12.46	6.70	7.15	7.14	1.29	0.93	0.94	0.76	0.67	0.68
9	7.43	7.92	7.99	3.61	3.74	3.75	3.47	2.99	2.96	3.13	2.97	2.96
10	15.32	17.43	17.16	11.20	11.81	11.80	0.93	0.94	0.95	0.52	0.52	0.53
	$c_1 = 2, c_2 = 1$						$c_1 = 2, c_2 = .01$					
1	3.34	2.18	2.15	1.66	0.79	0.78	2.99	1.48	1.44	1.47	0.08	0.08
2	4.14	3.94	3.82	2.14	1.07	1.06	0.54	3.37	3.27	1.86	0.14	0.13
3	2.44	1.67	1.65	1.17	0.71	0.70	1.81	0.45	0.44	0.87	0.07	0.07
4	2.05	1.70	1.69	0.98	0.70	0.70	1.32	0.36	0.35	0.66	0.07	0.07
5	1.08	1.17	1.16	0.50	0.51	0.51	0.27	0.13	0.13	0.11	0.05	0.05
6	1.74	1.14	1.14	0.78	0.52	0.52	1.29	0.13	0.13	0.55	0.05	0.05
7	1.90	1.57	1.56	0.91	0.72	0.72	1.22	0.31	0.31	0.56	0.07	0.07
8	1.48	1.38	1.38	0.70	0.60	0.60	0.81	0.18	0.18	0.38	0.06	0.06
9	1.41	1.30	1.29	0.67	0.59	0.58	0.69	0.14	0.14	0.30	0.06	0.06
10	1.22	1.38	1.38	0.56	0.59	0.59	0.26	0.15	0.15	0.10	0.06	0.06
	$c_1 = 4, c_2 = 1$						$c_1 = 4, c_2 = .01$					
1	2.99	1.48	1.44	1.47	0.08	0.08	2.99	1.45	1.41	1.47	0.01	0.01
2	3.54	3.37	3.27	1.86	0.14	0.13	3.54	3.36	3.25	1.87	0.05	0.05
3	1.81	0.45	0.44	0.87	0.07	0.07	1.80	0.38	0.37	0.86	0.01	0.01
4	1.32	0.36	0.35	0.66	0.07	0.07	1.31	0.28	0.27	0.66	0.01	0.01
5	0.27	0.13	0.13	0.11	0.05	0.05	0.24	0.06	0.06	0.10	0.01	0.01
6	1.29	0.13	0.13	0.55	0.05	0.05	1.29	0.06	0.06	0.54	0.01	0.01
7	1.22	0.31	0.31	0.56	0.07	0.07	1.20	0.24	0.24	0.55	0.01	0.01
8	0.81	0.18	0.18	0.38	0.06	0.06	0.79	0.09	0.09	0.37	0.01	0.01
9	0.69	0.14	0.14	0.30	0.06	0.06	0.68	0.06	0.06	0.29	0.01	0.01
10	0.26	0.15	0.15	0.10	0.06	0.06	0.23	0.07	0.07	0.09	0.01	0.01

Table 2
Absolute Relative Bias (%) of Different Estimators

Domain	1% Sample			5% Sample			1% Sample			5% Sample		
	SH	WOI	WOIM	SH	WOI	WOIM	SH	WOI	WOIM	SH	WOI	WOIM
	$c_1 = 0, c_2 = 1$						$c_1 = 0, c_2 = .01$					
1	4.34	2.40	2.51	1.87	0.26	0.27	2.66	1.58	1.54	1.22	0.03	0.03
2	8.88	3.46	4.39	2.18	0.30	0.23	3.15	3.40	3.31	1.38	0.04	0.04
3	3.13	3.47	2.74	0.51	1.12	1.15	1.44	0.31	0.32	0.68	0.01	0.01
4	1.57	0.51	0.53	0.50	0.21	0.22	1.11	0.29	0.30	0.53	0.03	0.03
5	0.13	0.33	0.35	0.20	0.16	0.18	0.10	0.03	0.02	0.05	0.01	0.01
6	1.09	0.14	0.04	0.02	0.39	0.42	1.09	0.03	0.03	0.43	0.02	0.01
7	1.20	1.09	1.59	0.54	0.28	0.30	0.99	0.22	0.23	0.43	0.01	0.01
8	0.40	0.04	0.12	0.20	0.53	0.54	0.55	0.00	0.01	0.28	0.03	0.03
9	1.03	0.47	0.36	0.24	0.04	0.01	1.01	0.35	0.37	0.45	0.14	0.14
10	1.05	2.27	2.03	0.04	0.30	0.29	0.08	0.02	0.01	0.06	0.01	0.01
	$c_1 = 2, c_2 = 1$						$c_1 = 2, c_2 = .01$					
1	2.40	1.37	1.33	1.13	0.01	0.01	2.47	1.43	1.39	1.15	0.01	0.01
2	3.00	3.28	3.16	1.33	0.02	0.01	3.06	3.34	3.24	1.36	0.03	0.03
3	1.53	0.39	0.38	0.70	0.04	0.04	1.46	0.35	0.34	0.65	0.01	0.01
4	1.00	0.25	0.25	0.53	0.04	0.04	1.01	0.23	0.23	0.49	0.00	0.00
5	0.10	0.02	0.03	0.04	0.00	0.01	0.10	0.01	0.02	0.04	0.00	0.00
6	1.16	0.01	0.01	0.47	0.02	0.02	1.15	0.01	0.00	0.46	0.00	0.00
7	1.00	0.27	0.27	0.42	0.00	0.00	0.95	0.21	0.21	0.41	0.00	0.00
8	0.48	0.04	0.04	0.25	0.01	0.01	0.57	0.04	0.04	0.26	0.00	0.00
9	0.64	0.06	0.05	0.27	0.02	0.02	0.61	0.01	0.00	0.26	0.00	0.00
10	0.01	0.02	0.02	0.02	0.00	0.00	0.06	0.01	0.01	0.03	0.00	0.00
	$c_1 = 4, c_2 = 1$						$c_1 = 4, c_2 = .01$					
1	2.47	1.43	1.39	1.15	0.01	0.01	2.48	1.43	1.39	1.15	0.00	0.00
2	3.06	3.34	3.24	1.36	0.03	0.03	3.07	3.35	3.24	1.36	0.04	0.04
3	1.46	0.35	0.34	0.65	0.01	0.01	1.45	0.34	0.34	0.64	0.00	0.00
4	1.01	0.23	0.23	0.49	0.00	0.00	1.01	0.24	0.24	0.49	0.00	0.00
5	0.10	0.01	0.02	0.04	0.00	0.00	0.11	0.01	0.02	0.04	0.00	0.00
6	1.15	0.01	0.00	0.46	0.00	0.00	1.15	0.01	0.00	0.46	0.00	0.00
7	0.95	0.21	0.21	0.41	0.00	0.00	0.94	0.20	0.20	0.41	0.00	0.00
8	0.57	0.04	0.04	0.26	0.00	0.00	0.58	0.04	0.05	0.26	0.00	0.00
9	0.61	0.01	0.00	0.26	0.00	0.00	0.60	0.00	0.00	0.25	0.00	0.00
10	0.06	0.01	0.01	0.03	0.00	0.00	0.06	0.01	0.01	0.03	0.00	0.00

3.2 Analysis of Results

The MARE values computed according to (3.3) and the ARB values from (3.4) for the three estimators and for different sample sizes are reported in Tables 1 and 2, respectively for a selection of pairs (c_1, c_2). The values of c_1 are chosen to represent, large means (as in the original population, $c_1 = 0$), moderate means ($c_1 = 2$) and small means ($c_1 = 4$), whereas, the values chosen for c_2 represent the original dispersion parameter ($c_2 = 1$) and a further smaller value ($c_2 = .01$). It may be interesting to note that increasing c_1 by 1 while keeping c_2 fixed reduces the coefficient of variation by a factor of 10.

Some of the MARE and ARB values reported in Tables 1 and 2 are also plotted for visual inspection in Figures 1 and 2 for 1% samples, respectively.

When comparing the MARE and ARB values, reductions in biases as well as in relative errors are observed in many cases for both 1% and 5% samples. It is found that, the MARE and ARB values decrease with decreasing values of mean and dispersion parameter σ. Reductions

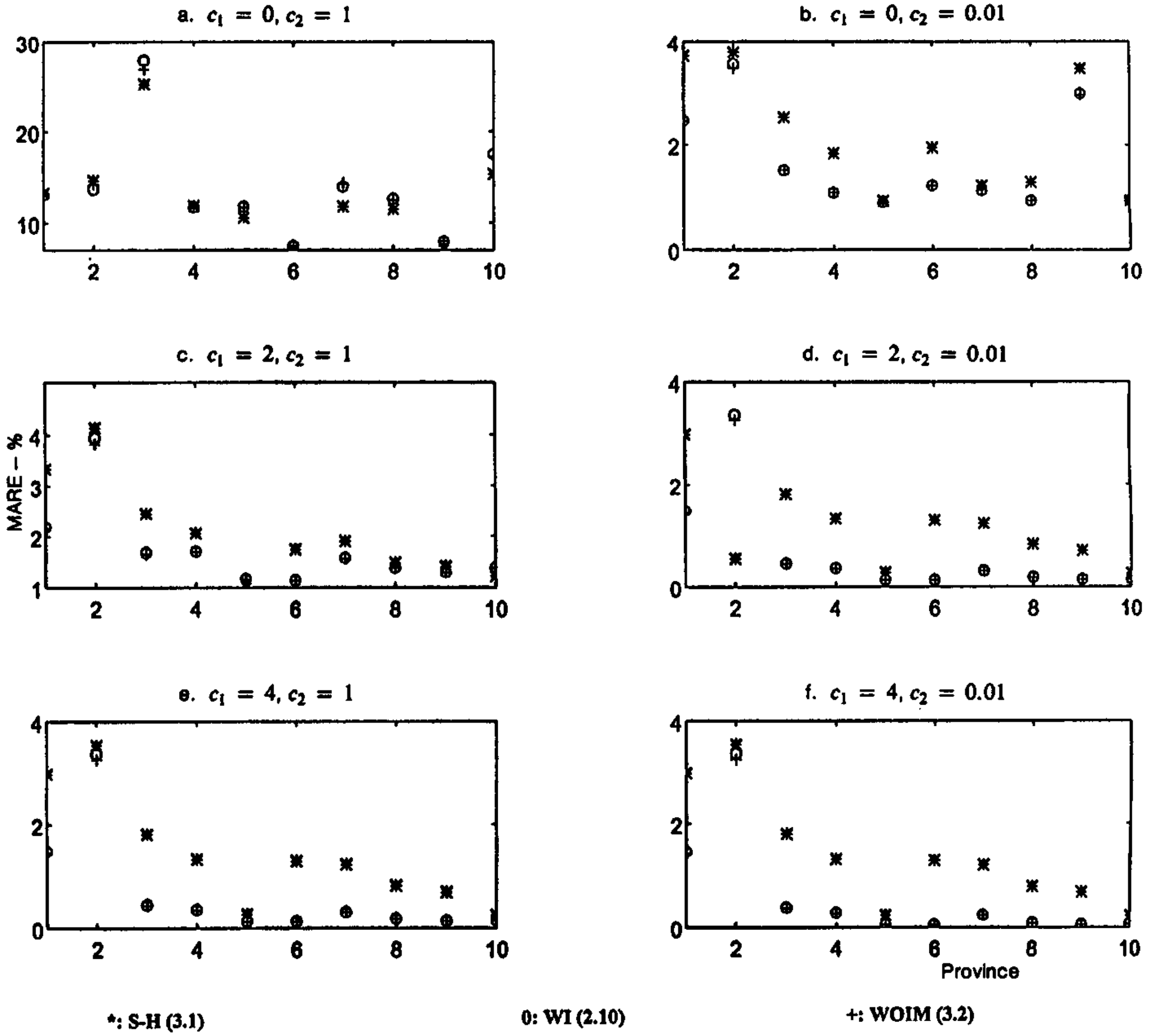

Figure 1. Mean absolute relative errors for different estimators for 1% sample.

are substantial, especially in case of 5% sample and/or when means are small. Note also that the reductions in bias are generally larger than reductions in the errors. We may note from Johnson and Kotz (1970, p. 141) that for fixed value of the mean, the standardized inverse Gaussian distribution tends to unit normal as the coefficient of variation tends to zero. Since larger gains in MARE and ARB values are noted for small values of the coefficient of variation, we conclude that proper modeling of the mean is important when the coefficient of variation is small for model based estimation.

We further find that $\hat{t}_{dWOI}$ and $\hat{t}_{dWOIM}$ have almost same MARE and ARB which indicates that the modification of the estimator in (2.10) is not necessary. It may be remarked that the estimator $\hat{t}_{dS-H}$, in contrast, has been demonstrated (see Hidiroglou and Särndal 1985) to be substantial improvement over the corresponding unmodified estimator due to Särndal (1984).

Owing to the criticism of $\hat{t}_{dWOI}$ and $\hat{t}_{dWOIM}$ as being model dependent, we want to defend these on the following grounds. The inverse Gaussian distribution offers a variety of shapes and may be able to approximate lognormal, gamma, Weibull and such other positively skewed shapes. If we suspect that the principal characteristic is positively skewed, then the methodology we discussed here is viable and useful.

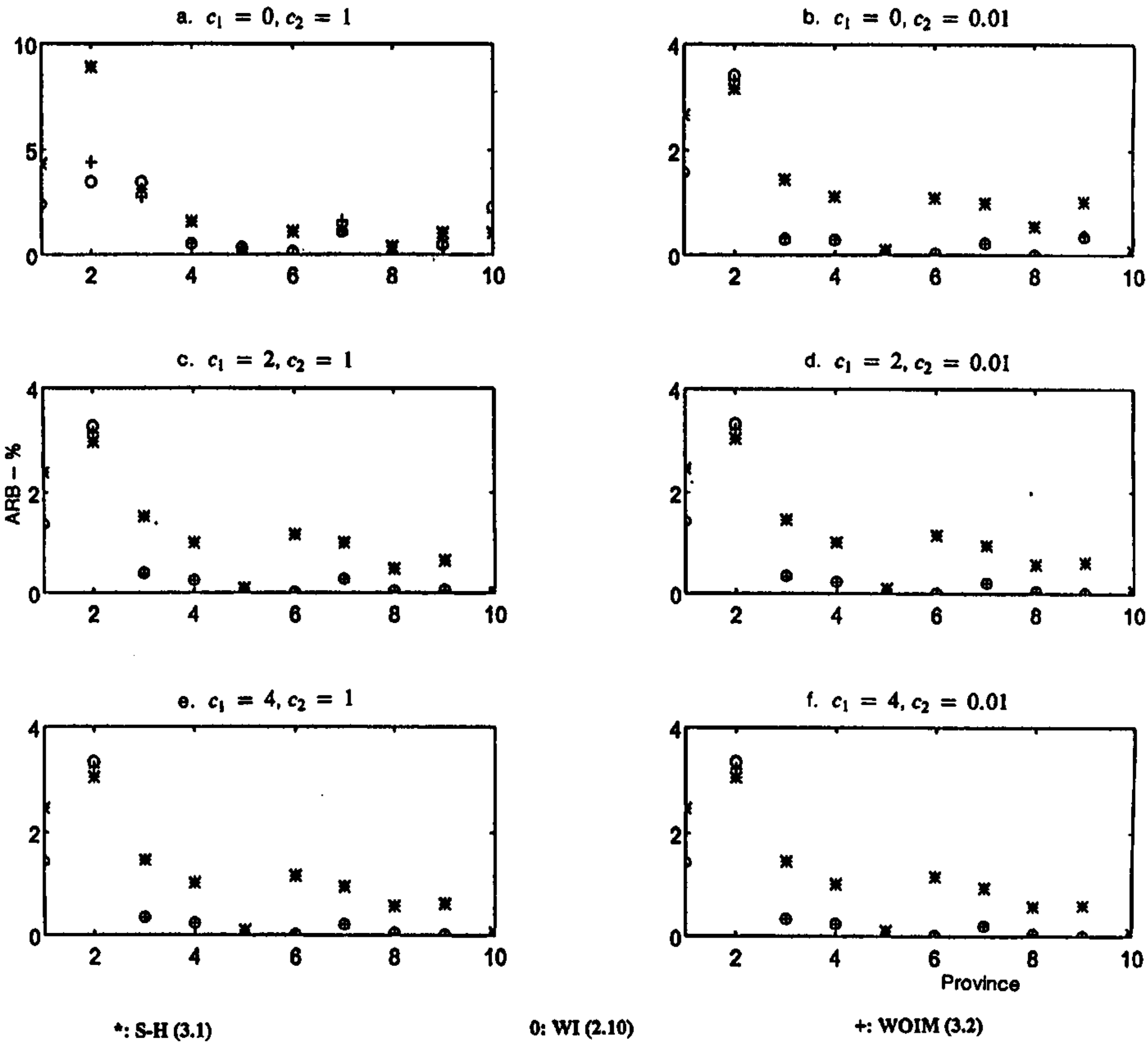

Figure 2. Absolute relative biases for different estimators for 1% sample.

4. SUMMARY AND CONCLUSIONS

The generalization of analysis of variance methodology for inverse Gaussian population for unbalanced design was considered. The models without interactions of factors were studied and applied to the problem of estimation of small area parameters in finite populations. Using Canadian survey data, synthetic populations were generated in a Monte Carlo study. Through this we demonstrated that the proposed estimators perform well under a variety of conditions when the population can be regarded as a random sample from some inverse Gaussian distribution.This approach offers a competitive choice for estimation of parameters in positively skewed survey data.

ACKNOWLEDGEMENTS

Support to Y.P. Chaubey and F. Nebebe by the Natural Sciences and Engineering Research Council of Canada is gratefully acknowledged. The authors thank the Editor, Professors J.N.K. Rao and A.B. Sim and the referees for their helpful comments.

APPENDIX A

Values of the Parameters for Generation of the IG Population

$\mu = 3.13241147 \times 10^{-5}, \quad \sigma = 2.5447984 \times 10^{-5}$

d	1	2	3	4	5
$10^6 \times \alpha_d$	3.1902855	2.8235779	1.5676078	.8056079	−.95350458

d	6	7	8	9	10
$10^6 \times \alpha_d$	−4.0661125	.49944356	.0061694263	−2.7414128	−1.1316622

g	1	2	3	4	5	6
$10^5 \times \beta_g$	1.0938451	.36781639	−.012707035	−.11561414	−.30936835	−1.023972

θ_{dg} values:

d/g	1	2	3	4	5	6
1	22,000.82	26,183.11	29,080.48	29,977.59	31,826.13	41,195.19
2	22,179.76	26,436.94	29,393.94	30,310.79	32,201.96	41,827.05
3	22,815.33	27,344.90	30,520.70	31,510.37	33,559.25	44,146.20
4	23,219.00	27,926.81	31,247.41	32,285.58	34,439.96	45,682.95
5	24,207.76	29,369.63	33,064.91	34,229.61	36,661.02	49,674.90
6	26,180.44	32,324.63	36,858.30	38,311.45	41,383.33	58,760.34
7	23,385.24	28,167.65	31,549.24	32,607.90	34,806.97	46,330.96
8	23,658.15	28,564.53	32,047.98	33,140.96	35,415.03	47,414.57
9	25,302.90	30,997.31	35,142.43	36,461.01	39,232.58	54,516.76
10	24,312.62	29,524.12	33,260.85	34,439.64	36,902.04	50,118.45

APPENDIX B

Values of the Cell Sizes N_{dg}

d/g	1	2	3	4	5	6	Total
1	627	360	277	84	215	110	1,673
2	285	212	198	72	68	83	918
3	597	483	616	148	204	231	2,279
4	729	397	568	151	239	219	2,303
5	1,372	761	1,216	202	473	511	4,535
6	1,177	888	1,795	517	707	800	5,884
7	639	432	673	165	236	222	2,367
8	850	512	888	264	349	297	3,160
9	700	699	1,350	385	696	572	4,401
10	456	540	1,083	342	393	407	3,221

REFERENCES

BHATTACHARYYA, G.K., and FRIES, A. (1986). On the inverse Gaussian multiple regression and model checking procedures. In *Reliability and Quality Control*, (A.P. Basu, Ed.). New York: North Holland, 86-100.

CHAUBEY, Y.P. (1991). A study of ratio and product estimators under super population. *Communications in Statistics*, A, 20 (5 and 6), 1731-1746.

CHOUDHRY, G.H., and RAO, J.N.K. (1988). Evaluation of small area estimators: An empirical study. Paper presented at the International Symposium on Small Area Statistics, New Orleans.

CHHIKARA, R.S., and FOLKS, J.L. (1989). *The Inverse Gaussian Distribution*. New York: Marcel Dekker, Inc.

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

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

FOLKS, J.L., and CHHIKARA, R.S. (1978). The inverse Gaussian distribution and its statistical applications-A review. *Journal of the Royal Statistical Society*, Series B, 40, 263-275.

FRIES, A., and BHATTACHARYYA, G.K. (1983). Analysis of two-factor experiments under an inverse Gaussian model. *Journal of the American Statistical Association*, 78, 820-826.

GONZALES, M.E., and HOZA, C. (1978). Small area estimation with application to unemployment and housing estimates. *Journal of the American Statistical Association*, 73, 7-15.

GHOSH, M., and RAO, J.N.K. (1994). Small area estimation: An appraisal. *Statistical Science*, 9, 55-76.

HIDIROGLOU, M.A., and SÄRNDAL, C.-E. (1985). An empirical study of some regression estimations for small domains. *Survey Methodology*, 6, 65-77.

HOLT, D., SMITH, T.M.F., and TOMBERLIN, T.J. (1979). A model-based approach to estimation for small subgroups of a population. *Journal of the American Statistical Association*, 74, 405-410.

IYENGAR, S., and PATWARDHAN, G. (1988). Recent developments in the inverse Gaussian distribution. In *Handbook of Statistics*, New York: Elsevier Science 479-490.

JOHNSON, N. L., and KOTZ, S. (1970). *Continuous Univariate Distributions-1, Distributions in Statistics*. New York: Wiley.

MACGIBBON, B., and TOMBERLIN, T.J. (1989). Small area estimates of proportions via empirical Bayes techniques. *Survey Methodology*, 15, 237-252.

MICHAEL, J.R., SCHUCANY, W.R., and HASS, R.W. (1976). Generating random variables using transformations with multiple roots. *American Statistician*, 30(2), 88-90.

PRASAD, N.G.N., and RAO, J.N.K. (1990). The estimation of the mean squared error of small-area estimators. *Journal of the American Statistical Association*, 85, 163-171.

PURCELL, N.J., and KISH, L. (1980). Postcensal estimates for local areas (or domains). *Bulletin of the International Statistical Institute*, 48, 3-18.

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

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

SÄRNDAL, C.-E., and RÅBÄCK, G. (1983). Variance reduction and unbiasedness for small domain estimators. *Statistical Review*, 21, (Essays in Honor of T.E. Dalenius), 33-40.

SCHAIBLE, W.L. (1979). A composite estimator for small area statistics. In *Synthetic Estimates for Small Areas*, NIDA Research Monograph 24, (J. Steinberg, Ed.). Rockville, MD: National Institute on Drug Abuse, 36-53.

STATISTICS CANADA (1987). Microdata file, Household Income, Facilities and Equipment (1987), Statistics Canada, Household Surveys Division.

STROUD, T.W.F. (1987). Bayes and empirical Bayes approaches to small area estimation. In *Small Area Statistics*, (R.Platek, J.N.K. Rao, C.-E. Särndal and M.P. Singh, Eds.). New York: Wiley, 124-137.

WHITMORE, G.A. (1983). A regression method for censored inverse Gaussian data. *The Canadian Journal of Statistics*, 11, 305-315.

Survey Methodology, June 1996
Vol. 22, No. 1, pp. 43–53
Statistics Canada

A Comparison of Some Weighting Adjustment Methods for Panel Nonresponse

LOU RIZZO, GRAHAM KALTON and J. MICHAEL BRICK[1]

ABSTRACT

In some surveys, many auxiliary variables are available for respondents and nonrespondents for use in nonresponse adjustment. One decision that arises is how to select which of the auxiliary variables should be used for this purpose and another decision involves how the selected variables should be used. Several approaches to forming weighting adjustments for nonresponse are considered in this research. The methods include those based on logistic regression models, categorical search algorithms, and generalized raking. These methods are applied to adjust for panel nonresponse in the Survey of Income and Program Participation (SIPP). The estimates from the alternative adjustments are assessed by comparing them to one another and to benchmark estimates from other sources.

KEY WORDS: Nonresponse bias; Panel surveys; Generalized raking; Benchmark estimates.

1. INTRODUCTION

Weights are commonly used in the analysis of survey data to compensate for unequal selection probabilities of the sampled elements, to compensate for unit nonresponse, and to make the weighted sample distributions for certain variables conform to known population distributions for those variables (thereby aiming to compensate for noncoverage and to improve the precision of the survey estimates) (Kish 1992). Corresponding to these three objectives, the weights are usually developed in three stages. First, a base weight is calculated for each sampled element as the inverse of the element's selection probability. Second, the base weights of responding sampling elements are multiplied by a nonresponse weight to compensate for the nonrespondents. Third, the adjusted weight is modified to make the weighted sample distributions for certain variables conform to external information on these distributions.

This paper deals with the nonresponse adjustment weights that attempt to compensate for unit nonresponse. A commonly used procedure for obtaining these weights is to divide the total sample into a set of weighting classes based on information known for both respondents and nonrespondents, and then to increase the base weights for the respondents in a weighting class to represent the nonrespondents in that class (Oh and Scheuren 1983; Kalton 1983). In many surveys little information is known about the nonrespondents, beyond the primary sampling units and strata from which they come. In this case, the choice of possible weighting classes is limited, and the procedure can be applied fairly straightforwardly.

In some surveys, however, there is an extensive amount of information available for the nonrespondents. This information may be available from the sampling frame (*e.g.*, when sampling employees from personnel files) or by matching sampled elements with administrative records. Also, in panel surveys and other surveys involving more than one stage of data collection, extensive information on nonrespondents at later stages is available from their responses at the early stages.

The major focus of this research is on methods for developing weighting adjustments for nonresponse when a large number of characteristics of the nonrespondents are known. In this situation, decisions about methods of adjusting for nonresponse involve selecting which auxiliary variables will be used and how they will be used to make the adjustments.

The main ideas are presented in this article by applying several different adjustment procedures in a specific panel survey, the Survey of Income and Program Participation (SIPP). The SIPP is an ongoing household panel survey conducted by the U.S. Bureau of the Census. The nonrespondents to a SIPP panel can be separated in two groups: those who fail to respond at the initial wave of data collection (initial wave nonrespondents), and those who respond at the initial wave but fail to respond at one or more of the subsequent waves of the panel for which they are eligible (panel nonrespondents). For the latter group, extensive information from the initial wave of data collection can be utilized in adjusting for panel nonresponse. The weighting adjustments studied here relate to the panel nonrespondents only. These adjustments modify the weights of panel respondents (*i.e.*, those who provide data for all waves for which they are eligible) to compensate for the panel nonrespondents.

In the SIPP, a national probability sample of households is interviewed each year, and all the adults aged 15 and over living in those households at the initial wave become panel members who are followed for the duration

[1] Lou Rizzo, Graham Kalton and J. Michael Brick, Westat Inc., 1650 Research Blvd., Rockville, MD 20850, U.S.A.

of the panel. Until now SIPP panels have had a lifetime of 2 ⅔ years, but this is being increased with the 1996 panel to 4 years. Interviews are conducted with panel members at four-month intervals to collect data about income amounts received, participation in income maintenance programs, and other factors that may affect their income and economic welfare. Data are also collected about children. See Nelson, McMillen and Kasprzyk (1985) and Jabine, King and Petroni (1990) for further information on the SIPP design.

The investigation reported here was conducted with the 1987 SIPP panel, using the panel's public use data file. That panel started with a sample of about 12,300 households and followed panel members for seven waves of data collection. The household nonresponse rate at the initial wave was 6.7 percent (Jabine *et al.* 1990). Including children, 30,841 individuals were living in the responding households at the initial wave. Of these individuals, 20.8 percent failed to provide data for all waves for which they were eligible, *i.e.*, they were panel nonrespondents.

In addition to selecting auxiliary variables and studying alternative methods of using those variables to form weighting adjustments for panel nonresponse, this research includes a comparative evaluation of the procedures. The evaluation is performed by comparing a range of estimates produced with the alternative methodologies with one another and with benchmark estimates. The final section of this article summarizes the results and draws conclusions about the effectiveness of the alternative weighting schemes investigated. Further details are given by Rizzo, Kalton, and Brick (1994).

2. PREDICTORS OF RESPONSE PROPENSITY

The first step in developing panel nonresponse adjustments is deciding which of the large number of items available from the first wave of data collection should be selected for use in the adjustment procedures. That selection is the focus of this section. The approach adopted is to choose items with responses that discriminate persons by their likelihood to respond at all later waves. Little (1986) calls this method a response propensity stratification method and shows that the large sample bias of estimates can be reduced by adjusting the base weight by the inverse of the probability that an element responds.

In the 1987 SIPP panel, there were 58 items available from the initial wave of data collection (Wave 1) that could be used as potential explanatory variables for panel nonresponse. All of the items used currently by the Bureau of the Census for the SIPP panel nonresponse adjustment were part of this set of 58, with the exception of the Metropolitan Statistical Area (MSA) status, which was suppressed from the public use data file because of disclosure concerns.

With panel response status (panel respondent *vs.* panel nonrespondent) as the dependent variable, logistic regression analysis was viewed as a natural method for selecting a model for panel nonresponse. However, before attempting this modeling, an initial screening of the variables was performed to reduce the large number of variables to a more manageable set. As a general guideline, items were retained for the logistic regression analysis if the difference in response rates between any two categories for the item was both statistically significant and at least four percentage points. For a variety of reasons, some items were retained even if they did not meet these requirements. For example, the difference in the panel response rates for males and females was less than 2 percent, but gender was nevertheless used in some subsequent analyses.

The screening process reduced the number of items for the logistic regression analysis from 58 to 31. The items retained were: tenure, public housing, household type, Census region, household education, household size, household income, whether householder holds financial instruments (bonds), gender, race, Hispanic origin, relationship to reference person (RRP), age, marital status, family type, education, student status, whether laid off work, personal income, whether holds multiple jobs, working class, whether a recipient of Medicare benefits, Medicaid, Women, Infants, and Children (WIC), Aid to Families with Dependent Children (AFDC), food stamps, general assistance, Social Security, other welfare, Veteran's status, and the number of imputed items at Wave 1.

The last item, the number of imputed items, was included as an index of cooperation at Wave 1. Other studies have found that individuals who are less cooperative at the initial wave of a panel survey are more likely to be nonrespondents at later waves (see, for example, Kalton, Lepkowski, Montanari and Maligalig 1990). As described below, this index turned out to be highly related to panel nonresponse.

2.1 Logistic Regression Analysis

Since all 31 items identified in the screening analysis were at least marginally correlated with panel nonresponse, they are all candidate variables for use in a weighting adjustment scheme to reduce the panel nonresponse bias in the survey estimates. However, the screening analysis was limited because it did not consider the interrelationships between the items and it retained too many variables for practical use in making the panel nonresponse adjustments. For example, two items that are highly associated with response status might also be highly correlated with each other, so that the use of one of the two might be sufficient in making the adjustments. To address this issue, the next step in selecting predictors of panel nonresponse was to investigate which combinations of the items could best predict panel response status.

Table 1

Parameter Estimates for the Logistic Regression Model

Predictors	Parameter Estimate
Intercept	-0.465
Age (χ^2 = 184.9, p-value < .0001).	
< 16	-0.179
16-24	0.446
25-50	0.187
51-71	-0.056
> 71	0.0
Race (χ^2 = 214.0, p-value < .0001).	
White	-0.351
Black	0.255
Other	0.0
RRP (χ^2 = 69.0, p-value < .0001).	
Family member	-0.251
Nonfamily member	0.0
Census region (χ^2 = 327.3, p-value < .0001).	
New England	0.009
Mid Atlantic	0.167
South Atlantic	0.027
East South Central	-0.231
North Central	-0.396
Mountain/West South Central	0.425
Pacific	0.0
Tenure (χ^2 = 207.2, p-value < .0001).	
Home owner	-0.154
Renter	0.331
Other	0.0
Items imputed (χ^2 = 434.2, p-value < .0001).	
0	-0.626
1	-0.244
2 to 3	0.296
> 3	0.0
Bond status (χ^2 = 97.1, p-value < .0001).	
No bonds	0.168
Some bonds	0.0
Layoff (χ^2 = 33.4, p-value < .0001).	
Not laid off	-0.179
Laid off	0.0
Food stamps (χ^2 = 39.3, p-value < .0001).	
Not recipient	-0.191
Recipient	0.0
Class of work (χ^2 = 31.4, p-value < .0001).	
Business	0.100
Other	0.103
Government	0.0
Education (χ^2 = 12.8, p-value = .0003).	
Last grade tenth or eleventh	-0.075
Other	0.0
Household income (χ^2 = 14.9, p-value = .0006).	
Less than $1,200/month	0.117
$1,200-$8,000/month	-0.088
Greater than $8,000/month	0.0
Gender (χ^2 = 10.3, p-value = .0013).	
Male	0.047
Female	0.0
RRP-Age < 16 Interaction (χ^2 = 10.1, p-value = .0015).	
Family member, child	0.096
Other	0.0

A logistic regression approach was used to the examine the joint relationships of several items with panel response status. The regression models were fitted using the Wave 1 survey weights that accounted for unequal selection probabilities and initial wave nonresponse. After examining a number of possible models, a model with thirteen main-effect variables and one interaction term was selected as a reasonable representation of the data.

Table 1 presents the parameter estimates for each level of each predictor variable in this model, together with Wald (χ^2) statistics for each predictor variable. The parameter value of the last level of each predictor variable (the benchmark level) is set to zero. The parameter estimates for the remaining levels of each predictor variable represent differences in response propensity from the benchmark level. As can be seen from the Wald statistics, all the predictor variables make highly significant contributions to the model.

A notable feature of this model is that it contains only one interaction term, the relationship to reference person/age under 16 interaction. All other interactions investigated had smaller χ^2 values than this one. Even the relationship to reference person/age under 16 interaction has a relatively low predictive power. In fact, this interaction and the last three predictor variables in Table 1 (education, household income, and gender) were not included in most of the weighting procedures discussed below because of their limited predictive power for panel response status. The weighting procedures are mostly based on a reduced main-effects model comprising the first ten predictor variables listed in Table 1.

3. ALTERNATIVE WEIGHT ADJUSTMENTS

The method used in the SIPP to adjust the weights for panel nonresponse is described by Chapman, Bailey, and Kasprzyk (1986). The method basically consists of forming nonresponse adjustment cells and then adjusting the weights by the inverses of the response rates in the cells. The cells are formed by the cross-classification of the responses from a set of Wave 1 variables thought to be correlated with panel response. Small cells are combined so that the resulting sample size in each collapsed cell is 30 or more. The reciprocal of the observed (weighted) response rate in each collapsed cell is the panel nonresponse adjustment for that cell. The panel nonresponse adjustment is then multiplied by the Wave 1 weight to create a nonresponse adjusted weight. The Wave 1 weight includes an adjustment for Wave 1 nonresponse, but it does not include the Wave 1 poststratification adjustment.

This section examines alternative methods for performing the panel nonresponse adjustments. These methods can be categorized into three groups:

- Logistic regression methods.
- CHAID methods.
- Generalized raking methods.

Each of the alternative approaches to nonresponse adjustment is discussed below. The procedures for developing the weighting adjustments are detailed along with important statistical properties of the adjustments.

3.1 Adjustments Based on Logistic Models

The first set of weighting adjustments we discuss is developed directly from the logistic regression model described in the previous section. This panel nonresponse weighting adjustment, called the *predicted logistic adjustment*, was computed by taking the inverses of the response rates predicted from the reduced main-effects logistic regression model for each of the cells in the crossclassification of the ten predictor variables in that model.

Since the parameters for computing the predicted response rates are estimated with a main-effects model from the marginal responses for the variables, the small sample sizes in the cells of the crossclassification of all the variables are not a concern. However, this benefit is gained by relying completely on the validity of the main-effects model, that is, by assuming that there are no interactions between the variables that need to be taken into account.

One approach to placing less reliance on the main-effects model is to base the adjustments on the observed response rates in cells that have sample sizes large enough to ensure the stability of the observed response rates and to base the adjustments on the predicted response rates in other cells. The second member of the class of alternative adjustments based on logistic regression uses this mixed strategy. In cells containing 25 or more sample persons, the nonresponse adjustment is the inverse of the observed cell response rate. In cells containing less than 25 sample persons, the nonresponse adjustment is the inverse of the predicted response rate for the cell. This adjustment is called the *mixed logistic adjustment*.

A third logistic nonresponse adjustment studied is similar to the current SIPP procedures. Initial cells were defined by the crossclassification of the ten independent variables used in the logistic regression. The cells were then collapsed until the sample size in each cell exceeded 30, and the inverse of the observed response rate within a collapsed cell was then used as the nonresponse adjustment. The strategy for collapsing cells was to group together cells with similar predicted response rates. This nonresponse adjustment is called the *collapsed logistic adjustment*. Although this adjustment is similar to the current SIPP panel nonresponse adjustment, there are some differences in the variables used to define the cells and the methods used to combine small cells are different.

For all three alternative weighting adjustments based on the logistic regression model, the observed and predicted response rates were computed from weighted counts of the number of cases rather than using the unweighted numbers, where the weights were the nonresponse adjusted Wave 1 weights. In practice, the weighted and unweighted adjustments were nearly the same.

3.1.1 Adjustments Based on CHAID Models

The second class of methods for adjusting for panel nonresponse involved using the CHAID categorical search algorithm to divide the data set into adjustment cells. The general approach was to define adjustment cells as combinations of responses to the predictor variables that had the greatest discrimination with respect to panel response rates, subject to the restriction that each cell should have a minimum sample size of at least 25 persons. The panel nonresponse adjustment was the inverse of the observed response rate in the cell.

The CHAID algorithm creates cells by splitting the data set progressively in a tree structure. The splitting along each newly created branch is performed by choosing the variable that maximizes a χ^2 criterion. When the split involves a polychotomous variable, the split may involve several branches. The χ^2 tests are modified using Bonferroni type adjustments to prevent variables from being chosen simply because they have more categories. CHAID is one version of the Automatic Interaction Detector (AID) developed for categorical variables. Kass (1980) presents the theory underlying the CHAID technique. Another version of the same methodology was used by Lepkowski, Kalton and Kasprzyk (1989) and Kalton, Lepkowski and Lin (1985) to model nonresponse in SIPP.

For the current analysis, two CHAID models were examined by including different sets of predictor variables. The first model included the seven most important predictors in the logistic regression model (age, relationship to reference person, race of householder, tenure, Census region, imputation flags, and bond-holding status), plus gender. This model resulted in 99 nonresponse adjustment cells. The nonresponse adjustment based on this model is called CHAID 1. The second CHAID model included the 13 predictor variables from the logistic regression model presented in Table 1. This model resulted in 142 nonresponse adjustment cells. The nonresponse adjustment for this model is called CHAID 2.

3.1.2 Adjustments Based on Generalized Raking

The third class of methods examined for adjusting for panel nonresponse was generalized raking. Unlike the other approaches, nonresponse adjustment cells were not developed by crossclassifying the predictor variables. Rather, raking was directly applied to force the panel respondents' marginal distributions for each of the predictor variables (computed using the adjusted weights) to equal the corresponding distributions for respondents and nonrespondents combined (computed using the original Wave 1 weights). Kalton and Kasprzyk (1986) refer to this method as sample based raking. The ten predictor variables from the reduced logistic regression model were used to define the marginal distributions. Hence, the raking problem was ten dimensional, with one dimension for each predictor variable.

Raking involves modifying the original weights in order to satisfy certain marginal constraints while minimizing the distance between the original and adjusted weights. Deville and Särndal (1992) describe some distance functions that may be used and derive the corresponding raking methodologies. The raking algorithm of Deming and Stephan (1942), which implicitly employs a distance function that leads to a multiplicative solution, is one form of generalized raking.

The CALMAR software described by Deville, Särndal and Sautory (1993) was used to compute the adjustments. Three different distance functions were examined: the multiplicative method, the linear method, and the truncated multiplicative method. The adjustments for all three distance functions were found to be nearly identical. This empirical result is consistent with results given by Deville and Särndal (1992) that show that the estimators using weights generated with different distance functions are asymptotically equivalent if the distance functions satisfy certain smoothness conditions. The three distance functions employed in this research satisfy those conditions. Since the adjustments were nearly identical for all three methods, only the weighting adjustment from the multiplicative method was retained for further evaluation. The resulting adjustment is called the *raking* adjustment.

3.1.3 Distributions of Nonresponse Adjustments

The adjustments for each of the six schemes described above were computed for the 1987 SIPP panel file. Table 2 summarizes the distributions of the resulting nonresponse adjustments. The summary is for the adjustments only, not the weights that are the products of the adjustments and the Wave 1 weights. Table 2 is divided into two parts: the upper part shows the mean, median, and extreme values for each adjustment distribution, as well as $(1 + CV^2)$, where CV is the coefficient of variation for each adjustment. The statistic $(1 + CV^2)$ serves as an indicator of the increase in variance of the estimates introduced by having variable nonresponse adjustment factors (see Kish 1992). The second part of Table 2 shows the correlations among the alternative forms of adjustment.

Since the overall weighted panel response rate is 0.794, the mean overall nonresponse adjustment would be $1/(0.794) = 1.26$ if the same adjustment were used for all persons. The mean weighting adjustments for the three weighting adjustments that use the inverses of cell response rates (collapsed logistic, CHAID 1 and CHAID 2) are necessarily equal to the overall nonresponse adjustment of 1.26. The mean weighting adjustments for the other schemes differ only minimally from the mean overall nonresponse adjustment.

For all six schemes, the distributions are positively skewed, with a few cases with large weights. By their nature, the various logistic and CHAID schemes cannot have adjustments less than 1.00, whereas the raking algorithm can, and does, do so. The median weights are similar among all schemes, but the maximum weights are not. The CHAID 2 scheme has a cell with a response rate of only 7 percent, leading to the largest maximum weight of 13.93. The raking scheme has the smallest maximum weight of 2.51.

Table 2

Distribution of Panel Nonresponse Adjustments

	Mean	Minimum	Median	Maximum	$1 + CV^2$
Predicted logistic	1.26	1.04	1.20	4.28	1.02
Mixed logistic	1.26	1.00	1.20	4.28	1.03
Collapsed logistic	1.26	1.00	1.20	3.43	1.02
CHAID 1	1.26	1.02	1.22	3.49	1.03
CHAID 2	1.26	1.01	1.19	13.93	1.04
Raking	1.26	0.91	1.23	2.51	1.02

Correlations

	Predicted Logistic	Mixed Logistic	Collapsed Logistic	CHAID 1	CHAID 2	Raking
Predicted logistic	1.00	0.96	0.73	0.73	0.63	0.95
Mixed logistic		1.00	0.73	0.72	0.63	0.90
Collapsed logistic			1.00	0.69	0.58	0.75
CHAID 1				1.00	0.81	0.73
CHAID 2					1.00	0.63
Raking						1.00

The values of $(1 + CV^2)$ are fairly consistent across the various adjustments. The CHAID 2 adjustment has the greatest value of $(1 + CV^2)$, primarily because of the presence of more outlying adjustments (such as the maximum value of 13.93). However, even for this method, the approximate increase in the variance of the survey estimates is only four percent. The raking adjustment has the smallest increase in variance (two percent), but this increase is not very different from that of the other methods.

The pairwise correlations between the six alternative sets of weights range from 0.58 to 0.96. Not surprisingly, the predicted logistic and mixed logistic weights are highly correlated. Given the similarity of the predicted main-effects logistic regression scheme to raking, it is also not surprising that their two sets of weights are highly correlated. The relatively high correlation between the raking weights and the CHAID 1 weight and the collapsed logistic weight is consistent with the earlier result showing no large interaction terms. The CHAID 2 weights have the lowest correlations with the other sets of weights, except for their correlation with the CHAID 1 weights. This finding is probably explained by the wide variability in the CHAID 2 weights resulting from the use of as many as 142 adjustment cells.

3.2 Final Panel Weights

The panel nonresponse adjustment weights discussed in the previous section represent the adjustments to the Wave 1 weights to compensate for panel nonresponse. The final panel weights that may be used in the analysis of the SIPP panel file are obtained by multiplying the panel nonresponse adjustment weights by the Wave 1 weights, and then applying poststratification to make weighted sample totals conform to totals derived primarily from the Current Population Survey (CPS). This procedure was applied for each of the six alternative panel nonresponse adjustment schemes.

The poststratification procedure used was equivalent to the current SIPP procedure, except that the latter procedure poststratifies by rotation groups whereas for the alternative weighting schemes the poststratification was performed on all rotation groups combined. The difference should not have an appreciable effect. After poststratification, the six alternative sets of final weights and the SIPP panel weights sum to the same control totals.

To compare the final panel weights for the six adjustment schemes with one another and with the current SIPP panel weight, the correlations between the weights were computed, along with the measure of variability used previously, $(1 + CV^2)$. The results are presented in Table 3. The estimates of the variability due to the weighting $(1 + CV^2)$ indicate similar increases of between 8 and 10 percent in the variances of survey estimates for all of the weighting schemes. The correlations between the alternative sets of final panel weights are all 0.85 or higher. Comparing these correlations to those in Table 2, it is clear that the correlations between the final weights are appreciably higher than those between the panel nonresponse adjustment weights. The correlations between the SIPP panel weight and the alternative final weights are consistently lower than any others, probably because the variables used in forming the nonresponse adjustments for this weight differed from those used for the alternative weights. The variables used in the alternative schemes that are not used in the SIPP panel weight are age, relationship to reference person, number of imputed items, class of work, and food stamp recipiency. Household size is the only variable other than MSA status (which was not available due to disclosure concerns) used in the SIPP panel weight but not used for the alternative schemes because it was not found to be significantly associated with response rates.

Table 3

Correlations Between Poststratified Weights with Variance Inflation Measures

	SIPP panel	Predicted Logistic	Mixed Logistic	Collapsed Logistic	CHAID 1	CHAID 2	Raking
SIPP panel	1.00	0.75	0.74	0.75	0.71	0.68	0.77
Predicted logistic		1.00	0.99	0.91	0.90	0.86	0.98
Mixed logistic			1.00	0.91	0.90	0.86	0.97
Collapsed logistic				1.00	0.89	0.85	0.93
CHAID 1					1.00	0.94	0.91
CHAID 2						1.00	0.87
Raking							1.00
$1 + CV^2$	1.08	1.09	1.09	1.08	1.09	1.10	1.08

4. COMPARING ESTIMATES USING ALTERNATIVE WEIGHTS

The previous section described the development of the alternative sets of final weights that may be used for the analysis of the SIPP panel file. All the final weighting schemes incorporate adjustments for unequal selection probabilities, nonresponse at the initial wave, panel nonresponse, and poststratification to external control totals. This section compares survey estimates obtained using the alternative weighting schemes with one another and with the corresponding estimates obtained using the SIPP panel weights. In addition, where possible, the various survey estimates are also compared with external estimates from other sources. Some of the external estimates are benchmark estimates obtained from administrative records or the Current Population Survey. Other external estimates are obtained from Wave 1 of the 1989 SIPP panel. Data collected in Wave 7 of the 1987 SIPP panel relate to the same time period as data collected in Wave 1 of the 1989 SIPP panel, and hence estimates obtained from these two data sources should be comparable.

In making comparisons with benchmark estimates, it needs to be recognized any differences observed may be explained by a variety of factors of which panel nonresponse is only one. For example, response errors and differences in definitions may explain differences between SIPP estimates and benchmark estimates. Thus the benchmark comparisons need to be treated with caution. Since the 1989 SIPP panel estimates are based on Wave 1 data, they are not subject to the panel nonresponse. Thus, differences between estimates obtained from the 1987 and 1989 SIPP panels are perhaps the most likely to be caused by a failure of the panel nonresponse adjustments to fully compensate for panel nonresponse bias. However, even in this case, alternative explanations such as panel conditioning could contribute to the differences (although Pennell and Lepkowski 1992, show that panel conditioning is not a major factor in most SIPP estimates).

Table 4 presents a variety of estimates from the 1987 SIPP panel file using the SIPP panel weight and the six alternative weighting schemes, and corresponding benchmark estimates and estimates from the 1989 SIPP panel where available. The estimates are percentages, except for the estimates of the mean number of months without health insurance, median household income, and annual wages. The estimates are for the total population, except for the employment estimates (percent employed, unemployed and out of the labor force), which are for persons over the age of 15, and for annual wages, which are for persons over the age of 14. The estimates are for three different time periods: June 1987, January 1989, and the calendar year of 1987. For example, the first three estimates in Table 4 are the estimated percentages of persons participating in the AFDC (Aid for Families with Dependent Children) program in June 1987, in January 1989, and at any time during the 1987 calendar year. A comparable estimate from the 1989 SIPP panel is available only for the January 1989 time period.

The most notable finding from Table 4 is the similarity of the estimates computed with all the weighting schemes from the 1987 panel. The percentage estimates in Table 4 are in fact given to two decimal places because the use of the conventional one decimal place would often show no difference between the alternative estimates. The largest difference occurs for the percentage employed in January 1989, where the estimate using the SIPP panel weight is 62.7 percent and the estimate using the mixed logistic regression weight is 62.3 percent. Even this largest of differences is relatively small, especially when considering that the estimated standard error for this estimate is 0.3 percent.

When the 1987 SIPP panel estimates are compared with the external estimates from the 1989 SIPP panel and from other sources, some of the differences are much larger and of substantive importance. To examine these differences in more detail, standardized differences between the alternative estimates and the benchmark estimates were computed and are shown in Table 5. A standardized difference is defined as the difference between the alternative estimate and the external estimate divided by the standard error of the difference.

The upper part of Table 5 shows the standardized differences when the 1989 SIPP panel is used to produce the external estimate. The standardized differences for most of the estimates are less than 2.0 in absolute value, indicating that the differences may be accounted for by sampling error. However, the standardized differences for the percentage unemployed and for the poverty rate are greater than 2.0 and highly significant. Thus, the alternative weighting adjustments do not succeed in bringing the 1987 survey estimates in line with the 1989 survey estimates for all characteristics.

The lower part of Table 5 shows the standardized differences when other benchmark estimates are used. These standardized differences are generally large and in many cases very large. Only a few are less than 2.0 and many are greater than 10.0. Given the much smaller standardized differences found in the upper part of Table 5 for similar statistics, it seems likely that factors other than panel nonresponse bias are largely responsible for the magnitude of these differences. The standardized differences based on these largely administrative data sources may signal important issues related to the quality of the data (from either the SIPP, the benchmark data source, or both), but they do not provide much help in assessing the effectiveness of alternative nonresponse adjustments in reducing panel nonresponse bias.

Table 4
Estimates for the Total Population from the 1987 SIPP Panel with Alternative Weighting Schemes and Estimates from Other Sources

	SIPP Panel	Predicted Logistic	Mixed Logistic	Collapsed Logistic	CHAID 1	CHAID 2	Raking	1989 SIPP Panel	Bench-mark
AFDC - June 1987	3.73	3.70	3.74	3.72	3.71	3.60	3.69		4.28[1]
AFDC - January 1989	3.10	3.12	3.14	3.12	3.14	3.02	3.10	3.56	4.24[2]
AFDC - Annual 1987	4.85	4.78	4.82	4.81	4.80	4.69	4.78		
Food stamps - June 1987	7.43	7.26	7.30	7.34	7.38	7.20	7.21		7.35[3]
Food stamps - January 1989	6.71	6.63	6.67	6.64	6.70	6.59	6.58	6.30	7.29[4]
Food stamps - Annual 1987	10.30	10.11	10.16	10.18	10.24	10.05	10.06		
Medicaid - January 1989	6.77	6.78	6.81	6.75	6.81	6.68	6.76	6.97	
Medicaid - Annual 1987	9.21	9.21	9.24	9.21	9.25	9.09	9.21		
SSI - June 1987	1.68	1.70	1.69	1.67	1.69	1.65	1.69		1.68[3]
SSI - January 1989	1.65	1.67	1.66	1.64	1.66	1.61	1.66	1.65	1.74[3]
SSI - Annual 1987	1.80	1.82	1.82	1.80	1.82	1.78	1.82		
Social security - January 1989	14.92	14.87	14.87	14.89	14.88	14.89	14.85	15.14	
Poverty rate - June 1987	10.88	10.75	10.79	10.76	10.79	10.69	10.74		
Poverty rate - January 1989	12.91	12.98	13.02	12.97	12.99	12.91	12.93	14.46	
Entering poverty 1987/1988	2.25	2.31	2.32	2.30	2.29	2.32	2.31		
Leaving poverty 1987/1988	2.69	2.63	2.64	2.60	2.62	2.63	2.63		
Mean months without health insurance - 1987	1.66	1.69	1.70	1.67	1.67	1.69	1.69		
Median household income - January 1989	2,601	2,600	2,597	2,607	2,607	2,607	2,602	2,550	
Annual wages 1987 (in trillions)	1.93	1.94	1.93	1.94	1.94	1.94	1.94		2.22[4]
Employed - January 1989	62.74	62.36	62.34	62.43	62.42	62.52	62.42	61.60	
Unemployed - January 1989	3.57	3.64	3.63	3.60	3.58	3.60	3.63	4.52	
Out of labor force - January 1989	33.69	34.01	34.03	33.96	34.01	33.88	33.95	33.88	
Married in 1987	1.39	1.41	1.40	1.39	1.39	1.39	1.41		1.86[5]
Divorced in 1987	0.51	0.50	0.50	0.49	0.50	0.51	0.49		0.90[6]
Changed address in 1987	12.88	13.32	13.32	13.19	13.36	13.37	13.33		17.99[6]

[1] Social Security Bulletin, Volume 52, No. 3.
[2] Social Security Bulletin, Volume 51, No. 7.
[3] USDA Food and Nutrition Service, unpublished data.
[4] U.S. Bureau of the Census, Current Population Reports, Consumer Income, P-60, No. 174.
[5] National Center for Health Statistics: Vital Statistics of the U.S., 1987, Volume III, Marriage and Divorce, DHHS Pub. No. (PHS) 91-1103.
[6] U.S. Bureau of the Census, Current Population Reports, Population Characteristics, P-20, No. 473.

Table 5

Standardized Differences Between 1987 SIPP Panel Estimates and Benchmark Estimates

	Bench-mark Estimate	SIPP Panel	Predicted Logistic	Mixed Logistic	Collapsed Logistic	CHAID 1	CHAID 2	Raking
1989 SIPP panel estimates								
AFDC	3.56	-1.58	-1.52	-1.43	-1.52	-1.44	-1.84	-1.57
Food stamps	6.30	1.02	0.82	0.92	0.86	1.01	0.73	0.69
Medicaid	6.97	-0.50	-0.47	-0.40	-0.53	-0.39	-0.70	-0.51
SSI	1.65	0.05	0.11	0.08	-0.03	0.07	-0.15	0.09
Social Security	15.14	-0.38	-0.46	-0.46	-0.42	-0.44	-0.42	-0.50
Poverty rate	14.46	-2.77	-2.64	-2.57	-2.67	-2.63	-2.78	-2.74
Median Income	2,550	2.05	2.01	1.89	2.30	2.30	2.29	2.09
Employed	61.60	2.42	1.60	1.56	1.76	1.72	1.95	1.73
Unemployed	4.52	-4.93	-4.59	-4.59	-4.76	-4.90	-4.78	-4.60
Out of labor force	33.88	-0.42	0.28	0.32	0.18	0.28	-0.01	0.15
Other benchmark estimates								
AFDC - June 1987	4.28	-2.55	-2.66	-2.49	-2.59	-2.65	-3.14	-2.71
AFDC - January 1989	4.24	-5.71	-5.62	-5.49	-5.63	-5.51	-6.10	-5.70
Food stamps - June 1987	7.35	0.27	-0.31	-0.16	-0.04	0.11	-0.50	-0.48
Food stamps - January 1989	7.29	-2.04	-2.32	-2.17	-2.26	-2.06	-2.44	-2.50
SSI - June 1987	1.68	0.00	0.13	0.08	-0.03	0.08	-0.20	0.11
SSI - January 1989	1.74	-0.57	-0.48	-0.53	-0.67	-0.54	-0.84	-0.50
Annual wages 1987	2.22	-16.12	-15.94	-16.38	-15.66	-15.61	-15.60	-15.78
Married in 1987	1.86	-5.11	-4.93	-4.98	-5.11	-5.10	-5.07	-4.95
Divorced in 1987	0.90	-7.15	-7.37	-7.36	-7.40	-7.32	-7.20	-7.40
Changed address in 1987	17.99	-11.49	-10.50	-10.51	-10.80	-10.42	-10.40	-10.49

5. DISCUSSION

Nonresponse weights are widely used to compensate for unit nonresponse in sample surveys. The basic requirement for this form of weighting is the availability of information on one or more auxiliary variables for both respondents and nonrespondents. In many surveys, this information is available for only a small number of auxiliary variables (such as the PSUs and strata from which the units were selected). In such surveys, the nonresponse weights can often be simply developed as weighting class adjustments for a set of classes based on the crosstabulation of the auxiliary variables.

There are, however, surveys in which data are available for a large number of auxiliary variables for possible use in developing nonresponse weights. This situation often applies when an administrative record system is used as the survey's sampling frame, with all the information in the system then being available for use in making nonresponse adjustments. It also applies when the survey data collection is conducted in two or more phases (*e.g.*, an initial screening interview followed by a detailed interview or some other form of data collection at a later time point) and when nonresponse adjustments are needed for later phases; in this case, data from prior phases of data collection may be used in compensating for nonresponse at later phases. A similar situation applies in panel surveys when adjustments are required for nonresponse at later waves of the panel, as discussed in this paper.

When a large number of auxiliary variables is available for all sampled units, two main choices need to be made. First, there is the choice of auxiliary variables to use in the adjustment. Second, there is the choice of the adjustment method to be applied.

The basic approach adopted in this study for choosing the auxiliary variables for use in the nonresponse adjustment was to identify the set of variables that were good predictors of panel nonresponse. With so many auxiliary variables available, the first step was a screening procedure to eliminate variables that were found to have little association with the panel nonresponse rate. Then, logistic regression models using predictor variables remaining from the screening were examined to identify the set of variables to be retained for use in adjusting the weights. Whether the number of auxiliary variables is reduced to a manageable set by this or some other approach (*e.g.*, by using the CHAID algorithm), this reduction is likely to be a necessary first step when there are many potential auxiliary variables available.

After selecting the subset of auxiliary variables, a wide variety of methods exists for creating the nonresponse adjustments. We examined panel nonresponse adjustments based on logistic regression models, categorical search models, and sample-based generalized raking. The final panel weights resulting from these adjustment schemes were highly correlated with one another and they yielded estimates that were very similar. None of the schemes produced estimates that were superior in terms of bias reduction.

In part, the high correlation of the final panel weights generated by the different adjustment schemes may be explained by the similarity of many of the adjustment schemes. In part, it may be explained by the final post-stratification weighting which raised the correlations between the weights. It may also be partly explained by the lack of large interaction effects between the auxiliary variables. If there were sizable interaction effects that were not included in the logistic modeling, then one might expect greater differences between the raking and predicted logistic weights on the one hand and the CHAID, mixed logistic, and collapsed logistic weights on the other hand. Thus, the similarity in weights produced by the alternative weighting schemes for the SIPP may not be as great in other circumstances.

A common concern that arises when many auxiliary variables are used to adjust the weights is that the adjusted weights might be highly variable, thus causing a serious loss of precision in the survey estimates. This proved not to be the case in the methods we evaluated. The variability of the weights with all the weighting schemes turned out to be similar, provided reasonable precautions were taken in creating the adjustments.

Although the empirical results do not show any appreciable differences in the estimates produced using the alternative weighting schemes and those produced using the SIPP panel weights, the correlations of the alternative adjusted weights and the current SIPP panel weight were found to be lower than the correlations among the alternative weights. This finding suggests that the choice of auxiliary variables is an important one, and probably more important than the choice of the weighting methodology. Although the more systematic methods used in this research for choosing the auxiliary variables did not result in major improvements over the current SIPP procedures, an analytic based choice of auxiliary variables may be more productive in other studies.

When a sizable number of auxiliary variables that are correlated to response propensity is available, it seems wise to use as many of them as possible in the nonresponse adjustment to serve as a safeguard in attempting to compensate for nonresponse bias. This general strategy should, however, be tempered by a careful assessment of the variation of the resulting weights in order to avoid too great a loss of precision in the survey estimates. In addition, a practical consideration that should be taken into account is the ease of implementation of the weighting methodology. If, as in this study, alternative weighting methodologies yield very similar weights and estimates, a method that is simple to apply may be preferable.

ACKNOWLEDGMENTS

We thank the referees for helpful comments on the earlier version of the paper. The paper reports research undertaken for the U.S. Bureau of the Census. The views expressed are the authors'. They do not necessarily reflect the views of the Bureau of the Census.

REFERENCES

CHAPMAN, D.W., BAILEY, L., and KASPRZYK, D. (1986). Nonresponse adjustment procedures at the U.S. Bureau of the Census. *Survey Methodology*, 12, 161-180.

DEMING, W.E., and STEPHAN, F.F. (1942). On a least squares adjustment of a sampled frequency table when the expected marginal totals are known. *Annals of Mathematical Statistics*, 11, 427-444.

DEVILLE, J.-C., and SÄRNDAL, C.-E. (1992). Calibration estimators in survey sampling. *Journal of the American Statistical Association*, 87, 376-382.

DEVILLE, J.-C., SÄRNDAL, C.-E., and SAUTORY, O. (1993). Generalized raking procedures in survey sampling. *Journal of the American Statistical Association*, 88, 1013-1020.

JABINE, T.B., KING, K.E., and PETRONI, R.J. (1990). *Survey of Income and Program Participation (SIPP): Quality Profile*. Washington, DC: U.S. Bureau of the Census.

KALTON, G. (1983). *Compensating for Missing Survey Data*. Ann Arbor, MI: Survey Research Center, University of Michigan.

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

KALTON, G., LEPKOWSKI, J.M., MONTANARI, G.E., and MALIGALIG, D. (1990). Characteristics of second wave nonrespondents in a panel survey. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 462-467.

KALTON, G., LEPKOWSKI, J., and LIN, T. (1985). Compensating for wave nonresponse in the 1979 ISDP Research Panel. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 372-377.

KASS, G.V. (1980). An exploratory technique for investigating large quantities of categorical data, *Applied Statistics*, 29, 119-127.

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

LEPKOWSKI, J., KALTON, G., and KASPRZYK, D. (1989). Weighting adjustments for partial nonresponse in the 1984 SIPP Panel. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 296-301.

LITTLE, R.J.A. (1986). Survey nonresponse adjustments for estimates of means. *International Statistical Review*, 54, 2, 137-139.

NELSON, D., McMILLEN, D., and KASPRZYK, D. (1985). *An Overview of the SIPP, Update 1*. SIPP Working Paper No. 8401. Washington, DC: U.S. Bureau of the Census.

OH, H.L., and SCHEUREN, F. (1983). Weighting adjustments for unit nonresponse. In *Incomplete Data in Sample Surveys, Volume 2: Theory and Bibliographies* (Eds. W.G. Madow, I. Olkin, and D. Rubin), 143-184. New York: Academic Press.

PENNELL, S.G., and LEPKOWSKI, J.M. (1992). Panel conditioning effects in the Survey of Income and Program Participation. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 556-571.

RIZZO, L., KALTON, G., and BRICK, M. (1994). Weighting adjustments for panel nonresponse in the SIPP. Final Report submitted to U.S. Bureau of the Census for SIPP Panel Nonresponse Project.

Survey Methodology, June 1996
Vol. 22, No. 1, pp. 55–64
Statistics Canada

Multiple Sample Estimation of Population and Census Undercount in the Presence of Matching Errors

YE DING and STEPHEN E. FIENBERG[1]

ABSTRACT

The multiple capture-recapture census is reconsidered by relaxing the traditional perfect matching assumption. We propose matching error models to characterize error-prone matching mechanisms. The observed data take the form of an incomplete 2^k contingency table with one missing cell and follow a multinomial distribution. We develop a procedure for the estimation of the population size. Our approach applies to both standard log-linear models for contingency tables and log-linear models for heterogeneity of catchability. We illustrate the method and estimation using a 1988 dress rehearsal study for the 1990 census conducted by the U.S. Bureau of the Census.

KEY WORDS: Capture-recapture census; Estimates for total population size; Log-linear models; Matching errors; Multiple recapture census.

1. INTRODUCTION

The multiple recapture census technique has been used in many fields to estimate the size of a closed population. Cormack (1968) and Seber (1982) give excellent reviews of many techniques used. Here we consider a sequence of samples, $s_1, \ldots, s_k$, where the members of i-th sample are uniquely labeled, for example, by tagging or marking, and then returned to the population (Darroch 1958). Usual multiple recapture census methods make the following assumptions.

(1) **Perfect matching.** Individuals in one list (information source, sample) can be matched with those in another list without error. In other words, there are no misclassification errors with respect to determining whether a particular individual has been recorded by both information sources or only one of them.
(2) **Independence.** The lists are independent of one another, that is, the probability of an individual being included in one list does not depend on whether the individual was included in previous lists.
(3) **Homogeneity (Equal Catchability).** All individuals in the population under study have equal probabilities of being observed (captured) in any list (sample).
(4) **Closure.** The population in question is "closed", so that there are no changes due to birth, death, emigration, or immigration during the period when the sampling takes place.

Darroch (1958) examined the multiple recapture census under these four assumptions. Fienberg (1972) adopted a log-linear model approach to allow for statistical dependence of specific types among samples, thereby dropping the independence assumption. Darroch, Fienberg, Glonek and Junker (1993) developed an extended log-linear model approach that allows for individual-level heterogeneity as well as dependence, but it requires at least three samples, *i.e.*, $k = 3$. In the context of the two-sample census approach used by U.S. Bureau of Census for census coverage evaluation, matching problems due to unavoidable mismatches and erroneous nonmatches have been explored by several authors. For example, Ding and Fienberg (1994) considered modeling matching errors in the two-sample census and developed systematic procedure for the estimation of population totals. The inclusion of a third sample, *e.g.*, drawn from the administrative records, in modeling and estimation of census coverage has been considered by the U.S. Bureau of Census in the past and remains an option to augment and evaluate the dual system approach. In this paper, we consider matching error models for the multiple sample census problem, allowing for both dependence and heterogeneity.

Here we view the observations from a multiple recapture census data as falling into a 2^k cross-classification, with absence or presence on the i-th sample defining the category for the i-th dimension. In this cross-classification, the cell corresponding to absence for all k samples is missing. The objective is to estimate the number of individuals in the population who are not observed, which corresponds to the missing cell in the 2^k incomplete contingency table. In Section 2, we investigate the effects of matching errors on the observed 2^k incomplete table. In Section 3, some models for matching errors are proposed to characterize an error-prone matching process. Based on these models and assumptions (3) and (4), we develop a procedure using log-linear model formulation for the estimation of the population size. In Section 5, we use the proposed methods to analyze data from 1988 Dress Rehearsal Census conducted by the U.S. Bureau of Census.

[1] Ye Ding, Research Scientist, Bureau of Biometrics, New York State Health Department, Concourse, Room C-144, Empire State Plaza, Albany, New York 12237, U.S.A.; Stephen E. Fienberg, Maurice Falk Professor of Statistics and Social Science, Department of Statistics, Carnegie Mellon University, Pittsburgh, Pennsylvania 15213, U.S.A.

2. MATCHING ERRORS IN MULTIPLE SAMPLE CENSUS

We begin by classifying matching errors into two broad categories, mismatches and erroneous nonmatches. To understand the nature of matching errors in multiple-sample census, we review the case of a three-sample census. Suppose that there are no missing data or errors in recording the information for any individual in the population and one takes three samples from the population, s_1, s_2, and s_3. For instance, suppose that, in sample s_1, individuals 1, 3, 4 and 7 are seen, individuals 3, 4, and 8 are seen in s_2, and individuals 4, 9, and 10 in s_3. In vector notation, we can represent this as $s_1 = (1, 3, 4, 7)$, $s_2 = (3, 4, 8)$ and $s_3 = (4, 9, 10)$. Matching errors are not present provided that there is complete and correct information available. We thus have the following incomplete 2^3 table corresponding to these three samples:

Table 1

Original Table without Matching Errors

	s_1			
	Present		Absent	
	s_2		s_2	
s_3	Present	Absent	Present	Absent
Present	1	0	0	2
Absent	1	2	1	-

Suppose further that, because of missing data or incorrect information, we actually observe

$$s_1 = (1, 3, 4, 7), \quad s_2 = (3^*, 4^*, 8), \quad s_3 = (4, 9, 10),$$

where 3* and 4* are individuals 3 and 4 but with incorrect information leading to two erroneous nonmatches when the samples are matched. Assuming no erroneous matches, we then observe the incomplete 2^3 table:

Table 2

Observed Table with Matching Errors

	s_1			
	Present		Absent	
	s_2		s_2	
s_3	Present	Absent	Present	Absent
Present	0	1	0	2
Absent	0	3	3	-

The effects of matching errors are obvious from a comparison of Table 1 and 2:

(i) The number of observations may increase for some cells while decreasing for the others, and as a consequence, the marginal totals and especially the total number of different individuals observed in the three samples may change, subject to the constraint that the total number of observations in each sample, x_{1++}, x_{+1+}, and x_{++1}, remain the same. Changes in the total number of different individuals in all samples make our problem distinct from the usual misclassification problem in the analysis of categorical data, in which the possibility of making mistakes in classifying individuals into respective categories is considered. (*e.g.*, see Chen 1979).

(ii) In parallel, there may be changes in some cell probabilities subject to the constraint that the probability of being captured in a sample, p_{1++}, p_{+1+}, and p_{1++}, is unchanged.

Because of the complexity of matching errors in the three-sample case, we need some special terminology for descriptive convenience. We say that an individual is at state 1 with respect to sample s_1 if the individual is observed in s_1 and at state 0 if not. We use a triple (i,j,k), $0 \leq i, j, k \leq 1$, to denote an individual at state i, j, and k with respect to s_1, s_2 and s_3, respectively. For instance, (1,0,0) is an individual observed only in s_1, and (1,1,1) is an individual captured in three samples. We define the level of an individual (i,j,k) as $i + j + k$, *i.e.*, the number of samples in which the individual is included. There are four different levels, 0, 1, 2 and 3. An individual has level 0 if and only if he/she is not captured by any sample, and has level 3 if he/she is in three samples. For a (1,1,0) individual, if the correct match is not made according to the matching rule, this individual decomposes into "two different" individuals, a (1,0,0) and a (0,1,0), assuming no erroneous matches. On the other hand, a (1,0,0) individual matched incorrectly with a (0,1,0) will produce a single observed (1,1,0) individual. For convenience, we call such a decomposition or combination a *transition*. Then transitions can only go from level 3 or 2 to the same (if there is no matching error) or lower levels in the absence of erroneous matches. More specifically, a (1,1,1) person may make a transition into one of 5 possible sets of individuals

$$\{(1,1,1)\}, \quad \{(1,0,0), (0,1,1)\}, \quad \{(0,1,0), (1,0,1)\}$$
$$\{(0,0,1), (1,1,0)\}, \quad \{(1,0,0), (0,1,0), (0,0,1)\}.$$

For level 2 individuals, (1,1,0) can decompose into $\{(1,0,0),(0,1,0)\}$ or stay at $\{(1,1,0)\}$, and similarly for $\{(0,1,1)\}$ and $\{(1,0,1)\}$. From above discussions, we summarize the effect of matching errors by the following diagram:

$$\boxed{\text{Table 1}} \rightarrow \{\text{Matching Process}\} \rightarrow \boxed{\text{Table 2}}$$

where Table 1 is the original 2^k incomplete table with no matching errors and Table 2 is the observed 2^k incomplete table in the presence of matching errors. Henceforth, we denote the cell probabilities and expected cell counts associated with Table 1 by $\{r_{ijk}\}$ and $\{l_{ijk}\}$ and those of Table 2 by $\{p_{ijk}\}$, $\{m_{ijk}\}$, for $1 \le i,j,k \le 2$.

3. SOME MODELS FOR MATCHING ERRORS

We now propose models to describe the matching errors, each of which allows us to formulate the reallocation of cell probabilities and expected cell counts associated with Table 1.

Model (1). In addition to the homogeneity and closure assumptions in §1, we assume that: (i) There are no erroneous matches in the matching process; (ii) Any individual will stay at his original state with probability θ, and transition to any of a possible set of individuals with probability $(1 - \theta)/(m - 1)$, where m is the number of all possible sets of individuals to which the individual may transition. For example, for a (1,1,1) person discussed late in last section, $m = 5$.

Under this model, for the three-sample census, we can express the probabilities for the table with matching errors, $\{p_{ijk}\}$, in terms of probabilities of the table with no matching errors, $\{r_{ijk}\}$:

$$p_{111} = \theta r_{111},$$

$$p_{112} = \frac{1-\theta}{4} r_{111} + \theta r_{112},$$

$$p_{121} = \frac{1-\theta}{4} r_{111} + \theta r_{121},$$

$$p_{211} = \frac{1-\theta}{4} r_{111} + \theta r_{211},$$

$$p_{122} = \frac{1-\theta}{2} r_{111} + (1-\theta) r_{112} + (1-\theta) r_{121} + r_{122},$$

$$p_{212} = \frac{1-\theta}{2} r_{111} + (1-\theta) r_{112} + (1-\theta) r_{211} + r_{212},$$

$$p_{221} = \frac{1-\theta}{2} r_{111} + (1-\theta) r_{211} + (1-\theta) r_{121} + r_{221}.$$

Let

$$\vec{p} = (p_{111}, p_{112}, p_{121}, p_{211}, p_{122}, p_{212}, p_{221})^T,$$

and

$$\vec{r} = (r_{111}, r_{112}, r_{121}, r_{211}, r_{122}, r_{212}, r_{221})^T,$$

then

$$\vec{p} = M_1 \times \vec{r}. \tag{1}$$

Here M_1 is a 7 by 7 matrix determined by the above seven equations derived under Model (1).It is straightforward to verify that the probability of catching any individual in each sample is fixed, *i.e.*, $p_{1++} = r_{1++} = p_1$, $p_{+1+} = r_{+1+} = p_2, p_{++1} = r_{++1} = p_3$. This must be the case because the sample capture probabilities do not depend on how the matching mechanism operates.

We can easily generalize this formulation to handle the k-sample case; however, the algebra involved is quite messy for large k. We can simplify this model by requiring that the transitions can go downwards by at most one level, thus yielding Model (2):

Model (2). In addition to the homogeneity and closure assumptions in §1, we assume that: (i) there are no erroneous matches in the matching process; (ii) a transitions can only go downwards by at most one level; (iii) any individual will stay at his original state with probability θ, and transition to any of a possible set of individuals with probability $(1 - \theta)/(m' - 1)$, where m' is the number of sets of individuals to which transitions are possible and allowed.

We first consider the three-sample case. A (1,1,1) individual can decompose into three individuals, *i.e.*, $(1,1,1) \mapsto \{(1,0,0), (0,1,0), (0,0,1)\}$ (we use "$\mapsto$" to denote for decomposition), if three presumed matches are not made. Assumption (ii) of Model (2) assumes that this triple error has negligible probability when compared with the transition in which only one of the matches is not made so that $(1,1,1) \mapsto \{(1,1,0),(0,0,1)\}$, or $(1,1,1) \mapsto \{(1,0,1),(0,1,0)\}$, or $(1,1,1) \mapsto \{(1,1,0),(0,0,1)\}$.

For three sample case, the parametric model for expressing $\{p_{ijk}\}$ in terms of $\{r_{ijk}\}$ is:

$$p_{111} = \theta r_{111},$$

$$p_{112} = \frac{1-\theta}{3} r_{111} + \theta r_{112},$$

$$p_{121} = \frac{1-\theta}{3} r_{111} + \theta r_{121},$$

$$p_{211} = \frac{1-\theta}{3} r_{111} + \theta r_{211},$$

$$p_{122} = \frac{1-\theta}{3} r_{111} + (1-\theta) r_{112} + (1-\theta) r_{121} + r_{122},$$

$$p_{212} = \frac{1-\theta}{3} r_{111} + (1-\theta) r_{112} + (1-\theta) r_{211} + r_{212},$$

$$p_{221} = \frac{1-\theta}{3} r_{111} + (1-\theta) r_{211} + (1-\theta) r_{121} + r_{221}.$$

Then

$$\vec{p} = M_2 \times \vec{r}, \tag{2}$$

where M_2 is a 7 by 7 matrix determined by the above seven equations derived under Model (2). Again, the capture probabilities are unchanged, *i.e.*, $p_{1++} = r_{1++} = p_1$, $p_{+1+} = r_{+1+} = p_2$, $p_{++1} = r_{++1} = p_3$.

For the k-sample problem, let $p_{\bar{1}}$ be the probability of being captured in all samples, *i.e.*, $p_{\bar{1}} = p_{111\ldots1}$, and let $p_{\bar{1},\bar{2}(h_1,h_2)}$ be the cell probability corresponding to absence in the h_1-th, and h_2-th sample and presence in the others, *etc.* Under Model (2), we have $p_{\bar{1}} = \theta r_{\bar{1}}$. For $i \leq k-2$, the probability of being missed by the h_1-th, h_2-th, ..., and h_i-th sample and captured by the others is

$$p_{\bar{1},\bar{2}(h_1,h_2,\ldots,h_i)} = \theta r_{\bar{1},\bar{2}(h_1,h_2,\ldots,h_i)} + \frac{1-\theta}{k-i+1} \sum_{j=1}^{i} r_{\bar{1},\bar{2}(\{h_1,h_2,\ldots,h_i\}\backslash h_j)}.$$

For $i = k-1$, the individual is included in only one sample. For example, the probability of being captured only by the first sample is

$$p_{1,\bar{2}} = r_{1,\bar{2}} + (1-\theta) \sum_{h \neq 1} r_{1,1(h),\bar{2}} + \frac{(1-\theta)}{3} \sum_{h_1,h_2 \geq 2} r_{1,1(h_1,h_2),\bar{2}} + \sum_{j=3}^{k-1} \sum_{h_1,h_2,\ldots,h_j \geq 2} \frac{(1-\theta)}{(j+1)} r_{1,1(h_1,h_2,\ldots,h_j),\bar{2}},$$

where $r_{1,1(h_1,h_2,\ldots,h_j),\bar{2}}$ is the cell probability in the original table which corresponds to presence in the first, h_1-th, h_2-th, ..., h_j-th sample and absence in the others. By symmetry, we can write down the expression for $p_{1(h),\bar{2}}$, the probability of being observed in the h-th sample only and missed in all others.

We can refine Model (2) by assuming unequal matching rates. For example, we consider two decompositions: $(1,1,1) \mapsto \{(1,1,0),(0,0,1)\}$ and $(1,1,0) \mapsto \{(0,1,0),(1,0,0)\}$. It is common for both cases that one presumed match is not made. They differ in that one has two sources of information for that match while the other has only one. It is reasonable to assume different matching error probabilities for the two cases instead of a common one as proposed in Model (2). This leads to:

Model (3). In addition to (i) and (iii) in Model (2), we assume

$$(1,1,1) \mapsto \begin{cases} (1,1,1) & \text{with probability } \alpha_1 \\ \{(1,1,0),(0,0,1)\} & \text{with probability } (1-\alpha_1)/3 \\ \{(0,1,1),(1,0,0)\} & \text{with probability } (1-\alpha_1)/3 \\ \{(1,0,1),(0,1,0)\} & \text{with probability } (1-\alpha_1)/3 \end{cases}$$

$$(1,1,0) \mapsto \begin{cases} (1,1,0) & \text{with probability } \alpha_2 \\ \{(0,1,0),(1,0,0)\} & \text{with probability } 1-\alpha_2 \end{cases}$$

$$(1,0,1) \mapsto \begin{cases} (1,0,1) & \text{with probability } \alpha_2 \\ \{(1,0,0),(0,0,1)\} & \text{with probability } 1-\alpha_2 \end{cases}$$

$$(0,1,1) \mapsto \begin{cases} (0,1,1) & \text{with probability } \alpha_2 \\ \{(0,1,0),(0,0,1)\} & \text{with probability } 1-\alpha_2 \end{cases}$$

and (1,0,0), (0,1,0), (0,0,1) stay the same with probability one.

Under this model, we can express the cell probability $\{p_{ijk}\}$ in Table 2 in terms of α_1, α_2 and the cell probabilities of Table 1, $\{r_{ijk}\}$. To do this, we need to consider all possible transitions that produce an individual that falls into the (i,j,k) cell in Table 2. For example, we consider an observed (1,0,0) individual. This person falls into cell (1,2,2) of Table 2. Let F be the event that an observed individual has a (1,0,0) status. Let E_{ijk} be the event that an individual falls into (i,j,k) cell in Table 1. Then

$$F = \bigcup_{\{i,j,k\}} (E_{ijk} \cap F).$$

According to Model (3), there are only four possible transitions as follows that can make F happen:

$$(1,1,1) \mapsto \{(1,0,0),(0,1,1)\},$$

$$(1,1,0) \mapsto \{(1,0,0),(0,1,0)\},$$

$$(1,0,1) \mapsto \{(1,0,0),(0,0,1)\},$$

$$(1,0,0) \mapsto \{(1,0,0)\}.$$

Therefore

$$F = (E_{111} \cap F) \cup (E_{112} \cap F) \cup (E_{121} \cap F) \cup (E_{122} \cap F).$$

By the definitions of cell probabilities of the two tables, $p(F) = p_{122}$, and $p(E_{ijk}) = r_{ijk}$. By the assumptions in Model (3), $p(F \mid E_{111}) = (1 - \alpha_1)/3$, $p(F \mid E_{112}) = p(F \mid E_{121}) = \alpha_2$, and $p(F \mid E_{122}) = 1$.

Since $E_{111} \cap F$, $E_{112} \cap F$, $E_{121} \cap F$ and $E_{122} \cap F$ are four mutually exclusive possibilities that F can happen, thus

$$\begin{aligned} p_{122} &= p(E_{111} \cap F) + p(E_{112} \cap F) \\ &\quad + p(E_{121} \cap F) + p(E_{122} \cap F) \\ &= p(F \mid E_{111}) \cdot p(E_{111}) + p(F \mid E_{112}) \cdot p(E_{112}) \\ &\quad + p(F \mid E_{121}) \cdot p(E_{121}) + p(F \mid E_{122}) \cdot p(E_{122}) \\ &= \frac{1 - \alpha_1}{3} r_{111} + (1 - \alpha_2) r_{112} + (1 - \alpha_2) r_{121} + r_{122}. \end{aligned}$$

In the same manner, we can derive the expressions of other cell probabilities of Table 2 to get

$$p_{111} = \alpha_1 r_{111},$$

$$p_{112} = \frac{1 - \alpha_1}{3} r_{111} + \alpha_2 r_{112},$$

$$p_{121} = \frac{1 - \alpha_1}{3} r_{111} + \alpha_2 r_{121},$$

$$p_{211} = \frac{1 - \alpha_1}{3} r_{111} + \alpha_2 r_{211},$$

$$p_{122} = \frac{1 - \alpha_1}{3} r_{111} + (1 - \alpha_2) r_{112} + (1 - \alpha_2) r_{121} + r_{122},$$

$$p_{212} = \frac{1 - \alpha_1}{3} r_{111} + (1 - \alpha_2) r_{112} + (1 - \alpha_2) r_{211} + r_{212},$$

$$p_{221} = \frac{1 - \alpha_1}{3} r_{111} + (1 - \alpha_2) r_{211} + (1 - \alpha_2) r_{121} + r_{221}.$$

Then

$$\vec{p} = M_3 \times \vec{r}, \tag{3}$$

where M_3 is a 7 by 7 matrix determined by the above seven equations derived under Model (3).

For $\alpha_1 = \alpha_2 = \theta$, we get the same formulation as under Model (2). For the special case with $\alpha_1 = \alpha_2 = 1$, $p_{ijk} = r_{ijk}$, reducing to the traditional problem. Again, the capture probabilities remain the same, *i.e.*, $p_{1++} = r_{1++}, p_{+1+} = r_{+1+}, p_{++1} = r_{++1}$.

4. ESTIMATING THE SIZE OF THE POPULATION

4.1 Log-linear Model Formulation

For purposes of exposition, we confine our attention to the three-sample census case, although extensions to the k-sample census for $k > 3$ are straightforward. As before, let l_{ijk} and m_{ijk} be expected cell counts for Table 1 and Table 2 respectively. The relationship between the cell probabilities and the expected cell counts is $l_{ijk} = r_{ijk}N$, and $m_{ijk} = p_{ijk}N$. Let

$$\vec{m} = (m_{111}, m_{112}, m_{121}, m_{211}, m_{122}, m_{212}, m_{221})^T,$$

and

$$\vec{l} = (l_{111}, l_{112}, l_{121}, l_{211}, l_{122}, l_{212}, l_{221})^T.$$

Since for each of the models we have proposed in the last section, there is a matrix M with entries depending on the matching probability parameters in the chosen model such that $\vec{p} = M \times \vec{r}$, multiplying through by N gives

$$\vec{m} = M \times \vec{l}. \tag{4}$$

For any log-linear model specified for Table 1, it is straightforward to obtain the parameterization for m_{ijk}. For example, for any of the models suggested in Fienberg (1972), we can write the expected counts in terms of functions of u-term parameters:

$$l_{ijk} = g_{ijk}(u, u_1(i), u_2(j), u_3(k), u_{12}(ij), u_{13}(ik), u_{23}(jk)), \tag{5}$$

and then obtain the parameterization of $\{m_{ijk}, (ijk) \neq (222)\}$ from (4).

4.2 Estimating the Size of the Population

We now consider the matching rates in our various models as known. To obtain the estimate of the population size, we proceed as follows. First, following Sanathanan (1972), we compute the maximum likelihood estimates of u-term parameters from l_c, the conditional likelihood associated with Table 2 given n,

$$l_c = n! \prod_{\{(ijk) \neq (222)\}} \frac{(q_{ijk})^{x_{ijk}}}{x_{ijk}!},$$

where $n = \sum_{\{(ijk)\neq(222)\}} x_{ijk}$, and $q_{ijk} = m_{ijk}/n$. Sanathanan (1972) shows that, under suitable regularity conditions, the conditional maximum likelihood estimates and the unconditional ones are both consistent and have the same asymptotic normal distribution. If we remove redundant u-term parameters using the constraints associated with the specified log-linear model for Table 1, then the problem is to find the maximum of l_c subject to the following single constraint:

$$\sum_{\{(ijk)\neq(222)\}} m_{ijk} = n.$$

Numerically, this is a nonlinearly constrained optimization problem. Rao (1957) studied regularity conditions under which there exist unique maximum likelihood estimates of the parameters in a multinomial distribution. His conditions are satisfied by the parameterization of $\{q_{ijk}\}$. Once the conditional maximum likelihood estimates of the u-term parameters are obtained, we use the loglinear model specified for Table 1 to compute the conditional maximum likelihood estimates of $\{l_{ijk}\}$, the expected cell counts of Table 1 including the expected count of the missing cell. Then our estimate of N is

$$\hat{N} = \sum_{\{ijk\}} \hat{l}_{ijk}.$$

In the case of no matching errors, with $\alpha_1 = \alpha_2 = 1$ in Model (3), $m_{ijk} = l_{ijk}$. Thus

$$\hat{N} = n + \hat{m}_{222},$$

i.e., we get back to the estimation method for the traditional multiple recapture census problem developed by Fienberg (1972) when the log-linear models in Fienberg (1972) are considered.

As we have discussed earlier, a log-linear model is specified for Table 1 and the observations are viewed as falling into Table 2, whose parametric model of the expected cell counts is specified by the log-linear model and a chosen model for matching errors. To assess the appropriateness of a log-linear model specified for Table 1, we can apply the usual Pearson and likelihood ratio goodness-of-fit tests, X^2 and G^2, discussed in Fienberg (1972), to Table 2. Each statistic has an asymptotic χ^2 distribution under the null hypothesis that the model fits, with degrees of freedom equal to $2^k - 1 -$ (number of independent parameters in the model).

5. ANALYSIS OF 1988 ST. LOUIS DRESS REHEARSAL CENSUS DATA

Dual System Estimation (DSE), based on the standard two-sample census, has been employed by U.S. Bureau of Census for census coverage evaluation since 1950. In 1988, the Census Bureau conducted a Dress Rehearsal Census for the 1990 decennial census at three sites: St. Louis, Missouri; Columbia, Missouri; and western Washington State. Zaslavsky and Wolfgang (1993) present data for a population subgroup from the Post Enumeration Survey (PES) in the dress rehearsal census in St. Louis which focuses on urban Black male adults who are believed to be underestimated by dual system methods. The resulting data consists of three sources: the C-sample is the census itself; the P-sample was compiled from the PES; a third source of information was the Administrative List Supplement (ALS), compiled from pre-census administrative records of state and federal government agencies, encompassing Employment Security, driver's license, Internal Revenue Service, Selective Service, and Veteran's Administrative records. The C-sample and P-sample provide data for the implementation of the usual DSE or capture recapture approach. The ALS data can be combined with the Census and the P-sample for analysis from a three-sample perspective, though it was originally intended to improve the coverage of the P-sample. In Table 3, we present three-sample data for PES sampling stratum 11 in St. Louis obtained by collapsing the original data in Table 1 of Zaslavsky and Wolfgang (1993) over four poststrata defined by owners/renters × age 20-29, 30-44.

Table 3

Three-Sample Data for Stratum 11, St. Louis

	Census			
ALS	Present P-sample		Absent P-sample	
	Present	Absent	Present	Absent
Present	300	51	53	180
Absent	187	166	76	–

Such triple-system data can be analyzed with the matching error Model (2) and data from a separate Matching Error Study (MES, or rematch study) associated with the same sampling poststratum. The MES is one of the operations conducted by the Census Bureau to evaluate the PES, and typically operates for a sample of cases, using more extensive procedures, highly qualified personnel and reinterviews to obtain estimates of the bias associated with the previous matching process. In the discussion of the Matching Error Study done in a 1986 test census in Los Angeles, Hogan and Wolter (1988) state that "The rematch was done independently of the original match, and the discrepancies between the match and the rematch results are adjudicated. Because of this intensive approach to the rematch, we believe the rematch results represent true match status, while differences between the match and rematch results represent the bias in the original match results."

Table 4

St. Louis Rematch Study: *P*-sample
Source: Mulry, Dajani and Biemer (1989)

Original Match Classification	Rematch Classification			
	Matched	Not Matched	Un-resolved	Total
Matched	2,667	7	8	2,682
Not matched	9	427	30	466
Unresolved	0	7	20	27
Total	2,676	441	58	3,175

The data from the MES thus provides a basis for estimating error rates in the original matching process. Mulry, Dajani and Biemer (1989) report the MES operation for the 1988 Dress Rehearsal and rematch data for all three test sites, and in Table 4, we reproduce those data relevant for our purposes.

Let α be the matching rate between the *C*-sample and the *P*-sample, and $\gamma = 1 - \alpha$ be the nonmatch error rate. We assume no errors in the rematch. Then from the data in Table 4, we can estimate α by $\hat{\alpha} = 2667/(2667 + 9) = 99.6637\%$, and γ by $\hat{\gamma} = 1 - \hat{\alpha} = .3363\%$. The parameter θ is a three-sample matching rate for the *C*-sample, *P*-sample and the ALS. It takes two matches, say, one between the *C*-sample and the *P*-sample, and the other one between the *P*-sample and the ALS, in order to reach a correct (1,1,1) three-sample classification. In the absence of evaluation of the match between the census and the ALS, we assume that these two matches are independent of each other and that the matching rate for the *P*-sample and ALS is the same for the *C*-sample and the *P*-sample. Thus we can use $\theta = \alpha^2$, and $\hat{\theta} = \hat{\alpha}^2 = 99.3285\%$. Based on other qualitative information, this seems to be unreasonably high match rate, and the match error rate for the census and the ALS is probably higher than the match error rate between the census and the *P*-sample. In the absence of better quantitative information, however, we proceed to use it in the calculations that follow.

Table 5

Estimates Under Various Models

Log-linear Model	Usual MLE		MLE Using Matching Error Model (2)	
	$\hat{N}$ (S.E.)	Fit (d.f.)	$\hat{N}$ (S.E.)	Fit (d.f.)
[C] [P] [A]	1091.48 (11.24)	248.31 (3)	1083.58 (10.93)	244.56 (3)
[CP] [A]	1204.14 (23.31)	90.60 (2)	1194.73 (22.86)	87.30 (2)
[PA] [C]	1108.34 (13.77)	247.93 (2)	1100.03 (13.40)	244.53 (2)
[CA] [P]	1068.87 (10.47)	230.66 (2)	1061.09 (10.10)	226.42 (2)
[CP] [CA]	1271.11 (52.55)	87.16 (1)	1256.77 (50.97)	84.37 (1)
[CP] [PA]	1598.88 (106.26)	17.55 (1)	1585.03 (104.93)	15.88 (1)
[CA] [PA]	1080.47 (13.38)	230.43 (1)	1072.19 (12.88)	226.44 (1)
[CP] [CA] [PA]	2360.82 (363.25)	– (0)	2309.55 (352.36)	– (0)

Table 5 gives the estimates of the population size for various log-linear models with estimates of standard errors and goodness-of-fit statistics. Standard errors are computed with the delta method as discussed in Fienberg (1972). The assumption of independence between the census and the *P*-sample has been questioned for the use of the DSE. The dual system method has limited capacity to test this assumption and to adjust for potential dependency, while both can be handled through log-linear models for three or more samples. There are four models listed in Table 5 that assume independence between the census and the *P*-sample: the independence model [C] [P] [A], [PA] [C], [CA] [P], and [CA] [PA]. All of them fit the data poorly. The three models with the interaction term for the census and the *P*-sample, [CP] [A], [CP] [CA], and [CP] [PA] fit the data much better. With the addition of an interaction term linking the census and the ALS, model [CP] [CA] fits only slightly better than [CP] [A], indicating that the census and the *P*-sample are together nearly independent from the ALS. The model [CP] [PA] fits the data the best, suggesting that the usual independence assumption for the DSE is invalid and that there is dependence between the *P*-sample and the ALS. For all seven non-saturated log-linear models, we obtain better fits under matching error Model (2), though only slightly so, due to the high match rate for the data from the 1988 U.S. Census Dress Rehearsal. For the [CP] [PA] model, there is a .8738% difference in the estimate of *N* associated with the nonmatch rate of .3363%. If the nonmatch rate had been 10%, *i.e.*, a 90% match rate, and assuming that the difference in the estimate of *N* is approximately linear in the nonmatch rate, there would have been a 26% difference between the usual maximum likelihood estimate of *N* and our estimate.

Table 6

Dual-System Data for Stratum 11, St. Louis

P-sample	Census		
	Present	Absent	Total
Present	487	129	616
Absent	217	–	
Total	704		

Table 6 presents the usual dual system data for stratum 11, St. Louis. The number of people in both the census and the *P*-sample is $y_{11} = 300$, the number of those in the census only is $y_{12} = 217$, and number in the *P*-sample only is $y_{21} = 129$. The total census count is $y_{1+} = y_{11} + y_{12} = 704$, the total *P*-sample count is $y_{+1} = y_{11} + y_{21} = 616$, the dual system estimate is $\widehat{\text{DSE}} = y_{1+}y_{+1}/y_{11} = 893$ (p. 232, Bishop, Fienberg and Holland 1975), and the estimated variance of $\widehat{\text{DSE}}$ is $\text{Var}(\widehat{\text{DSE}}) = y_{1+}y_{+1}y_{12}y_{21}/y_{11}^3 = 105.4$ (p. 233, Bishop *et al.* 1975). The standard error is $\text{SE}(\widehat{\text{DSE}}) = 10.27$.

The census undercount for the population estimate $\widehat{\mathrm{DSE}}$ is $(\widehat{\mathrm{DSE}} - y_{1+})/\widehat{\mathrm{DSE}} \times 100\% = 21.17\%$. For our best fitting model, the census undercount is $(\hat{N} - y_{1+})/\hat{N} = 55.97\%$ for the estimate $\hat{N} = 1599$ assuming no matching error and 55.58% for $\hat{N} = 1585$ from matching error Model (2). Thus there is a 55.97% − 55.58% = 0.39% upward bias by ignoring matching errors. This is quite close to the figure of 0.37% computed in Ding and Fienberg (1994) for the 1986 Los Angeles test census data using a two-sample match rate of 99.4734%, as compared to 99.6637% here for the St. Louis data. Our estimates show that the urban Black male adults targeted in the St. Louis Dress Rehearsal were heavily undercounted by the census, and that the undercount is severely underestimated by the usual dual-system or capture-recapture estimator of the population size. A third and qualitatively different sample might work well for this demographic group.

The homogeneity of the capture probabilities is one of the assumptions in the standard approach to the estimation of the size of a closed population. Darroch *et al.* (1993) developed a quasi-symmetry model and a partial quasi-symmetry model to allow for varying catchability of individuals. The quasi-symmetry model assumes that the pattern of heterogeneity is the same for all three samples, the partial quasi-symmetry model assumes that the pattern of heterogeneity is the same for two samples but different for the third sample. This is a sensible model given that the third sample is qualitatively quite different from the census and the PES and this model is equivalent to a combination of dependence and heterogeneity. For the multinomial cell probabilities including the missing cell, $R = (r_{111}, r_{112}, \ldots, r_{222})$, both are log-linear models of the form $\log R = A\beta$ with an appropriately chosen design matrix A and a vector of parameters β. The design matrices for both models are given in Darroch *et al.* (1993).

Table 7

Heterogeneous Catchability Models

Log-Linear Model	MLE from Darroch *et al.* (1993)		MLE Using Matching Error Model (2)	
	$\hat{N}$ (S.E.)	Fit (d.f.)	$\hat{N}$ (S.E.)	Fit (d.f.)
Full quasi-symmetry	1923.63 (216.84)	133.54 (2)	1906.61 (213.47)	133.50 (2)
Partial quasi-symmetry	2576.54 (413.28)	11.70 (1)	2557.08 (409.39)	11.72 (1)

Our proposed method can readily incorporate heterogeneous catchability to estimate the population size by assuming a heterogeneity model for Table 1 and then adopting the conditional likelihood estimation (Sanathanan 1972). Table 7 presents estimates from fitting the quasi-symmetry model and the partial quasi-symmetry model for the data from stratum 11. Again, the effect of the matching errors in this analysis is not substantial due to the high matching rate. The partial quasi-symmetry model fits much better than the quasi-symmetry model, indicating there seems to be plausible heterogeneity and the pattern of heterogeneity seems different in the ALS. The lack of fit of the independence model might also be explained in part by the dependence among the samples (in particular between the census and the P-sample) and in part by heterogeneous catchability.

The partial quasi-symmetry model incorporates the [CP] dependence and thus is an alternative to the model [CP] [PA] in Table 5. The two models yield similar fits to the data, but they give dramatically different estimates of N, with the model incorporating heterogeneity having a much larger estimate accompanied by a much larger estimated standard error. This suggests that there is a considerable instability associated with heterogeneity parameters and, although the two models are not nested and thus not directly comparable, it seems reasonable to opt for the smaller and more stable estimate which does not incorporate heterogeneity.

Darroch *et al.* (1993) considered four substrata for stratum 11 in their analysis. The two cross-classification variables for the four substrata O2, R2, O3 and R3 are whether residents owned or rented homes and whether they were age 20-29 or 30-44. The data for the four substrata are given in Table 8 where 1 corresponds to presence in a sample and 0 is for absence. We have reanalyzed them for comparison. Table 9 and Table 10 give estimates for both heterogeneity models. As pointed out earlier, the high match rate yields similar estimates and fits for models incorporating matching errors. The partial quasi-symmetry model shows significant improvement in fits over the full quasi-symmetry model with the best fits obtained for R2 and R3. If we add the estimates of N across the four substrata, the total for the matching error version of partial quasi-symmetry is $\hat{N} = 2980.8$, more than 16% larger than the estimate from the collapsed model in Table 7. Of course, the standard error of the estimate has increased by a similar magnitude.

Table 8

Three-Sample Data for Four Substrata of Stratum 11
Source: Table 2, Darroch *et al.* (1993)

Sample			Substratum			
C	*P*	*A*	O2	R2	O3	R3
0	0	1	59	43	35	43
0	1	0	8	34	10	24
0	1	1	19	11	10	13
1	0	0	31	41	62	32
1	0	1	19	12	13	7
1	1	0	13	69	36	69
1	1	1	79	58	91	72

Table 9

Estimates for Full Quasi-Symmetry

Sub-stratum	MLE from Darroch *et al.* (1993)		MLE Using Matching Error Model (2)	
	$\hat{N}$ (S.E.)	Fit (d.f.)	$\hat{N}$ (S.E.)	Fit (d.f.)
O2	780.83 (294.81)	11.70 (2)	777.98 (293.99)	11.69 (2)
R2	394.34 (56.45)	41.09 (2)	391.14 (55.29)	41.02 (2)
O3	765.45 (254.57)	25.99 (2)	759.97 (252.44)	25.98 (2)
R3	361.83 (47.33)	59.31 (2)	358.71 (46.20)	59.22 (2)

Table 10

Estimates for Partial Quasi-Symmetry

Sub-stratum	MLE from Darroch *et al.* (1993)		MLE Using Matching Error Model (2)	
	$\hat{N}$ (S.E.)	Fit (d.f.)	$\hat{N}$ (S.E.)	Fit (d.f.)
O2	605.66 (212.63)	7.51 (1)	601.44 (210.93)	7.52 (1)
R2	652.34 (205.12)	0.04 (1)	646.59 (202.58)	0.04 (1)
O3	1124.00 (473.26)	8.27 (1)	1126.90 (476.54)	8.22 (1)
R3	611.78 (200.82)	2.92 (1)	605.91 (198.26)	2.92 (1)

6. SUMMARY

In this paper, we have presented models for matching errors and models for the estimation of the population total and census undercount in a multiple sample census. We have illustrated our methods by reanalyzing census coverage data from the 1988 St. Louis Dress Rehearsal census. Two sources of information are considered in our analysis, the data from a Matching Error Study (MES), and triple-system data with every individual cross-classified according to presence or absence in each of three samples: the census, a post enumeration survey (*P*-sample) and an administrative list supplement. We imbed the standard log-linear model formulation of Fienberg (1972) into our estimation procedure to account for statistical dependency together with matching errors and to allow for formal goodness-of-fit test of various models. Our method applies to any model of a log-linear form and we have illustrated how heterogeneity models can be incorporated into our approach to allow for both matching errors and heterogeneous catchability.

Our matching error models assume that false matches are negligible. Sensitivity analysis in Ding (1990) shows that when both the false nonmatch rate and the false match rate are the same order of magnitude, the matching bias is dominated by the false nonmatch rate (see also Fay, Passel, Robinson and Cowan 1988, p. 53). This is because the capture probabilities in the census and the post enumeration survey are high, and thus a comparable change in both the false nonmatch and false match rates has substantially more impact on false nonmatches than false matches. For the 1986 Los Angeles test census data, the estimates of false nonmatch rate and false match rate computed in Ding and Fienberg (1994) are about 0.5% and 0.8%, respectively. Based on these empirical findings, we have some reason to believe that, at least in the census application described here, our models for false nonmatch errors are reasonable approximations to reality.

We have analyzed the St. Louis triple-system data with an estimate of the matching rate taken from the MES. Matching rates may not be homogeneous over different population strata, and we suggest that the MES data associated with the same sampling stratum be used. We have developed formulation in §3 for the *k*-sample census, and our approach can be readily applied to a *k*-sample census with $k \geq 4$.

REFERENCES

BISHOP, Y.M.M., FIENBERG, S.E., and HOLLAND, P.W. (1975). *Discrete Multivariate Analysis: Theory and Practice.* Cambridge, MA: M.I.T. Press.

CHEN, T.T. (1979). Log-linear models for categorical data with misclassification and double sampling. *Journal of American Statistical Association*, 74, 481-488.

CORMACK, R.M. (1968). The statistics of capture-recapture methods. *Oceanography and Marine Biology, Annals Review*, 6, 455-506.

DARROCH, J.N. (1958). The multiple-recapture census, I: estimation of a closed population. *Biometrika*, 45, 343-359.

DARROCH, J.N., FIENBERG, S.E., GLONEK, G.F.V., and JUNKER, B.W. (1993). A three-sample multiple-recapture approach to census population estimation with heterogeneous catchability. *Journal of American Statistical Association*, 88, 1137-1148.

DING, Y. (1990). Capture-recapture census with uncertain matching. Ph.D. dissertation, Department of Statistics, Carnegie Mellon University, Pittsburgh, Pennsylvania.

DING, Y., and FIENBERG, S.E. (1994). Dual system estimation of census undercount in the presence of matching error. *Survey Methodology*, 20, 149-158.

FAY, R.E., PASSEL, J.S., ROBINSON, J.G., and COWAN, C.D. (1988). The coverage of population in the 1980 census. Bureau of the Census, U.S. Department of Commerce.

FIENBERG, S. E. (1972). The multiple recapture census for closed populations and incomplete 2^k contingency tables. *Biometrika*, 59, 591-603.

HOGAN, H., and WOLTER, K. (1988). Measuring accuracy in a Post-Enumeration Survey. *Survey Methodology*, 14, 99-116.

MULRY, M.H., DAJANI, A., and BIEMER, P. (1989). The Matching Error Study for the 1988 Dress Rehearsal. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 704-709.

RAO, C.R. (1957). Maximum likelihood estimation for the multinomial distribution. *Sankhyā*, 18, 139-148.

SANATHANAN, L. (1972). Estimating the size of a multinomial population. *Annals of Mathematical Statistics*, 43, 142-152.

SEBER, G.A.F. (1982). *The Estimation of Animal Abundance and Related Parameters*. New York: MacMillan.

ZASLAVSKY, A.M., and WOLFGANG, G.S. (1993). Triple System Modeling of Census, Post-Enumeration Survey and Administrative List Data. *Journal of Business and Economic Statistics*, 11, 279-288.

Survey Methodology, June 1996
Vol. 22, No. 1, pp. 65–75
Statistics Canada

Applying the Lavallée and Hidiroglou Method to Obtain Stratification Boundaries for the Census Bureau's Annual Capital Expenditures Survey

JOHN G. SLANTA and THOMAS R. KRENZKE[1]

ABSTRACT

The Lavallée-Hidiroglou (L-H) method of finding stratification boundaries has been used in the Census Bureau's Annual Capital Expenditures Survey (ACES) to stratify part of its universe in the pilot study and the subsequent preliminary survey. This iterative method minimizes the sample size while fixing the desired reliability level by constructing appropriate boundary points. However, we encountered two problems in our application. One problem was that different starting boundaries resulted in different ending boundaries. The other problem was that the convergence to locally-optimal boundaries was slow, *i.e.*, the number of iterations was large and convergence was not guaranteed. This paper addresses our difficulties with the L-H method and shows how they were resolved so that this procedure would work well for the ACES. In particular, we describe how contour plots were constructed and used to help illustrate how insignificant these problems were once the L-H method was applied. This paper describes revisions made to the L-H method; revisions that made it a practical method of finding stratification boundaries for ACES.

KEY WORDS: Convergence; Contour plots; Economic surveys.

1. INTRODUCTION

The primary objectives of the sample design of the Census Bureau's Annual Capital Expenditures Survey (ACES) are to meet desired reliability levels using operationally-feasible methodology and to stay within budget limitations. To achieve these goals, we implemented a stratified simple random sample design using a modified version of Lavallée and Hidiroglou's (L-H) (1988) approach of finding stratum bounds. This stratification method for skewed populations obtains optimal boundary points by minimizing the total sample size given a desired coefficient of variation (c.v.). Survey managers associated with a single-purpose survey having access to a single stratifier can benefit from its operational ease and cost reductions.

We considered several papers that documented other methods for finding size stratum boundaries. Hess, Sethi, and Balakrishnan (1966) compared several stratifying techniques. The popular Dalenius and Hodges method (Cochran 1977, p. 129) was considered easy to implement in our case but was initially ruled out because it was not designed with certainty strata in mind. Sethi's method (1963) of using standard distributions was not used because we thought it would be cumbersome to identify the distribution and sub-optimal to use standard distributions for each of the 80 ACES industries. Eckman's rule (1959) of equalizing the product of stratum weights and stratum range seemed to require rather ominous calculations.

The L-H method was the most appealing to our application. Designed specifically for skewed populations, which is often the case for economic surveys, it creates a boundary that defines the take-all stratum, and the optimal boundary point(s) for the take-some strata. It sometimes will create additional take-all strata if through Neyman Allocation, the stratum sample size is greater than or equal to the stratum size.

The L-H method goes through an iterative algorithm beginning with computing or arbitrarily setting the initial stratum boundaries. Then, stratum statistics are computed such as, the stratum size, mean, and the variance. These parameters are entered into boundary formulas that were derived from minimizing the sample size subject to a desired cv. If the new boundaries do not converge then the stratum statistics are calculated for the newly defined size strata. The cycle continues until the boundaries converge.

Schneeberger (1979) discussed the problem of finding optimal stratification boundaries. Schneeberger shows in the paper that when expressing this problem as a non-linear program, when solved by a gradient method, the solution may be relative or global minima, maxima, or saddle points of the variance of the sample mean. Detlefsen and Veum (1991) document this as a shortcoming of the L-H method when testing its application for the Census Bureau's Monthly Retail Trade Survey. In the L-H method, they found that many times the resulting boundaries differed substantially from where the initial boundaries were set,

[1] John G. Slanta, Manufacturing and Construction Division; Thomas R. Krenzke, Decennial Statistical Studies Division, U.S. Bureau of the Census, Washington, D.C. 20233, U.S.A.

so the minimum sample size attained was a local minimum. Geometrically, the sample size as a function of two strata boundaries, appears like a landscape with one or more bowl-shaped valleys. The L-H method begins in a region and descends until it reaches the lowest point. If more than one minimum exists, it will not continue to search for the global minimum. Therefore, one objective is to have initial boundaries that are in the neighborhood of the global minimum. Using starting boundaries resulting from a technique such as the Dalenius and Hodges method may help satisfy this desire.

Detlefsen and Veum (1991) also noted instances of slow or non-convergence. However, they also noted that convergence occurred faster when the number of strata was reduced and when starting boundaries were the same as the previous survey's sample selection boundaries. In order to defend ourselves against infinite loops in the computer program or a large number of iterations, we decided on doing two things. First, we implemented a sample design in which the L-H method would create sets of only three size strata. Second, we decided to implement stopping rules so that when the convergence rate appeared to slow down, the program stopped processing.

In this work, we give background information on the ACES and briefly describe the way the L-H method was applied. We show how contour plots and three-dimensional plots gave us justification for using the L-H method to get the final boundaries. We show how the contour plots address the convergence problem by showing how constraints can be setup to be met after each iteration. This would protect us against slow or non-convergence under the assumption that the marginal gain achieved is not worth the extra effort.

2. ACES BACKGROUND

The 1992 ACES was designed by the Census Bureau to be a large-scale operational test of the sampling, processing, programming, data entry, editing, and estimation procedures which extended beyond a 1991 pilot study, to prepare for the 1993 full-scale survey. Capital expenditure estimates for domestic activities were published at conglomerated industry levels from the 1992 survey. In addition, the 1991 and 1992 preliminary surveys provided valuable capital expenditure data that will be used in future sample design enhancements.

The sampling unit for the ACES was the company which may be comprised of several establishments. The sampled population included all active companies with five or more employees from all major industry sectors except Government. These sectors include mining, construction, manufacturing, transportation, wholesale and retail trade, finance, services, and a portion of the agriculture sector that includes agricultural services, forestry, fishing, hunting, and trapping. Only companies with domestic activity were included in the sampling frame. The Research and Methodology Staff of the Census Bureau's Industry Division constructed the sampling frame, selected the sample, and generated estimates.

The ACES sampling frame was constructed from the Census Bureau's Standard Statistical Establishment List (SSEL) in November 1992 using final 1991 data for single unit (SU) establishments and 1990 data for establishments associated with multiunit (MU) firms. Major exclusions from the frame were public administration, U.S. Postal Service, international establishments, establishments in Puerto Rico, Guam, Virgin Islands, and the Mariana Islands. EI Submasters which are SU records on the SSEL that are associated with MU establishments, establishments associated with agricultural production, and private households were also excluded from the frame.

The establishment-based file was consolidated into a company-based file. In addition, the 4-digit Standard Industrial Classification (SIC) codes for each company were recoded into ACES categories. The 80 ACES categories consisted of either 3-digit SICs or combinations of 3-digit SICs. The ACES sampling frame included approximately two million companies.

3. THE L-H METHOD APPLIED TO THE ACES

The universe of companies was classified into two major strata. Stratum I was an arbitrarily defined take-all stratum that consisted of large companies with more than 500 employees and over $100 million in assets. Stratum I companies were not classified into one ACES industry. For the estimated industry level payroll totals used in the calculation of the industry-level sample sizes, stratum I companies could contribute to more than one ACES industry depending on the number of different ACES industries the companies have payroll in, identified in the SSEL.

Stratum II contained companies that had five or more employees and had less than 500 employees. Stratum II companies were classified into one industry, even if engaged in more than one activity. Each company had frame information available for each of the ACES industries the company had activity in. However, the company's payroll contributed only to estimated total payroll for the industry that the company was classified in. Subsequently, within stratum II, for each ACES industry category, three size strata were created based on total company annual payroll using the L-H method.

A concern with the sample design is the result of companies being misclassified due to the measure of size being used. We classified each stratum II company into its highest payroll industry; however, companies self-report their capital expenditures into ACES industries on the ACES questionnaire. Companies may report in multiple industries. If too many companies self-report into industries other than where they were classified, then control on the reliability of the estimates is lost.

A similar concern is that the variation in payroll is not the same as the variation in expenditures. Since sample size is directly related to the variance, sample sizes may be different than what is really required. Therefore, since the correlation between payroll and expenditures is not high, the chances that reliability constraints will be met will diminish.

The application of the L-H method to the ACES 1992 preliminary survey sample design involved splitting stratum II into one take-all size stratum and two take-some size strata for each ACES industry. The boundaries were derived for each industry by taking the partial derivative of the sample size with respect to a boundary while fixing the other boundary. However, in practice, we allowed both boundaries to move simultaneously. This results in an iterative process of minimizing the sample size for each industry subject to c.v. constraints. Within stratum II for each ACES industry and assuming Neyman Allocation (Detlefsen and Veum 1991), the sample size equation that is minimized is,

$$n = n_{TA} + \frac{N\left(\sum_{j=1}^{2} W_j S_j\right)^2}{\frac{cv^2 Y^2}{N} + \sum_{j=1}^{2} W_j S_j^2}, \tag{1}$$

where, n_{TA} is the number of companies in the take-all size stratum within stratum II defined by the L-H method, N is the number of stratum II companies in the ACES industry of interest, $W_j = N_j/N$ is the stratum proportion, N_j is the number of stratum II companies for size stratum j, cv is the desired coefficient of variation for the ACES industry of interest, Y is the total payroll for stratum I and II for the ACES industry of interest defined by,

$$Y = \sum_{k=1}^{N_I} y_k + \sum_{j=1}^{3} \sum_{i=1}^{N_j} y_{ji},$$

N_I is the number companies in stratum I, and S_j is the standard deviation of payroll from the SSEL for size stratum j in stratum II defined by,

$$S_j = \sqrt{\frac{\sum_{i=1}^{N_j} (y_{ji} - \bar{Y}_j)^2}{N_j - 1}},$$

where, y_{ji} is the payroll value of company i of size stratum j for the ACES industry of interest, and $\bar{Y}_j$ is the mean of payroll for size stratum j.

The reliability level for each industry was an expected c.v. of 5% on payroll. It was not known, however, what standard errors would result for capital expenditures, as no capital expenditures data exist for the frame records. Companies responding in ACES industries different from the ones they contributed to in the sample design also caused the c.v.'s to fluctuate. The total number of companies selected for the ACES 1992 preliminary survey was 11,194, consisting of 1,500 stratum I companies and 9,694 stratum II companies.

4. CONVERGENCE INTO NEIGHBORHOODS

One of the problems with the L-H method is that it sometimes takes a large number of iterations before the boundaries converge; sometimes they never converge. Generally after just a few iterations, a large proportion of the improvement in the sample size has already occurred. Our goal was to be able to implement stopping rules so that when an area around a local minimum is reached, we can stop processing. This prompted our use of contour plots in analyzing the effect the boundaries have on the resulting sample size. It also allowed us to get a graphical view of the neighborhoods around the local minima. We will use two distributions to illustrate the benefits of reviewing contour plots.

4.1 Non-Skewed Distribution

The first example is a non-skewed distribution from Schneeberger's paper. This distribution is symmetric at $x = 1$ as shown in Figure 1.

$$f(x) = \begin{array}{lcl} 0 & & x \le 0 \\ 2x & & 0 < x \le 0.5 \\ 2(1-x) & \text{for} & 0.5 < x \le 1 \\ 2(x-1) & & 1 < x \le 1.5 \\ 2(2-x) & & 1.5 < x \le 2 \\ 0 & & 2 < x \end{array}$$

Schneeberger's objective was to find boundaries for three take-some strata using a gradient method. Using the objective function of $z = (\sum W_h \sigma_h)^2$, the results attained are listed in Table 1.

Table 1

Optimum Boundaries for Non-Skewed Distribution

	b_1	b_2	Optimum Point
(2a)	.50241	1.03985	Minimum
(2b)	.70910	1.29090	Saddle Point
(2c)	.96015	1.49759	Minimum

Source: Schneeberger (1979).

Table 2
L-H Boundaries for Three Take-Some Strata for Non-Skewed Distribution

N	Starting Method	1st Iteration			Iteration Within 5% of Sample Size				Final Iteration			
		b_1	b_2	n	b_1	b_2	n	iter.#	b_1	b_2	n	iter.#
50	$N_1 = N_2 = N_3$	.59	1.41	10.89	.66	1.34	9.98	2	.70	1.31	9.77	4
100	$N_1 = N_2 = N_3$	.59	1.41	12.60	.66	1.34	10.91	2	.70	1.30	10.55	5
200	$N_1 = N_2 = N_3$	.59	1.41	13.42	.66	1.34	11.43	2	.71	1.29	10.99	6
1000	$N_1 = N_2 = N_3$	.59	1.41	13.85	.66	1.34	11.75	2	.71	1.29	11.37	7
5000	$N_1 = N_2 = N_3$	.59	1.41	14.12	.66	1.34	11.84	2	.71	1.29	11.45	9
50	Dalenius-Hodges	.70	1.40	10.09	.70	1.40	10.09	1	.77	1.37	9.63	4
100	Dalenius-Hodges	.70	1.40	10.90	.84	1.40	10.14	7	.93	1.47	9.65	13
200	Dalenius-Hodges	.70	1.40	11.42	.83	1.40	10.44	7	.95	1.49	9.96	17
1000	Dalenius-Hodges	.70	1.40	11.86	.86	1.42	10.67	8	.96	1.50	10.27	23
5000	Dalenius-Hodges	.70	1.40	11.95	.86	1.42	10.74	8	.96	1.50	10.34	28
50	Off Line	.50	1.30	10.87	.57	1.20	9.43	3	.55	1.11	9.11	6
100	Off Line	.50	1.30	11.95	.57	1.18	10.04	3	.53	1.07	9.65	8
200	Off Line	.50	1.30	12.64	.56	1.14	10.28	4	.51	1.05	9.96	12
1000	Off Line	.50	1.30	13.24	.56	1.14	10.59	4	.50	1.04	10.27	18
5000	Off Line	.50	1.30	13.37	.56	1.14	10.67	4	.50	1.04	10.34	24

We generated five datasets of different sizes (*e.g.*, $N = 50$, 100, 200, 1000, and 5000) using the formula, $F(x) = (j - 1/2)/N$. For this example, we adapted the L-H method to construct three take-some strata and no take-all stratum in order to compare our results with the results in the Schneeberger paper. With our application of estimating totals, when minimizing the sample size subject to a c.v. $= 0.05$, the L-H method ran for each of the five population sizes using three different starting techniques. The results are given in Table 2.

There are three main points from the information in Table 2. First, the algorithms convergence depends on the population size. The underlying theory of the L-H method is based on continuous distributions. Our examples and any survey application has discrete data from finite populations. It is also apparent that as N gets larger, the resulting boundaries get closer to where the minimum is under an infinite population size. Figure 2 shows the roughness of the sample size surface when N is small (*i.e.*, $N = 50$). The resulting surface illustrates the saddle in three dimensions in Figure 2. In this graph, the axes are the lower and upper boundaries and the surface is the resulting sample sizes. This graph shows the saddle-point, the two local minima, and it also gives a picture of the magnitude of the sample size reductions as a result of shifting the boundaries. In contrast, Figure 3 shows the smoothness of the surface when N is large (*i.e.*, $N = 5000$). From this, it seems that the roughness of the sample size surface and consequently the population size has an effect on where the boundaries converge.

The second point of this example reemphasizes that the ending boundaries are dependent on the starting boundaries. For this example, Schneeberger describes that with a starting point symmetric to $x = 1$, where $b_1 = 1 - \lambda$ and $b_2 = 1 + \lambda (0 < \lambda < 1)$ which defines the line $b_2 = 2 - b_1$, the gradient method moves the gradient along the line $b_2 = 2 - b_1$ into the saddle-point. When we set the starting boundaries on this line, which occurred when we started with the condition $N_1 = N_2 = N_3$, the L-H method also converged to the saddle point (see Table 1). With starting boundaries from the Dalenius-Hodges method, which are not on the line in the case where $b_2 > 2 - b_1$, the L-H method converged to a minimum (2c). The Dalenius-Hodges method works well in this example because of the three take-some strata. With starting boundaries which are not on the line in the case where $b_2 < 2 - b_1$ (specifically, $b_1 = .5$ and $b_2 = 1.3$), the L-H method converges to a different minimum (2a). This problem is not unique to the L-H method, as Schneeberger points out that the gradient method's resulting boundaries are also dependent of the starting boundaries.

The third point of this example is that there seems to be relatively large reductions in sample size in the first few iterations and then there are several iterations where there are small reductions in sample size. Results are shown in Table 2 from the iteration in which the algorithm produced a sample size within 5% of the final sample size. This implies that the L-H algorithm quickly goes to a neighborhood around an optimal boundary. While close to an optimal sample size, there seems to be a wide range of boundary points resulting in a small range of sample sizes. The point is that stopping rules can save computing time while not relinquishing any real reduction in sample size, since sample size is in integer values.

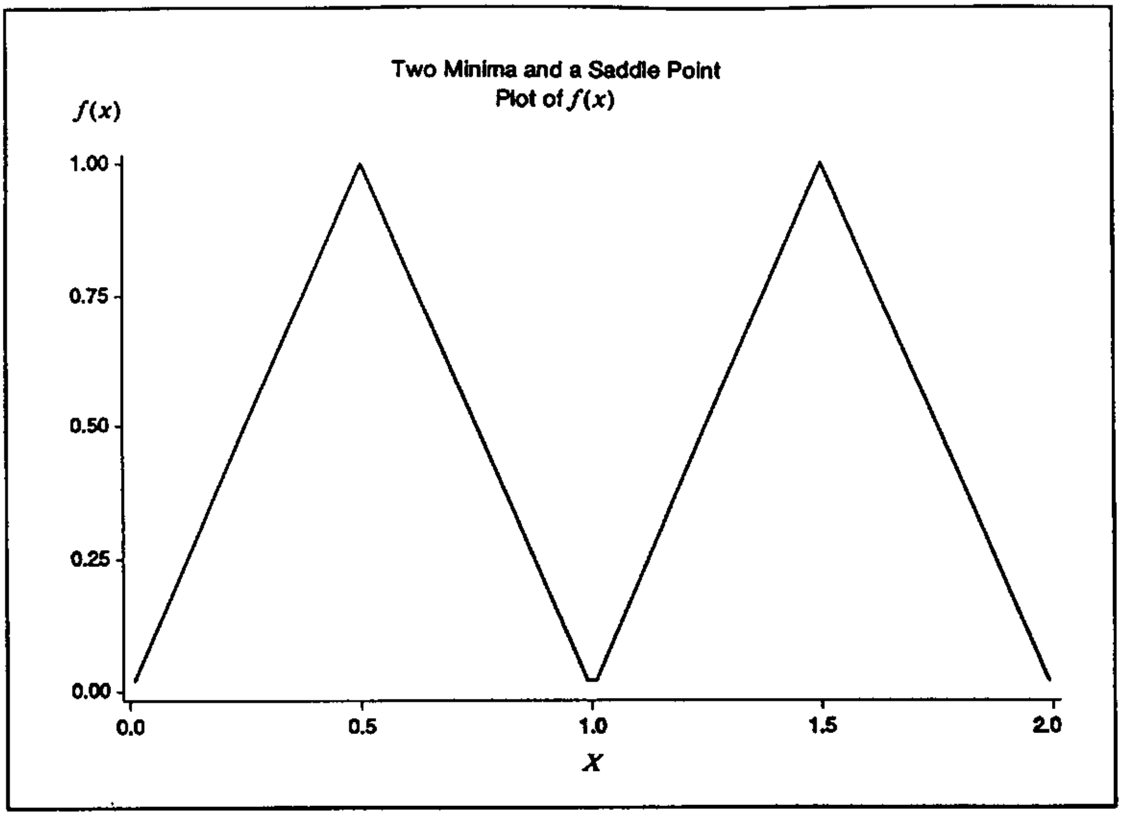

Figure 1. Graph of non-skewed distribution.

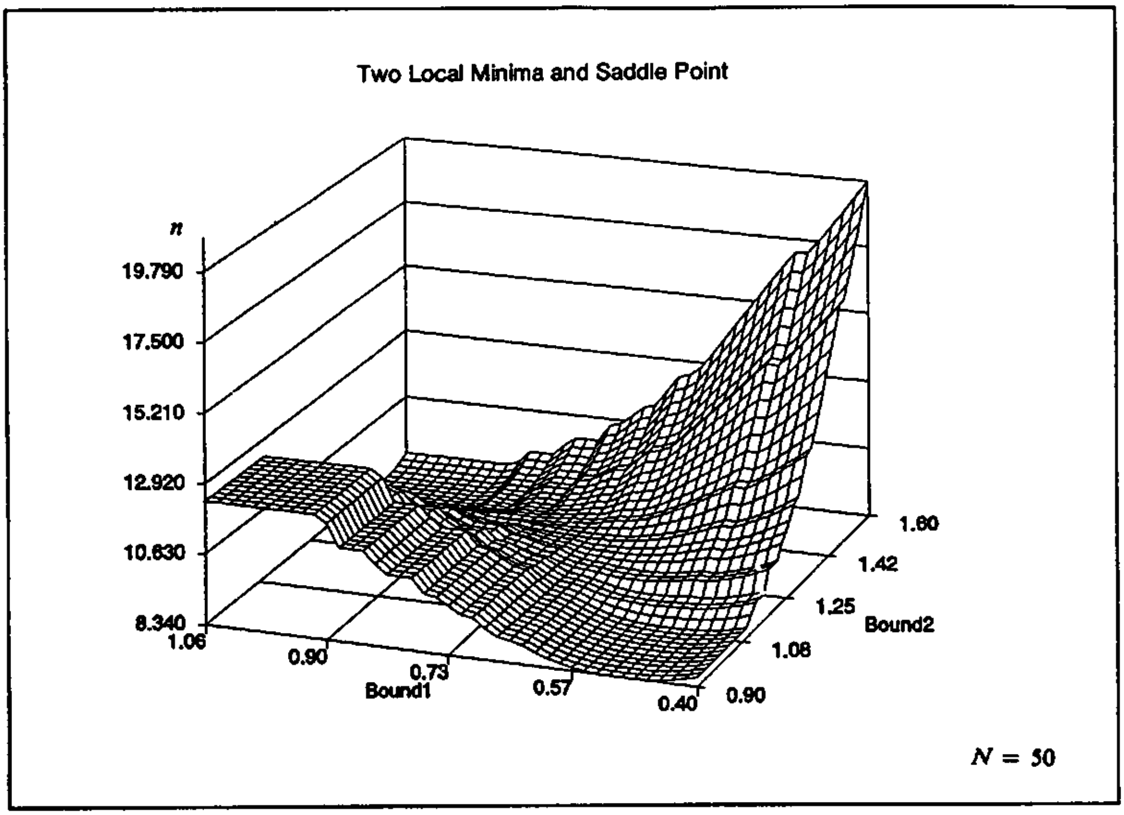

Figure 2. Sample size surface for non-skewed distribution ($N = 50$).

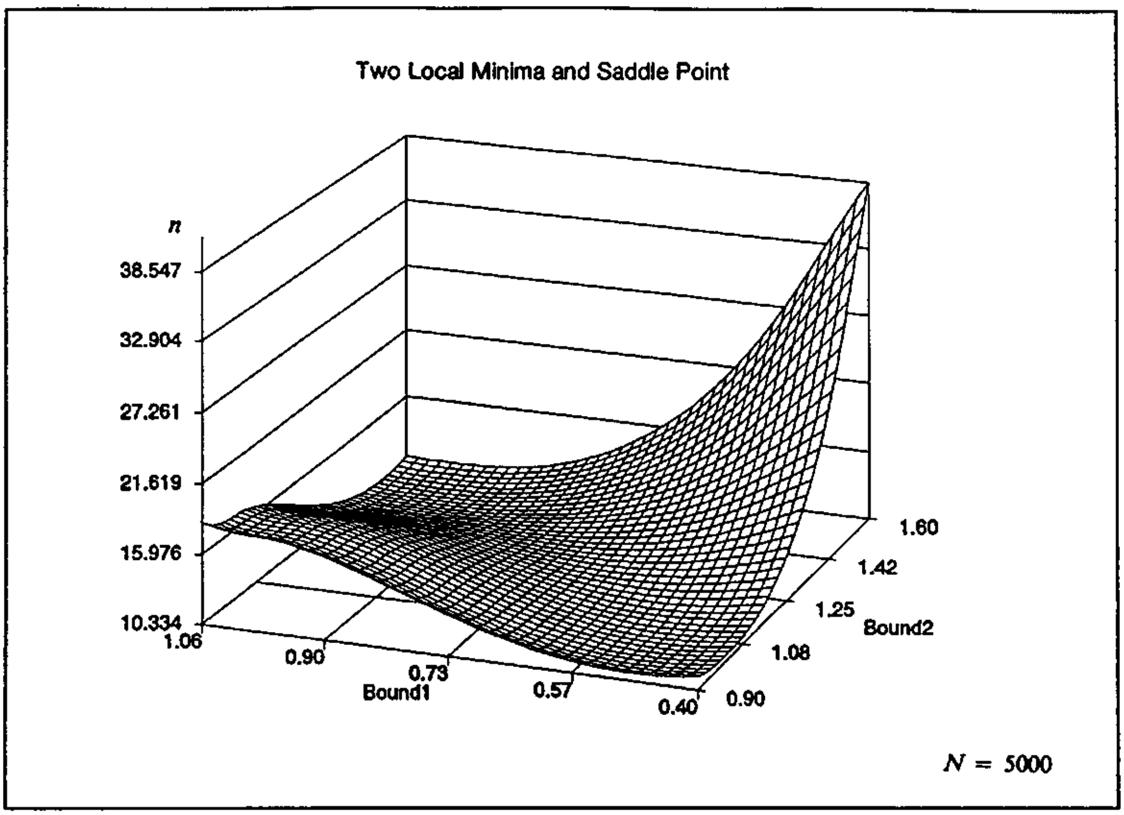

Figure 3. Sample size surface for non-skewed distribution ($N = 5000$).

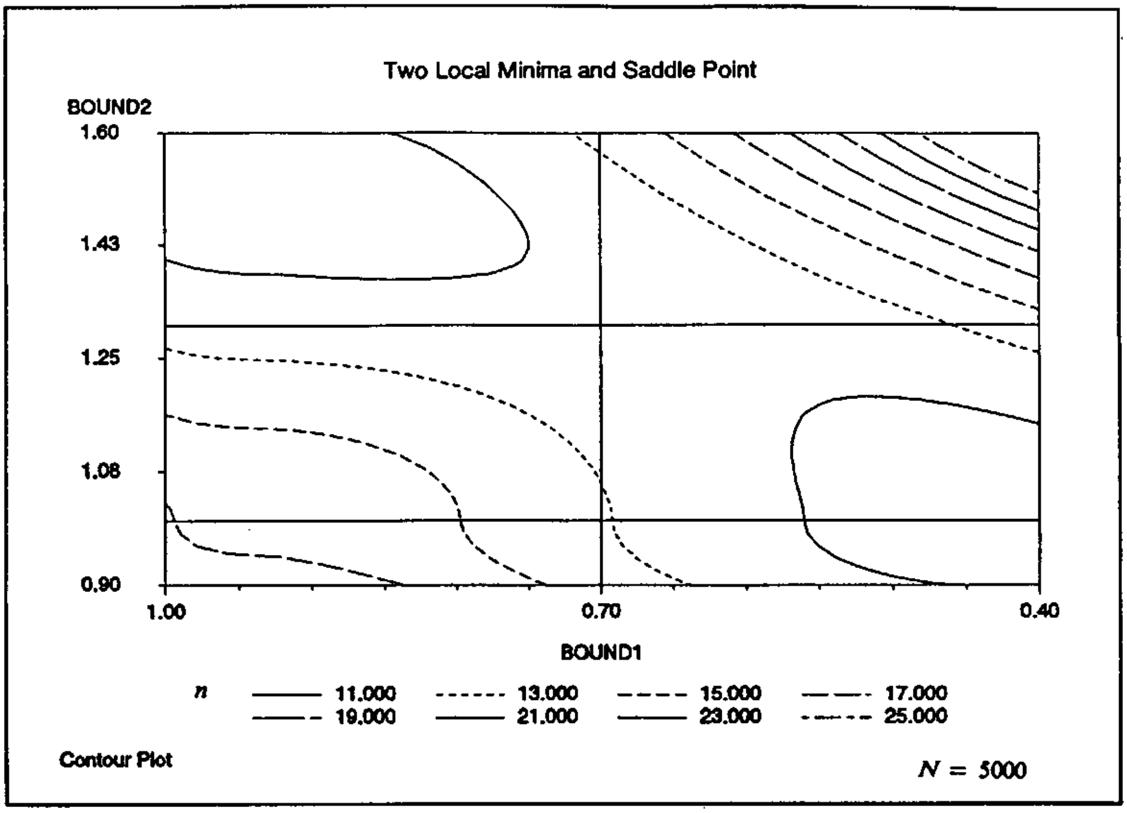

Figure 4. Contour plot for non-skewed distribution ($N = 5000$).

A contour plot of the surface shown in Figure 3 is given in Figure 4. Again, the axes are the lower and upper boundaries and the surface is defined by the resulting sample size. The lines in the plot represent a sample size value. The space between the lines gives an area that contains a range of sample size values. For example, a solid line represents a sample size of 11 and a series of short dash marks represents a sample size of 13. The area in between the solid line and the line of short dash marks contains sample sizes in the range of 11 to 13. This contour plot shows a marginal improvement in the sample size by illustrating that when an area around the bottom of the surface is reached, moving on is unnecessary. At this point, most of the improvement on the sample size from iteration to iteration is less than a value of one. It becomes apparent that after the first few iterations, the improvement of the sample size from iteration to iteration reduces quickly. For instance, in Table 2, where $N = 5000$ and where the Dalenius-Hodges method was used for the starting boundaries, the first eight iterations accounted for 74% of the total reduction in the sample size from iteration 1 to the 28th and final iteration.

4.2 A Skewed Distribution

Economic data are usually highly skewed and therefore it is more appealing to have a take-all stratum. The next example comes from the Pareto distribution, which is a very typical distribution of economic universes, where there are a large number of small companies and a small number of large companies.

The Pareto distribution function is defined as $F(x) = 1 - 1/(1 + x)^b$, $0 \leq x < \infty$. From this we again generated five datasets of different sizes using the formula $F(x) = (j - 1/2)/N$. We let the values of b change as the population size changed. This was done so as to keep the upper tail of the finite discrete distribution roughly the same proportion to the entire population for each population size. To do so, the parameter b was chosen in such a way that about 90% of the total sum could be accounted for in the top 20% of all possible sampling units. Since the datasets contain a finite number of discrete values there was no problem deriving variances of different strata when values of b were less than 2.

Table 3 gives the L-H results for different population sizes and starting points. The first group uses starting values which yield equal stratum populations ($N_1 = N_2 = N_3$). The second group uses the Dalenius-Hodges method to obtain all initial boundaries. The third group obtains starting boundaries by first using a method for determining the take-all boundary as presented by Hidiroglou (1986) and uses the Dalenius-Hodges method for the other boundary. Again it can be observed that the sample size surface given strata boundaries is much more choppier for smaller population sizes (see Figure 5). For example, when $N = 50$ and b_1 is fixed, there was only one sample size when b_2 varied between 11.8 and 14.7. This is because there were no values within this range in the population. As the population size increases, the data values are closer together, and the sample surface becomes very smooth (see Figure 6).

Table 3

L-H Boundaries for Skewed Distribution (one take-all stratum, two take-some strata)

N	Starting Method	1st Iteration					Iteration Within 5% of Sample Size					Final Iteration					
		b	b_1	b_2	n_{TA}	n	b_1	b_2	n_{TA}	n	iter.#	b	b_1	b_2	n_{TA}	n	iter.#
50	$N_1 = N_2 = N_3$	.80	.63	2.81	17	17.2	1.66	10.20	7	9.6	5	.80	2.44	11.81	7	9.4	9
100	$N_1 = N_2 = N_3$	.90	.56	2.33	34	34.3	1.61	10.29	11	15.8	5	.90	2.58	12.44	10	15.1	12
200	$N_1 = N_2 = N_3$	.90	.56	2.36	67	67.2	2.35	17.04	15	21.8	6	.90	3.61	20.46	13	20.9	13
1000	$N_1 = N_2 = N_3$	1.00	.50	2.00	333	334.2	3.35	30.58	32	53.0	7	1.00	4.93	36.32	27	51.3	18
5000	$N_1 = N_2 = N_3$	1.05	.47	1.85	1665	1667.2	4.67	64.33	62	113.5	7	1.05	7.39	79.38	50	108.8	22
50	Dalenius-Hodges	.80	1.25	8.04	9	10.5	1.76	10.37	7	9.5	3	.80	2.44	11.81	7	9.4	6
100	Dalenius-Hodges	.90	1.39	8.98	13	16.6	1.62	10.16	11	15.8	2	.90	2.58	12.44	10	15.1	9
200	Dalenius-Hodges	.90	1.82	11.66	20	24.3	2.45	17.29	15	21.7	3	.90	3.61	20.46	13	20.9	10
1000	Dalenius-Hodges	1.00	2.37	17.28	55	65.6	3.15	29.70	33	53.5	3	1.00	4.93	36.32	27	51.3	15
5000	Dalenius-Hodges	1.05	3.09	26.27	155	175.0	4.98	66.28	60	112.3	4	1.05	7.39	79.38	50	108.8	19
50	Hidiroglou 1986	.80	.94	6.50	10	11.3	1.58	10.02	7	9.6	3	.80	2.44	11.81	7	9.4	7
100	Hidiroglou 1986	.90	.74	6.17	17	19.6	1.66	10.38	11	15.8	4	.90	2.58	12.44	10	15.1	11
200	Hidiroglou 1986	.90	1.39	9.55	24	27.2	2.50	17.58	14	21.5	4	.90	3.61	20.46	13	20.9	10
1000	Hidiroglou 1986	1.00	2.02	15.13	62	71.3	3.34	30.54	32	53.0	4	1.00	4.93	36.32	27	51.3	15
5000	Hidiroglou 1986	1.05	3.24	28.72	142	164.1	5.11	67.05	59	112.0	4	1.05	7.39	79.38	50	108.8	19

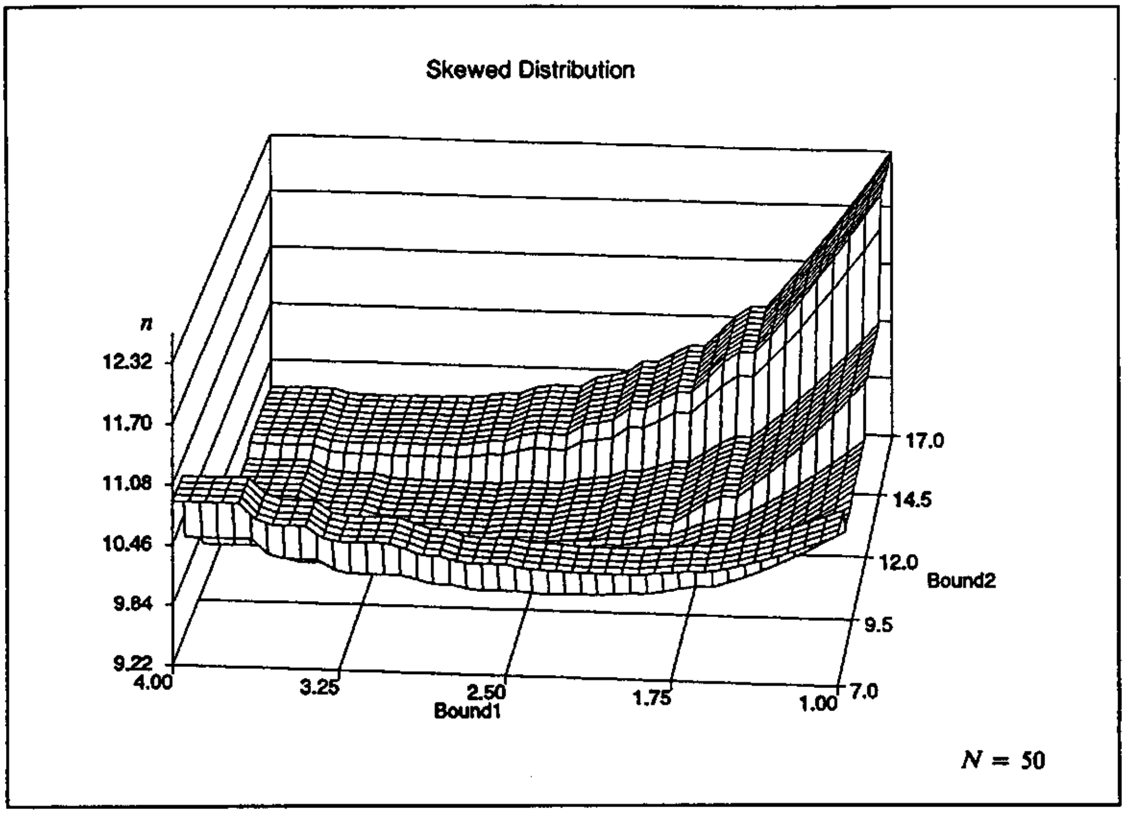

Figure 5. Sample size surface for skewed distribution ($N = 50$).

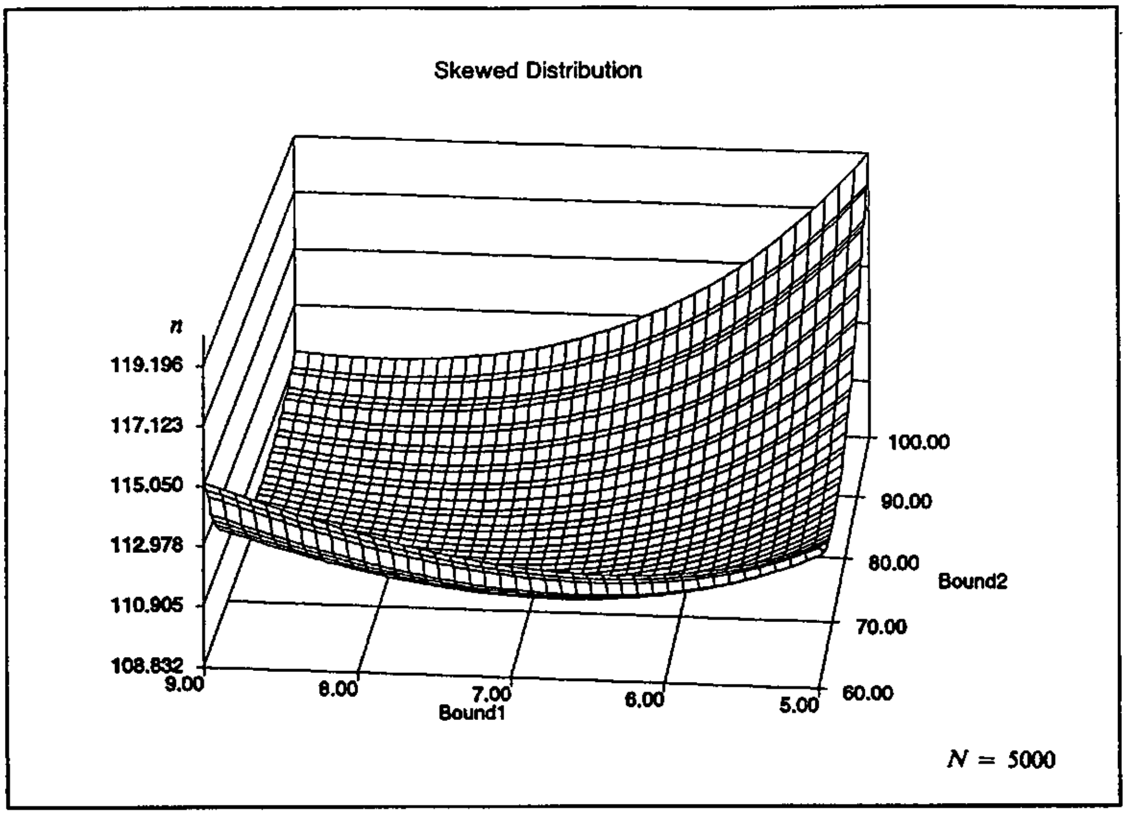

Figure 6. Sample size surface for skewed distribution ($N = 5000$).

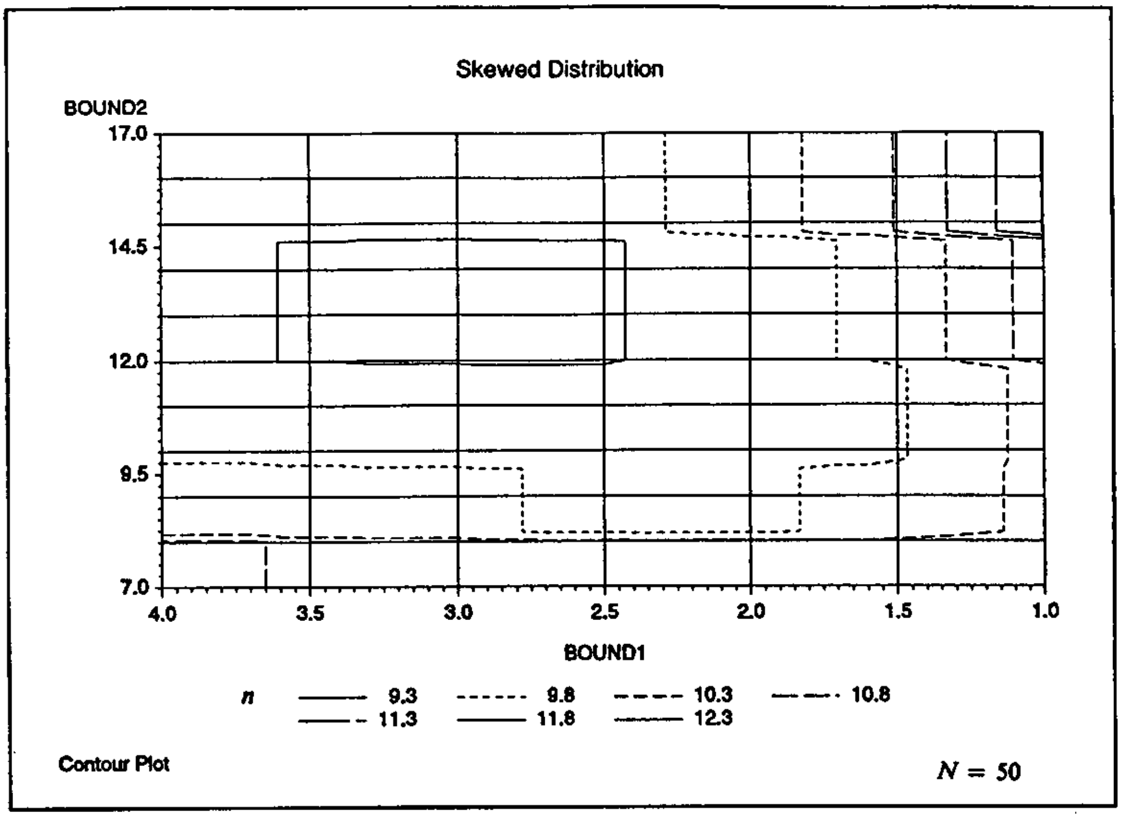

Figure 7. Contour plot of skewed distribution ($N = 50$).

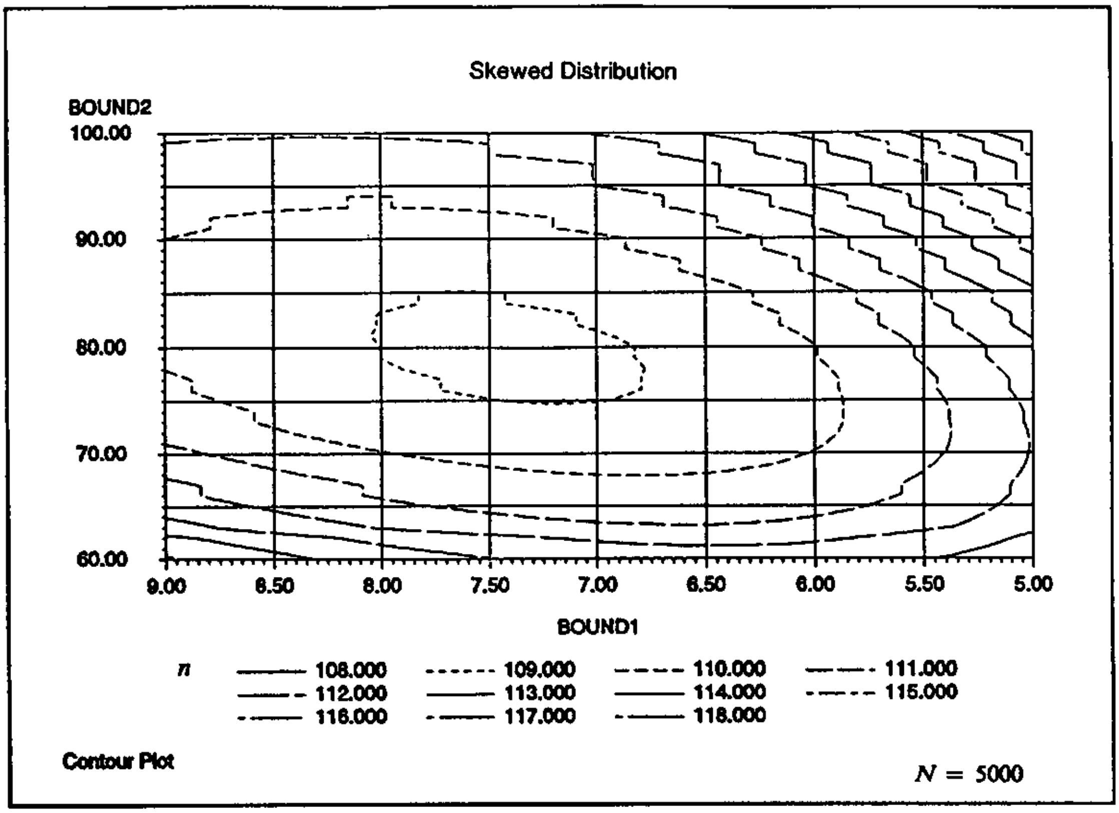

Figure 8. Contour plot of skewed distribution ($N = 5000$).

The contour plot for $N = 50$ (Figure 7) has erratic shapes defined by straight lines for contour markings. The contour plot for $N = 5000$ (Figure 8) has almost smooth concentric ellipses for contour markings. It would appear to be a desirable quality for the contour markings to be the same shape and concentric. This would imply that the global minimum is the only local minimum.

The contour plot for $N = 50$ demonstrated the case where the L-H method didn't converge to optimal boundaries. Since, for this example, we let the L-H program run until it converged the question may arise as to why the L-H method didn't converge to the optimal boundaries. The easiest way to explain this is by viewing Figure 5. We can see that when the population size is small then the sample size surface is not as smooth as in Figure 6. We see several major ridges in Figure 5 that are caused by wide gaps in the skewed discrete data ($x_{43} = 9.71$, $x_{44} = 11.81$, $x_{45} = 14.79$, $x_{46} = 19.29$). This means that for a given b_1, any value of b_2 between 11.81 and 14.79 would yield the same sample size. When we ran the L-H program for different starting boundaries other than the three listed in Table 3 we came up with the final boundaries as in Table 3 along with other boundaries and their corresponding sample sizes. It appears that the L-H method converges to a low region on one of the major ridges, provided that the region is in the neighborhood of the optimal boundaries. The minimum sample size is 9.22 and the L-H method in Table 3 yielded a sample size of 9.36. The smallest whole integer sample size for each result that meets or exceeds the constraint is 10. Here again we see that the L-H method performs exceptionally well even with discrete distributions that have small population sizes as we see that the boundaries converge within the neighborhood containing the optimal solution.

Another observation to be pointed out is that there is a broad range of values that the boundaries can take on while keeping the integer value of the sample size the same. As the size of the neighborhood expands, the range of boundary values extends as well. It should also be pointed out that even though the range of b_1 values for a given neighborhood is smaller than the range of values for b_2, there are far more sampling units in the range of b_1 than b_2 because of the skewed distribution.

5. SUMMARY

The graphs presented here have shown that a wide range of boundary values result in a small range of sample sizes when in a neighborhood around an optimal value (the bowl shape bottom of the graphs). Any extraordinary improvement on the sample size, *i.e.*, a small marginal gain, might not be worth the extra effort to obtain. This marginal gain may or may not even improve the sample size since the sample size is really an integer and the marginal gain might only be a small fraction. The L-H method proved very effective in obtaining boundary values in a desired neighborhood around an optimal value, and did it relatively fast.

By measuring the rate of convergence using the sample size instead of boundary values we were better able to determine when a desired neighborhood around an optimal value was reached. This is because boundary values vary greatly in such a neighborhood while sample size (which is of main interest) varies slightly. When the improvement in sample size from iteration to iteration was marginal or nonexistent we immediately terminated the program under the assumption that we reached the desired neighborhood. The following stopping rules are recommended. Stop processing when:

1) the difference between the new upper boundary and the previous iteration's upper boundary is less than one. The whole number, one, is used in our case since payroll values are only available to us in whole number values and any shifting of boundaries of a value less than one does not affect any companies;
2) the difference between the new lower boundary and the previous iteration's lower boundary is less than one;
3) the difference between the new sample size and the previous iteration's sample size is less than a small arbitrary value. We recommend a number less than one since sample sizes are usually rounded up and any fractional improvement on the sample size is negligible. One should be careful when choosing this value since it is possible that the sample size reduction rate may increase from iteration to iteration because the slope of the surface changes;
4) the program goes into the 30th iteration. Of course, this is an arbitrary value and may depend on the number of times (industries) one has to apply the L-H method.

Another note is that small population sizes may cause convergence of the boundaries to a point suboptimal, as shown in the examples. Graphs of the sample size surface show a rough surface for small populations and a smooth surface for large populations. It is this rough surface due to the discrete nature of the small population that contribute, in part, to where the L-H method converges.

Another point in conclusion, in our application, the Dalenius-Hodges method assumes that all resulting strata will be sampled. The L-H method is written to construct an analytical take-all substratum. Therefore, the top stratum developed by the Dalenius-Hodges method, when creating the initial boundaries for ACES industries, will be top-heavy since it will not be sampled. Improvements in the sample size were noticed from the Dalenius-Hodges method to the first iteration of the L-H method in this situation. The error that occurs is that the starting boundaries may lead to a local minimum that is not the best solution.

ACKNOWLEDGEMENTS

The authors are grateful to Michel Hidiroglou for useful comments and discussion. We also thank Carol (Veum) Caldwell, Easley Hoy, the referees from *Survey Methodology*, and the Research and Methodology Branch managers of the Manufacturing and Construction Division for helpful comments during review.

REFERENCES

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

DETLEFSEN, R., and VEUM, C. (1991). Design issues for the retail trade sample surveys of the U.S. Bureau of the Census. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 214-219.

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

HESS, I., SETHI, V.K., and BALAKRISHNAN, T.R. (1966). Stratification: A practical investigation. *Journal of the American Statistical Association*, 61, 74-90.

HIDIROGLOU, M.A. (1986). The construction of a self-representing stratum of large units in survey design. *The American Statistician*, 40, 27-31.

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

SCHNEEBERGER, H. (1979). Saddle-points of the variance of the sample mean in stratified sampling. *Sankhŷa*, Series C, 41, 92-96.

SETHI, V.K. (1963). A note on optimum stratification of populations for estimating the population means. *American Journal of Statistics*, 5, 20-23.

Survey Methodology, June 1996
Vol. 22, No. 1, pp. 77–83
Statistics Canada

A New Method to Reduce Unwanted Ripples and Revisions in Trend-Cycle Estimates From X-11-ARIMA

ESTELA BEE DAGUM[1]

ABSTRACT

The estimation of the trend-cycle with the X-11-ARIMA method is often done using the 13-term Henderson filter applied to seasonally adjusted data modified by extreme values. This filter however, produces a large number of unwanted ripples in the final or "historical" trend-cycle curve which are interpreted as false turning points. The use of a longer Henderson filter such as the 23-term is not an alternative for this filter is sluggish to detect turning points and consequently is not useful for current economic and business analysis. This paper proposes a new method that enables the use of the 13-term Henderson filter with the advantages of: (i) reducing the number of unwanted ripples; (ii) reducing the size of the revisions to preliminary values and (iii) no increase in the time lag to detect turning points. The results are illustrated with nine leading indicator series of the Canadian Composite Leading Index.

KEY WORDS: Trend-cycle; X-11-ARIMA; Turning points; Leading economic indicators.

1. INTRODUCTION

The estimation of the trend-cycle with the X-11-ARIMA seasonal adjustment method (Dagum 1980, 1988) as well as the U.S. Bureau of the Census X-11 variant (Shiskin, Young and Musgrave 1967) is done by the application of linear filters due to Henderson (1916). These Henderson filters are applied to seasonally adjusted series where the irregulars have been modified to take into account the presence of extreme values. The length of the filters is automatically selected on the basis of specific values of noise to signal ratios (I/S) being the most commonly chosen the 13-term filter.

The problem of trend-cycle estimation has attracted the attention of several authors, among others, Rhoades (1980); Cholette (1981, 1982); Kenny and Durbin (1982); Castles (1987); Dagum and Laniel (1987); Cleveland, Cleveland, McRae and Terpenning (1990); Wallgren and Wallgren (1990); Gray and Thomson (1990); Findley and Monsell (1990); Scott (1990); and Kenny (1993). Nevertheless, most statistical agencies (excepted the Australian Bureau of Statistics) concentrate their publications on seasonally adjusted series and only very few provide some sort of information on the trend-cycle, usually under the form of graphs.

There are several reasons for limiting the publication of trend-cycle estimates. In the majority of the cases, the seasonally adjusted data are already smooth enough as to be able to provide a clear signal of the short-term trend. But for highly volatile series where further smoothing is required the main objections for trend-cycle estimation are: (1) the size of the revisions of the most recent values (generally much larger than for the corresponding seasonally adjusted estimates) and (2) the presence of short cycles or ripples (9 and 10 months cycles) in the final trend-cycle curve when the 13-term Henderson filter is applied. On this regard, Kenny (1993) has argued that the presence of ripples in the final estimates of the trend-cycle leads to a large number of false turning points, making the 13-term filter unsuitable for monitoring turning points. He has proposed the use of the 23-term Henderson filter with the object of obtaining a much smoother trend. However, it is well known that this longer filter is sluggish to detect turning points and, hence not useful for current economic and business analysis. For this latter viewpoint, the 13-term filter is preferable but it produces ripples which can be interpreted as false turning points (an unwanted property).

The main purpose of this study is to introduce a new method by which the 13-term Henderson filter can be used with the advantages of: (1) reducing the number of unwanted ripples, (2) reducing the size of the revisions made to the most recent estimates when new observations are added to the series, and (3) not increasing the time lag to detect turning points.

2. TREND-CYCLE CASCADE FILTERS

The 13-term Henderson filter is the most often selected and combined with the standard seasonal filters (5- and 7-term moving averages) produces a symmetric cascade filter for final or central values (at least four years from each end of the series) with a gain as exhibited in Figure 1.

Figure 1 also shows the gain functions of other filter convolutions, namely: (1) short seasonal filters with the 9-term Henderson filter and (2) long seasonal filters with the 23-term Henderson filter. It is apparent that cycles of 9 and 10 months (in the 0.08-0.16 frequency band) will not be suppressed by any of the cascade filters, particularly,

[1] Estela Bee Dagum, Faculty of Statistical Sciences, University of Bologna, Via delle Belle Arti 41, (40126) Bologna, Italy.

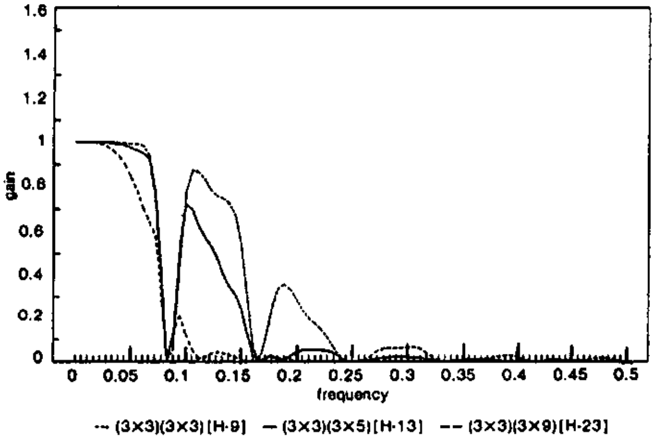

Figure 1. Trend-cycle symmetric cascade filters.

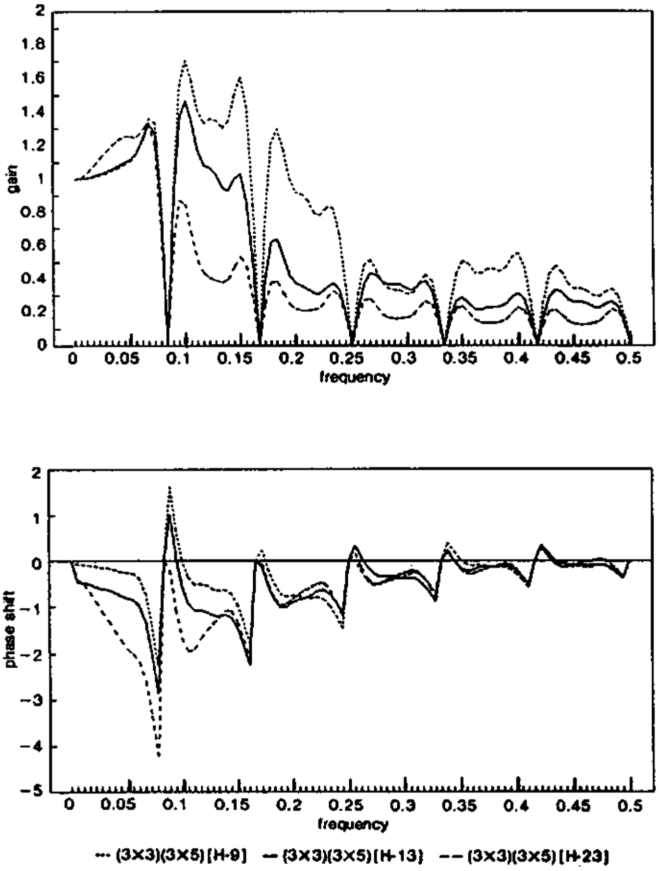

Figure 2. Trend-cycle concurrent cascade filters. Standard seasonal m.a. combined with three Henderson filters.

those using the 9- and 13-term Henderson filters. In fact, the symmetric trend-cycle cascade filter that results from the 9-term Henderson passes about 90% of the power of these short cycles; 72% and 21% are passed by the 13- and 23-term Henderson filters, respectively.

For the concurrent trend-cycle filters which are applied to the last available observation, the peak reached at the frequency band corresponding to 9 and 10 months cycles is even larger (see Figure 2). Furthermore, all these asymmetric filters introduce phase shift, being near to two months for the 23-term (the largest), one month for the 13-term, and one-half month for the 9-term filter.

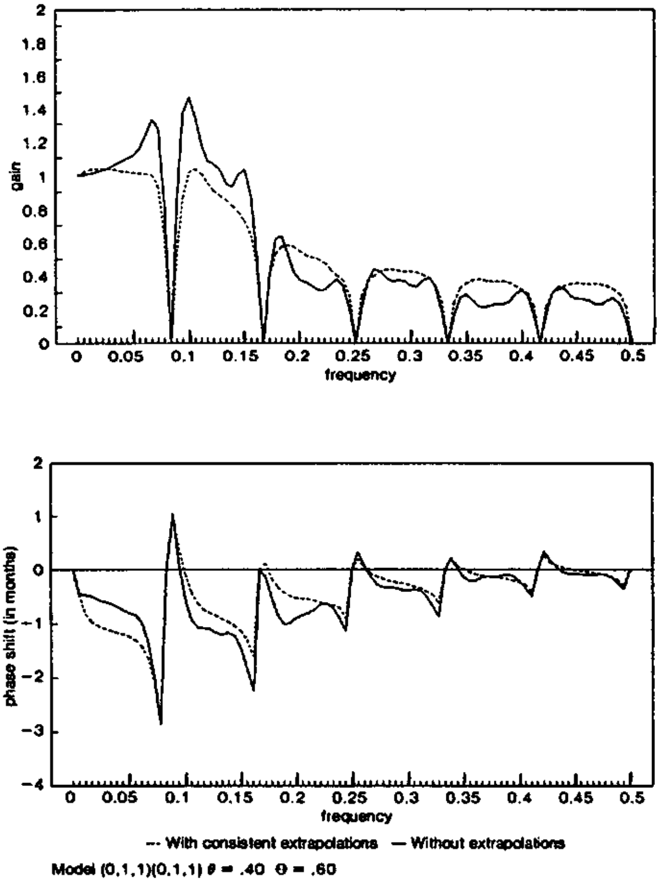

Figure 3. Trend-cycle concurrent cascade filters, (3 × 3) (3 × 5) [H − 13], with and without ARIMA extrapolations.

Figure 3 shows how the use of ARIMA extrapolations makes the gain of the concurrent cascade filters (using the 13-term Henderson) to resemble the symmetric one although at the expense of a small increase in phase shift. The extrapolations are from an ARIMA model (0,1,1) $(0,1,1)_s$ where the regular moving average parameter is $\theta = 0.40$ and the seasonal moving average parameter is $\Theta = 0.60$.

Although not shown for space reasons,the gain and phase shift of this trend-cycle concurrent filter fall between the other two combinations.

When ARIMA extrapolations are used, the gain of the concurrent filter converges very fast to that of the final. Dagum and Laniel (1987) show that after three more observations are added to the series, the gain of the asymmetric trend-cycle filter is very close to the symmetric one. The properties of these filters are also extensively discussed in Dagum, Chhab and Chiu (1993, 1996).

The presence of ripples in the final trend-cycle estimates will be produced by the 13-term Henderson filter only if some power is present in the input to the filter at the 0.08-0.16 frequency band. The input to the filter is the seasonally adjusted data with extreme values replaced.

In most empirical cases, the presence of unwanted ripples occurs in periods of high volatility when the observed data are mostly influenced by outliers which can be falsely interpreted as turning points. Although the seasonally adjusted series are modified by extreme values, there is a need for further smoothing which can be done either by applying a longer Henderson filter or by being stricter with the replacement of outliers. Since we want to keep the advantage of a short filter to detect turning points faster, the latter approach is the one followed here.

In the current procedure, the default sigma limits for the replacement of extreme values are ± 1.5 sigma and ± 2.5 sigma. Values greater than ± 2.5 sigma receive a zero weight and those smaller than ± 1.5 sigma a weight of one (full weight). Values falling within the boundaries are assigned a linearly graduated weight between zero and one.

3. A NEW METHOD

The new method here proposed, basically consists of: (1) extending a smoothed seasonally adjusted series (modified by extreme values with zero weight) with ARIMA extrapolations, and (2) applying the 13-term Henderson filter to the extended series using stricter sigma limits for the identification and replacement of extreme values.

Experimentation with real data showed that the power spectrum of the seasonally adjusted series at the 0.08-0.16 frequency band was drastically reduced only when strict sigma limits such as ±0.7 sigma and ±1.0 sigma were used. Hence, when applying the 13-term Henderson filter, the trend-cycle curve did not exhibit unwanted ripples while still maintaining its good property of rapid detection of turning points. Under the assumption of normality, these new sigma limits imply that 48% of the irregulars will be modified, 32% will get zero weight and will be replaced by the mean value and 16% will get graduated weights from zero to one.

The extension of the smoothed seasonally adjusted series with ARIMA extrapolations is needed to reduce the size of the revisions for the most recent estimates of the trend-cycle.

The implementation of this new procedure in the context of the X-11-ARIMA and X-11 methods must be done in two steps as follows:

(1) Produce the best seasonally adjusted series selecting appropriate options for the estimation of the components, that is, seasonality, trend-cycle, trading-day variations and Easter effects plus permanent or temporary priors, if applicable. The seasonally adjusted values are printed in Table D11. The seasonally adjusted series is modified by extreme values with zero weights using the default sigma limits and printed in Table E2. When the estimates of the published seasonally adjusted series for the current year are modified according to some revision practices, then this published revised series should be resubmitted to the X-11-ARIMA program to obtain the corresponding output shown in Table E2.

(2) The output from Table E2 is extended with one year of extrapolations from an ARIMA model. The ARIMA model found adequate with many real series is the (0,1,1) (0,0,1) model. Although the output from Table E2 does not contain seasonality, the seasonal moving average parameter (often of very small value) is needed to correct for some sort of seasonal autocorrelation in the data. The extended series is then run with the X-11-ARIMA program using the Summary Measures option and requesting strict sigma limits ($\pm 0.7\sigma$ and $\pm 1.0\sigma$) and the 13-term Henderson filter. The new trend-cycle estimates are printed in Table D12.

4. EMPIRICAL RESULTS

The new method for trend-cycle estimation is tested with nine leading indicator series of the Canadian Composite Leading Index. In the so called "filtered" version of the Canadian Composite Leading Index published by Statistics Canada, each of the components series as well as the Index itself are smoothed applying to the seasonally adjusted data asymmetric filters based on ARMA models developed by Rhoades (1980). The spectral properties of these ARMA trend-cycle filters are similar to those of the end point of the 9- 13- and 23-term Henderson filters depending on the ARMA model chosen (see Cholette 1982). (Although a comparison with the ARMA filters is not done in this paper, it is likely that the new approach will also give improved results.) Most of the series are highly volatile and all lead at turning points in the business cycle. The series are:

TSE300 Stock Price Index (TSE300)
House Spending Index (HSI)
Money Supply (M1)
Business and Personal Services Employment (BPSE)
Average Workweek in Manufacturing (AWM)
Retail Sales of Furniture and Appliances (RSFA)
Retail Sales of Durable Goods (RSDG)
New Orders for Durable Goods (NODG)
Shipments to Inventories Ratio (SIR).

The advantages of the new procedure versus the currently available in X-11-ARIMA are evaluated as follows.

4.1 Reduction of Ripples in the Final Trend-Cycle Estimates

To calculate the reduction of ripples we first introduce the definition of a turning point within the context of trend-cycle data. A turning point is generally defined as a point in time t when a series, say Y_t is larger (smaller) than or equal to the preceding k and subsequent m observations of the series. That is,

$$Y_{t-k} \leq \ldots \leq Y_{t-1} > Y_t \geq Y_{t+1} \geq \ldots \geq Y_{t+m}$$

defines a downturn and

$$Y_{t-k} \geq \ldots \geq Y_{t-1} < Y_t \leq Y_{t+1} \leq \ldots \leq Y_{t+m}$$

defines an upturn.

From the viewpoint of seasonally adjusted series and trend-cycle data, there is no general consensus for what values of k and m, a turning point has occurred. Rhoades (1980) defines a turning point for $k = 1$ and $m = 0$; Wecker (1979) defines a turning point to be the second of two (or more) successive declines or increases, *i.e.*, for $k = 2$ and $m = 2$; Zellner, Hong and Min (1991), LeSage (1991) and Pfeffermann and Bleuer (1992) have chosen $k = 3$ and $m = 0$. These definitions do not necessarily correspond to those of cyclical turning points for business cycle analysis but any one can be useful to calculate the number of unwanted ripples as long as two turning points (a downturn and an upturn) occur within a period of ten months or less. We use here the turning point definition for which $k = 3$ and $m = 0$ given the smoothness of the trend-cycle data.

Table 1 shows the number of ripples present in the trend cycle estimates from the standard and the modified 13-term Henderson filter for the period January 1981-December 1993.

Table 1

Number of Unwanted Ripples in the Trend-Cycle Data Using the 13-Term Henderson Filter for the Period 1981-1993

Series	Standard Procedure	Modified Procedure
NODG	9	2
HSI	8	4
RSDG	8	4
BPSE	8	5
AWM	7	1
SIR	5	1
TS300	4	2
M1	4	2
RSFA	4	0

The results show that the reduction is larger for those series with a large number of ripples and significant in all cases.

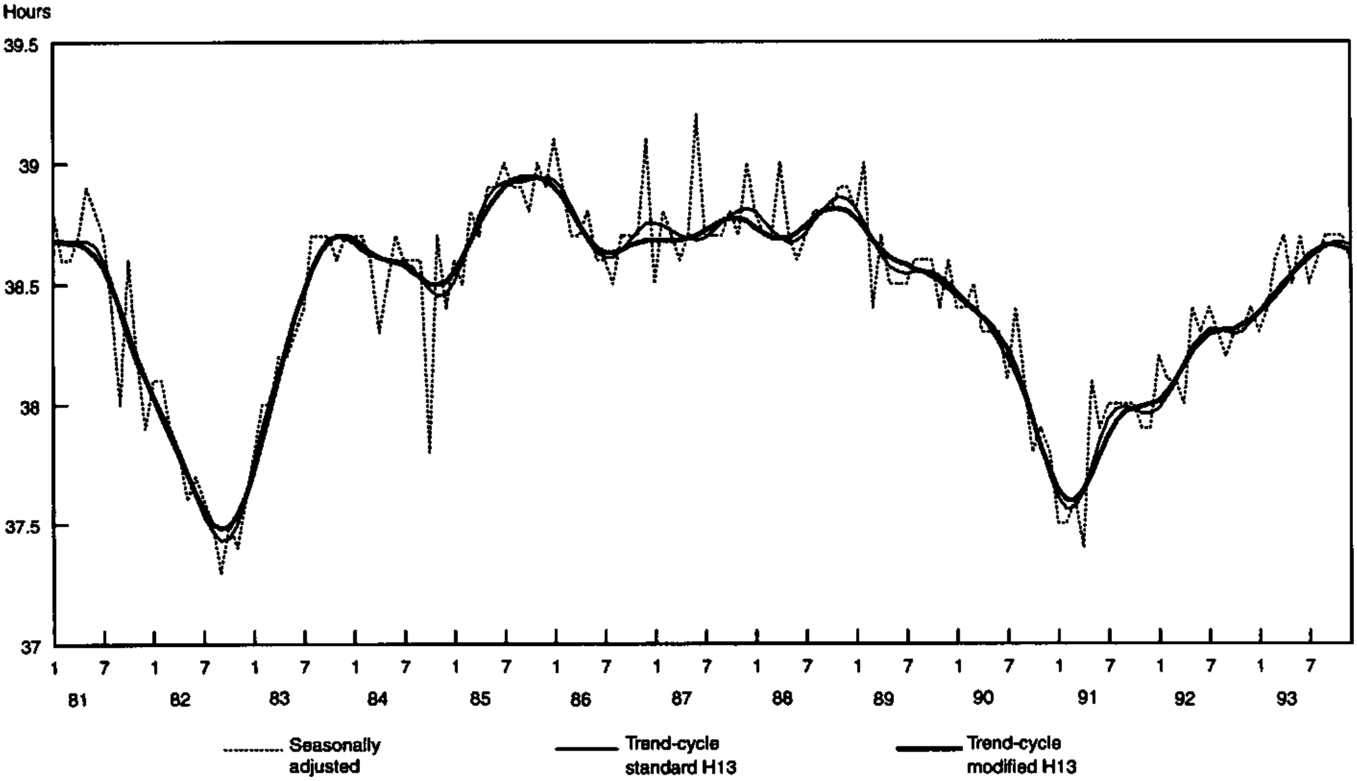

Figure 4. Average work week manufacturing.

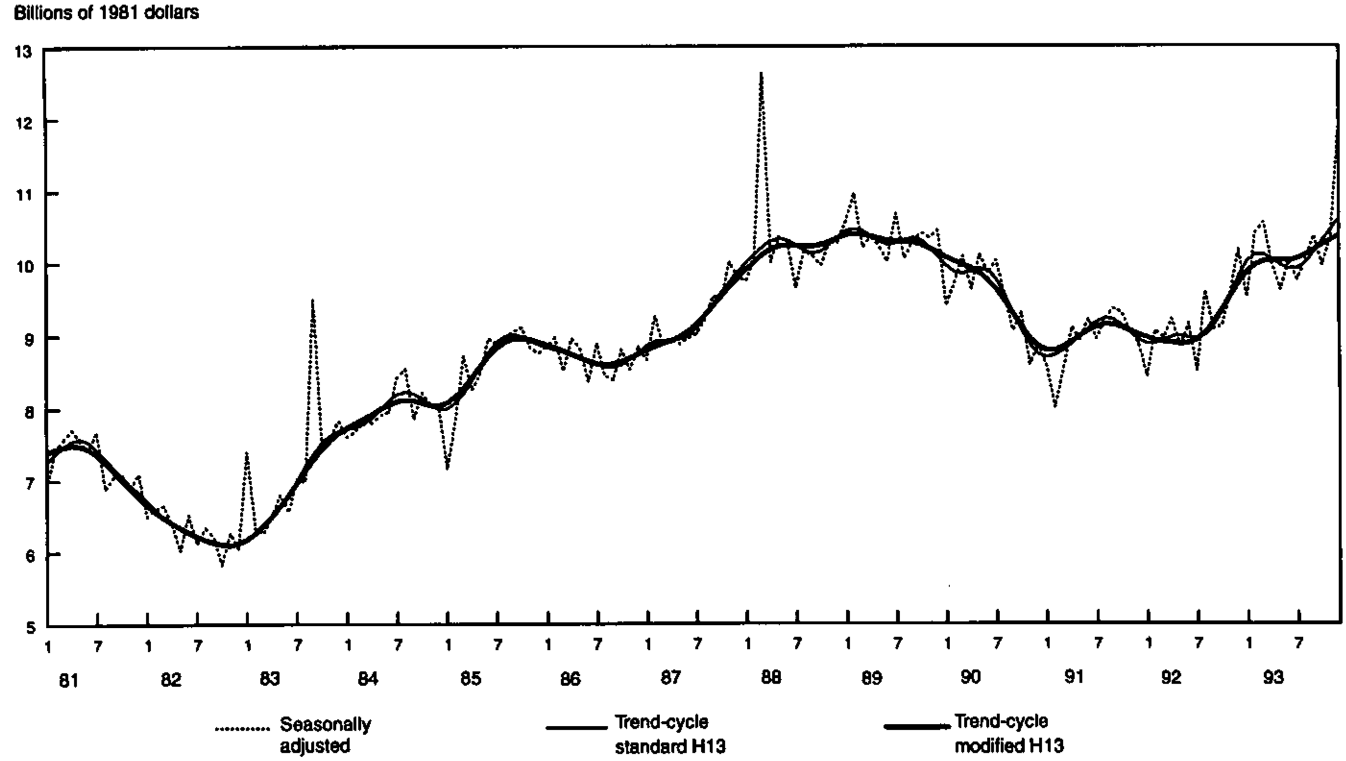

Figure 5. New orders for durable goods.

For illustrative purposes, Figures 4 and 5 for AWM and NODG respectively, exhibit the seasonally adjusted values and the trend-cycle data of both the standard and modified procedures. It is apparent that the new method reduces the ripples in the trend-cycle data with respect to those shown by the standard procedure. In fact, the modified trend-cycle data resembles that of the 23-term Henderson filter but with larger penetration into peaks and troughs of cycles of long duration.

4.2 Turning Point Detection

It is important that the reduction of ripples in the final estimates of the trend-cycle is not achieved at the expense of increasing the lag in detecting turning points which is the main limitation of the 23-term Henderson filter.

To study the revision path of the trend-cycle for any given point in time, the estimates were computed for all end points and previous time points. The revision path of the modified trend-cycle values showed that the identification of cyclical turning points is done with an average lag similar to the standard approach. Depending on the series, the lag was either equal or plus minus one month. For illustrative purposes, Figures 6a. exhibits the revision path of the modified trend-cycle values of New orders for durable goods for the cyclical turning point of February 1991. Successive updates are carried out using data up to March 1991, April 1991 and so on. The turning point is recognized in April, after 2 months whereas it takes

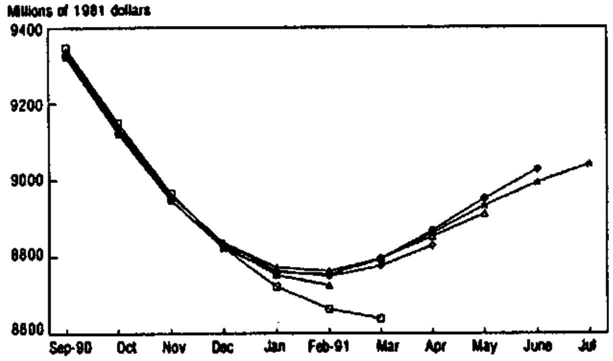

Figure 6a. New orders for durable goods. Trend-cycle modified H13 revisions path.

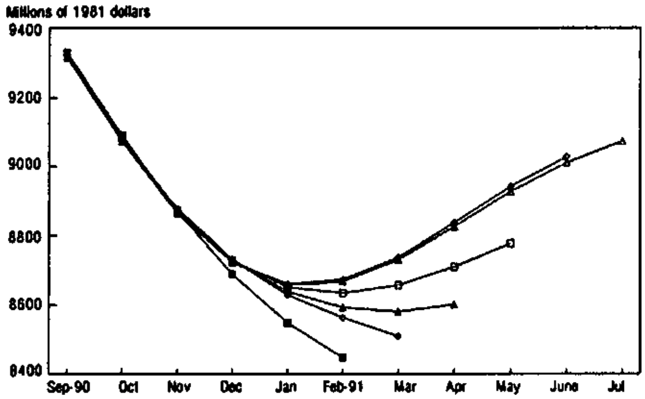

Figure 6b. New orders for durable goods. Trend-cycle standard H13 revisions path.

3 months for the standard procedure as exhibited in Figure 6b. Furthermore, it is shown that successive revisions of the trend-cycle estimates keep generally very close to the final values. The lines which protude, indicating a large revision, can be explained in terms of the underlying data which seem to indicate an increasing decline contradicted by the following values.

Figures 7a. and 7b. for the Average work week in manufacturing reveal that the turning point February-March 1991 is detected three months later by both procedures.

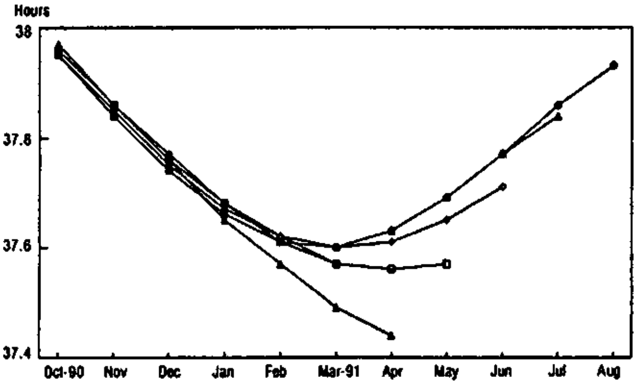

Figure 7a. Average work week manufacturing. Trend-cycle modified H13 revisions path.

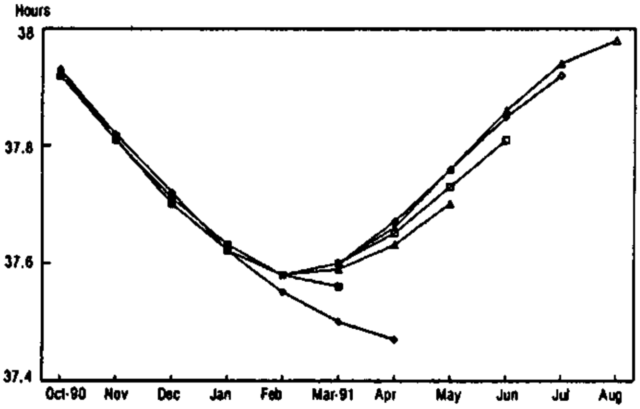

Figure 7b. Average work week manufacturing. Trend-cycle standard H13 revisions path.

4.3 Reduction of Revisions of Concurrent Trend-Cycle Estimates

Another important aspect to take into consideration is to reduce the total revision of the most recent estimate of the trend-cycle which is of preliminary character. Theoretically, the final trend-cycle value is obtained after the series is extended with four years of data but the size of the revisions is negligible after three more months.

Table 2 shows the mean absolute percent revision of the concurrent trend-cycle estimates over a four year period from January 1988 untill December 1991. The results indicate that for six of the nine cases analyzed the total revisions of the concurrent trend-cycle values using the modified procedure are much smaller compared to the standard, only for two series they are slightly larger.

Table 2

Mean Absolute Percent Total Revision of Concurrent Trend-Cycle Values Using the 13-Term Henderson Filter

Series	Standard Procedure (1)	Modified Procedure (2)	Ratio (2)/(1)
NODG	1.55	1.10	0.73
RSFA	0.62	0.47	0.76
RSDG	0.77	0.62	0.80
SIR	0.87	0.70	0.80
AWM	0.13	0.12	0.92
TS300	1.12	1.07	0.95
M1	0.35	0.35	1.00
HSI	2.09	2.20	1.05
BPSE	0.40	0.42	1.05

5. CONCLUSION

This paper introduced a new method for trend-cycle estimation which enables the use of the 13-term Henderson filter with the advantages of: (i) reducing the number of unwanted ripples in the final trend-cycle curves, (ii) reducing the size of the revisions to preliminary concurrent values, and (iii) not increase the time lag in turning point detection.

The new method basically consists of extending a smoothed seasonally adjusted series (modified by extreme values with zero weight) with one year of ARIMA extrapolations, and then applying the 13-term Henderson filter using strict sigma limits for the identification and replacement of outliers.

The procedure is illustrated with nine leading indicator series of the Canadian Composite Leading Index and the results are highly satisfactory.

ACKNOWLEDGEMENTS

This article is based on work supported by Statistics Canada. Views expressed are those of the author and do not necessarily reflect those of Statistics Canada. I am indebted to Marietta Morry and Norma Chhab of the Time Series Research Centre for many stimulating discussions and their collaboration in the work that led to this paper. I am also thankful to an anonymous referee for helpful comments on an earlier version.

REFERENCES

CASTLES, I. (1987). A Guide to Smoothing Time Series Estimates of Trend. Catalogue No. 1316.0, Australian Bureau of Statistics.

CHOLETTE, P.A. (1981). A comparison of various trend-cycle estimators. In *Time Series Analysis*. (O.D. Anderson and M.R. Perryman, Eds). Amsterdam: North-Holland, 77-87.

CHOLETTE, P.A. (1982). Comparaison de deux estimateurs des cycles économiques. Research Paper No. 82-09-OO1F, Time Series Research and Analysis Centre, Statistics Canada.

CLEVELAND, R., CLEVELAND, W.S., McRAE, J.E., and TERPENNING, I. (1990). STL: A seasonal-trend decomposition procedure based on Loess. *Journal of Official Statistics*, 6, 3-33.

DAGUM, E.B. (1980). The X-11-ARIMA Seasonal Adjustment Method. Catalogue No. 12-564E. Statistics Canada.

DAGUM, E.B. (1988). The X-11-ARIMA/88 Seasonal Adjustment Method - Foundations and User's Manual. Time Series Research and Analysis Centre, Statistics Canada.

DAGUM, E.B., and LANIEL, N. (1987). Revisions of trend-cycle estimators of moving average seasonal adjustment methods. *Journal of Business and Economic Statistics*, 5, 177-189.

DAGUM, E.B., CHHAB, N., and CHIU, K. (1993). Linear properties of the X-11-ARIMA seasonal adjustment method. *Proceedings of the Business and Economic Statistics Section, American Statistical Association.*

DAGUM, E.B., CHHAB, N., and CHIU, K. (1996). Derivation and properties of the Census X-11 variant and the X-11-ARIMA linear filters. *Journal of Official Statistics*, (forthcoming).

FINDLEY, D.F., and MONSELL, B.C. (1990). Comment (on Cleveland *et al.* 1990). *Journal of Official Statistics*, 6, 55-59.

GRAY, A.G., and THOMSON, P.J. (1990). Comment (on Cleveland *et al.* 1990). *Journal of Official Statistics*, 6, 47-54.

HENDERSON, R. (1916). Note on graduation by adjusted average. *Transactions of the Actuarial Society of America*, 17, 43-48.

KENNY, P. (1993). Trend presentation. T02919, SMQ, Branch, Central Statistical Office, London, England.

KENNY, P.B., and DURBIN, J. (1982). Local trend estimation and seasonal adjustment of economic and social time series. *Journal of the Royal Statistical Society*, Series A, 145, 1-41.

LeSAGE, J.P. (1991). Analysis and development of leading indicators using a Bayesian turning-points approach. *Journal of Business and Economic Statistics*, 9, 305-316.

PFEFFERMANN, D., and BLEUER, S.R. (1992). Probabilistic detection of nonseasonal turning points in economic time series estimated from sample surveys. Internal report, Methodology Branch, Statistics Canada, Ottawa.

RHOADES, D. (1980). Converting timeliness into reliability in economic time series or minimum phase shift filtering of economic time series. *Canadian Statistical Review*, 6-13.

SCOTT, S. (1990). Comment (on Cleveland *et al.* 1990). *Journal of Official Statistics*, 6, 59-62.

SHISKIN, J., YOUNG, A.H., and MUSGRAVE, J.C. (1967). The X-11 Variant of Census Method II Seasonal Adjustment. Technical Paper No. 15, U.S. Bureau of the Census.

WALLGREN, B., and WALLGREN, A. (1990). Comment (on Cleveland *et al.* 1990). *Journal of Official Statistics*, 6, 39-46.

WECKER, W. (1979). Predicting the turning points of a series. *Journal of Business*, 52, 35-50.

ZELLNER, A., HONG, C., and MIN, C. (1991). Forecasting turning points in international output growth rates using Bayesian exponentially weighted autoregression, time-varying parameter, and pooling techniques. *Journal of Econometrics*, 48, 275-304.

Survey Methodology, June 1996
Vol. 22, No. 1, pp. 85–94
Statistics Canada

A Moving Stratification Algorithm

YVES TILLÉ[1]

ABSTRACT

A general algorithm with equal probabilities is presented. The author provides the second order inclusion probabilities that correspond to the algorithm, which generalizes the selection-rejection method, so that a sample may be drawn using simple random sampling without replacement. Another particular case of the algorithm, called moving stratification algorithm, is discussed. A smooth stratification effect can be obtained by using, as a stratification variable, the serial number of the observation units. The author provides approximations of first and second order inclusion probabilities. These approximations lead to a population mean estimator and to an estimator of the variance of this mean estimator. The algorithm is then compared to a classical stratified plan with proportional allocation.

KEY WORDS: Selection algorithm; Equal probability sampling; Strata.

1. INTRODUCTION

When a file is ordered according to an auxiliary variable that is close to the variable of interest, how can a sample be selected using such information? One solution to the problem consists of making a stratified selection. However, making such a selection requires that a delicate problem be resolved, namely subdividing the population into strata. Another simple solution that is both quick and efficient consists of making a systematic selection. The algorithm can be written in a few lines. Moreover, the way in which the file is ordered can be put to good use. However, a systematic selection has one major flaw, namely that estimating the variance of total or mean estimators requires one or several hypotheses concerning the population. It will be shown that there is another simple selection algorithm with which a sample can be drawn in one pass using the file ordering system. For this algorithm, an estimator of the variance of a total or mean estimator is provided, requiring no modelling of the population.

A general selection algorithm providing equal first order inclusion probabilities is presented in section 2. First and second order inclusion probabilities are provided. In section 3, the proposed algorithm is shown to generalize the selection-rejection method so that a simple random sample can be drawn without replacement along with the stratified plan with proportional allocation. Finally, in section 4, the moving stratum method is defined and, in section 5, conclusions are drawn.

2. PRESENTATION OF THE GENERAL ALGORITHM

2.1 The Algorithm

Let us consider a finite population $U = \{1, \ldots, i, \ldots, N\}$; we write $y_1, \ldots, y_i, \ldots, y_N$, the N values assumed by variable y for N observation units of U. The mean of the values assumed by variable y for the population is written as

$$\bar{y} = \frac{1}{N} \sum_{i \in U} y_i.$$

A random sample s of fixed size n is drawn from this population. The random variables indicating the presence of observation units in s are written as I_i, $i \in U$. The first order inclusion probability is written as $\pi_i = \Pr(i \in s) = E(I_i)$, $i \in U$ and the second order inclusion probability as $\pi_{ik} = E(I_i I_k)$, $i \neq k \in U$. The algorithm is very short. It resembles the algorithms of Fan, Fuller and Rezucha (1962), Bebbington (1975), McLeod and Bellhouse (1983) and Sunter (1977, 1986). Only N, n and the b_i, $i = 0, \ldots, N - 1$ need to be known. The other variables are working variables.

General Algorithm

```
$j <= 0;$
$i <= 0;$
Repeat for $i = 0, \ldots, N - 1$
    $u <=$ a random number with a uniform distribution $[0,1]$;
        if $\frac{(b_i + i)n/N - j}{b_i} > u$ then
            select record $i + 1$;
            $j <= j + 1$;
        otherwise, pass the record $i + 1$;
    $i <= i + 1$.
```

At each step, j represents the number of records already selected and i the number of records passed (selected or not). For each iteration, a decision is made about selecting the record $i + 1$. If the record is selected, it becomes the $(j + 1)$-th in the sample. The coefficients b_i, $i = 0, \ldots, N - 1$, are strictly positive real numbers. These

[1] Yves Tillé, Laboratoire de Méthodologie du Traitement des Données, C.P. 124, Université Libre de Bruxelles, avenue Jeanne, 44, 1050 Bruxelles, Belgique, E-mail ytilleb@ulb.ac.be

quantities must meet certain conditions discussed below if the plan is to be of fixed size or if the units are to be selected with equal probability. The choice of different values for b_i, $i = 0, \ldots, N-1$, will make it possible to generate several special cases of the general algorithm.

If b_i are strictly positive reals such that $b_i \leq N - i$, then the sample size is equal to or smaller than n. In fact, assuming we have already drawn n units from the population at step i and that $b_i \leq N - i$, then

$$\frac{(b_i + i)n/N - n}{b_i} = \frac{n}{N} - \frac{n}{b_i}\frac{N-i}{N} \leq \frac{n}{N} - \frac{n}{N-i}\frac{N-i}{N} = 0.$$

It becomes impossible to draw a further unit. It will be assumed in everything that follows that $b_i \leq N - i$. Moreover, if $b_i \leq N - i$, $i = 1, \ldots, N - n - 1$ and if $b_i = N - i$, $i = N - n, \ldots, N - 1$, the sample is of fixed size n. Note that these conditions for obtaining a sample of fixed size are sufficient but not necessary.

Three particular cases of the algorithm are examined below. These three cases are defined by three choices of coefficient b_i, $i = 0, \ldots, N-1$. Before examining these particular choices, we will determine the first and second order inclusion probabilities without loss of generality.

2.2 First Order Inclusion Probabilities

We write n_i, the number of units selected after passing i records. We see immediately that $n_1, \ldots, n_i, \ldots, n_N$ is a Markov chain. In fact, we directly derive from the algorithm that

$$\Pr[n_i = j \mid n_1, \ldots, n_{i-1}] = \Pr[n_i = j \mid n_{i-1}].$$

The random variables

$$c_i = \frac{(b_i + i)n/N - n_i}{b_i}, \; i = 0, \ldots, N-1,$$

can sometimes assume values greater than 1 or less than 0. Since $\max(0, n - N + i) \leq n_i \leq \min(i, n)$, then $\Pr[0 \leq c_i \leq 1] = 1$ if

$$b_i \geq \begin{cases} \min\left(i\dfrac{N-n}{n}, N - i\right) & \text{if } n \leq N/2 \\ \min\left(i\dfrac{n}{N-n}, N - i\right) & \text{if } n > N/2 \end{cases}, \quad i = 0, \ldots, N-1. \quad (1)$$

Again conditions (1) are sufficient but not necessary. We can therefore construct b_i which do not meet these conditions but which provide c_i in $[0,1]$. The case dealt with in section 3.2 (stratification) represents one example.

The following example also provides c_i in $[0,1]$ without meeting condition (1): let us consider $N = 12$, $n = 4$ and $b_0 = b_1 = b_3 = b_4 = b_6 = 6$, $b_2 = b_5 = 7$, $b_i = N - i, i = 12 - i, i = 7, \ldots, 11$. We have $c_0 = 1/3$, $c_1 = (7 - 3n_1)/18$, $c_2 = (3 - n_2)/7$, $c_3 = (3 - n_3)/6$, $c_4 = (10 - 3n_4)/18$, $c_5 = (4 - n_5)/7$, $c_6 = (4 - n_6)/6$, $c_7 = (4 - n_7)/5$, $c_8 = (4 - n_8)/4$, $c_9 = (4 - n_9)/3$, $c_{10} = (4 - n_{10})/2$, $c_{11} = (4 - n_{11})$. We note that $n_1 \leq 1$, $n_2 \leq 2$, $n_3 \leq 3$. If $n_3 = 3$ then $c_3 = 0$ and therefore $n_4 \leq 3$. We then have $n_5 \leq 4$ and if $n_5 = 4$ then $c_5 = 0$ and therefore $n_6 \leq 4$. This last comment is true for all c_i that follow. We therefore note that all c_i are in $[0,1]$ whereas $b_4 = 6$ does not meet condition (1).

In order to simplify the demonstrations which follow, it will be assumed that

$$\Pr[0 \leq c_i \leq 1] = 1, \; i = 0, \ldots, N-1.$$

We will return to the problem of c_i values greater than 1 or smaller than 0 later on. If

$$\Pr[0 \leq c_i \leq 1] = 1, \; i = 0, \ldots, N-1,$$

we have

$$E[I_{i+1} \mid n_1, \ldots, n_i] = E[I_{i+1} \mid n_i] = \frac{(b_i + i)n/N - n_i}{b_i}.$$

It can be shown easily by recursion that if $\Pr[0 \leq c_i \leq 1] = 1$, $i = 0, \ldots, N-1$, $E[n_i] = i\,n/N$, $i = 0, \ldots, N$. Therefore,

$$\pi_i = E[I_i] = E[n_i] - E[n_{i-1}] = \frac{n}{N}. \quad (2)$$

2.3 Second Order Inclusion Probabilities

Four results provided by lemmas 1, 2 and 3 are needed in order to determine second order inclusion probabilities.

Lemma 1 If $\Pr[0 \leq c_i \leq 1] = 1$, $i = 0, \ldots, N-1$, then

$$E[n_{i+k} \mid n_i] = (i + k)\frac{n}{N} + \left(n_i - i\frac{n}{N}\right)\prod_{\ell=i}^{i+k-1}\frac{b_\ell - 1}{b_\ell},$$

$$i = 1, \ldots, N-1, k = 1, \ldots, N-i.$$

This lemma can be demonstrated by recursion if it is assumed to be true for $k - 1$. Using lemma 1, the following lemma is readily obtained by subtraction:

Lemma 2 If $\Pr[0 \le c_i \le 1] = 1,\ i = 0, \ldots, N-1$, then

$$E[I_{i+k} \mid n_i]$$

$$= \frac{n}{N} - \left(n_i - i\frac{n}{N}\right)\frac{1}{b_{i+k-1}}\prod_{\ell=i}^{i+k-2}\frac{b_\ell - 1}{b_\ell},$$

$$i = 1, \ldots, N-1,\ k = 1, \ldots, N-i.$$

It is assumed by convention that an empty product has a value of 1.

Lemma 3 If $\Pr[0 \le c_i \le 1] = 1,\ i = 0, \ldots, N-1$, then

$$\mathrm{Var}[n_i] = \frac{n}{N}\frac{N-n}{N}\sum_{j=1}^{i}\prod_{\ell=j}^{i-1}\frac{b_\ell - 2}{b_\ell}, i = 1, \ldots, N. \quad (3)$$

The demonstration is provided in the appendix.

Finally, the second order inclusion probability is provided by the following proposition:

Proposition 1 If $\Pr[0 \le c_i \le 1] = 1,\ i = 0, \ldots, N-1$, then

$$E[I_{i+k}I_{i+1}]$$

$$= \frac{n^2}{N^2} - \frac{n}{N}\frac{N-n}{N}\frac{1}{b_{i+k-1}}$$

$$\times\left(1 - \frac{1}{b_i}\sum_{j=1}^{i}\prod_{\ell=j}^{i-1}\frac{b_\ell - 2}{b_\ell}\right)\prod_{\ell=i+1}^{i+k-2}\frac{b_\ell - 1}{b_\ell},$$

$$i = 0, \ldots, N-2,\ k = 2, \ldots, N-i. \quad (4)$$

The demonstration is provided in the appendix.

Corollary 1 If $\Pr[0 \le c_i \le 1] = 1,\ i = 0, \ldots, N-1$, then

$$\pi_{ik} = \frac{n^2}{N^2} - \frac{n}{N}\frac{N-n}{N}\left(1 - \frac{1}{b_{i-1}}\sum_{j=1}^{i-1}\prod_{\ell=j}^{i-2}\frac{b_\ell - 2}{b_\ell}\right)$$

$$\times\frac{1}{b_{k-1}}\prod_{\ell=i}^{k-2}\frac{b_\ell - 1}{b_\ell},\ i = 1, \ldots, N-1, k > i.$$

2.4 The Horvitz-Thompson Estimator and its Variance

The Horvitz-Thompson estimator is the simple sample mean since the first order inclusion probabilities are all equal

$$\hat{\bar{y}}_\pi = \frac{1}{n}\sum_{i\in s} y_i.$$

If the design is of fixed size, we can use the Yates and Grundy variance formula (1953)

$$\mathrm{Var}[\hat{\bar{y}}_\pi] = \frac{1}{2N^2}\sum_{i\in U}\sum_{\substack{k\in U\\k\neq i}}\left(\frac{y_i}{\pi_i} - \frac{y_k}{\pi_k}\right)^2(\pi_i\pi_k - \pi_{ik}). \quad (5)$$

Since $\pi_i = n/N,\ i = 1, \ldots, N$ and assuming that

$$\gamma_{ik} = 1 - \pi_{ik}\frac{N^2}{n^2},$$

we can write

$$\mathrm{Var}[\hat{\bar{y}}_\pi] = \frac{1}{N^2}\sum_{i\in U}\sum_{\substack{k\in U\\k\neq i}}(y_i - y_k)^2\gamma_{ik}. \quad (6)$$

The variance estimator is provided by

$$\widehat{\mathrm{Var}}[\hat{\bar{y}}_\pi] = \frac{1}{2N^2}\sum_{i\in s}\sum_{\substack{k\in s\\k\neq i}}\left(\frac{y_i}{\pi_i} - \frac{y_k}{\pi_k}\right)^2\frac{\pi_i\pi_k - \pi_{ik}}{\pi_{ik}}. \quad (7)$$

This can be written here as

$$\widehat{\mathrm{Var}}[\hat{\bar{y}}_\pi] = \frac{1}{2n^2}\sum_{i\in s}\sum_{\substack{k\in s\\k\neq i}}(y_i - y_k)^2\frac{\gamma_{ik}}{1 - \gamma_{ik}}.$$

3. APPLICATION 1: SIMPLE AND STRATIFIED RANDOM SELECTIONS

3.1 Simple Design

The simplest selection algorithm, the selection-rejection method described in Fan, Fuller and Rezucha (1962, method 1), Beddington (1975) and Deville and Grosbras (1987, p. 210), is of course a particular case of the general algorithm. We need only take

$$b_i = N - i,\quad i = 0, \ldots, N-1.$$

We always have $0 \le c_i \le 1$. The first order inclusion probabilities always have a value of n/N. Calculations for second order inclusion probabilities follow from proposition 1. Assuming $k > i$, on the basis of corollary 1, we can find the second order inclusion probabilities of the simple design:

$$\pi_{ik} = \frac{n(n-1)}{N(N-1)}.$$

We also recall some classical results concerning the simple design that we will be using later on. The estimator for $\bar{y}$ is therefore the mean of the sample

$$\hat{\bar{y}}_{srs} = \frac{1}{n} \sum_{i \in s} y_i. \tag{8}$$

The variance of this estimator is provided by

$$\mathrm{Var}[\hat{\bar{y}}_{srs}] = \frac{\sigma_y^2}{n} \frac{N - n}{N - 1} \tag{9}$$

where

$$\sigma_y^2 = \frac{1}{N} \sum_{i \in U} (y_i - \bar{y})^2. \tag{10}$$

An unbiased estimate of this variance is

$$\widehat{\mathrm{Var}}[\hat{\bar{y}}_{srs}] = \frac{s_y^2}{n} \frac{N - n}{N} \tag{11}$$

where

$$s_y^2 = \frac{1}{n - 1} \sum_{i \in s} (y_i - \hat{\bar{y}}_{srs})^2. \tag{12}$$

3.2 Stratified design

The stratified design can also be defined using the general algorithm. The stratification variable in this case is the serial number of the individual. Let us consider the particular case of a stratified design of H strata with proportional allocation where all the strata are of the same size. The strata are such that the individuals of a given stratum are adjacent in the data file. It is also assumed that N/H is an integer. This stratified design is obtained by simply taking

$$b_i = \left\{(N - i - 1) \bmod \frac{N}{H}\right\} + 1,\ i = 0, \ldots, N - 1.$$

4. APPLICATION 2: MOVING STRATIFICATION

4.1 The Problem

The file is assumed to be ordered according to an auxiliary variable that is close to the variable of interest. The problem is as follows: how can we draw a random selection that yields a small variance for the Horvitz-Thompson estimator of a mean? Looking at the formulation of the Yates-Grundy variance (5), we see that there are two distinct answers to this question.

The first solution consists of selecting with unequal probabilities using first order inclusion probabilities that are proportional to the variable of interest. If such a selection could be made, all quantities

$$\left(\frac{y_i}{\pi_i} - \frac{y_k}{\pi_k}\right)^2$$

would be zero and therefore the variance would be zero.

The second solution consists of using second order inclusion probabilities. A good selection could be one where π_{ik} are close to $\pi_i\pi_k$ if y_i is very different from y_k. On the other hand, if y_i is very close to y_k, we can select second order inclusion probabilities π_{ik} that are clearly smaller than $\pi_i\pi_k$. Thus, where quantities

$$\left(\frac{y_i}{\pi_i} - \frac{y_k}{\pi_k}\right)^2$$

would be large (respectively small), quantities $\pi_i\pi_k - \pi_{ik}$ would be small (respectively large). We would thus have a small variance.

The second solution we have just described is in fact often used. It is the basic idea for stratification. Our objective is to apply this idea to the construction of a sequential selection algorithm that is easy to implement. Such an algorithm could be applied to any file without the need to know anything save the size of the population. It would therefore apply to very large files. We could thus benefit from the information provided by this auxiliary variable like for stratification, without the need to actually subdivide into strata.

4.2 The Method

We first define M the length of the moving stratum within the population. M represents, in a way, the size of the stratum within the population and is such that $N/n \leq M \leq N$. The algorithm of the moving stratum is defined by

$$b_i = \min(M, N - i),\ i = 0, \ldots, N - 1.$$

There is, however, one problem. Quantities c_i defined by

$$c_i = \begin{cases} \dfrac{(M + i)n/N - n_i}{M} & \text{if} \quad i \leq N - M \\ \dfrac{n - n_i}{N - i} & \text{otherwise,} \end{cases}$$

are not always in $[0,1]$.

In fact, let us assume that, before the $(N - M)$-th step of the algorithm, c_i is positive and very close to zero and that through some bad luck the unit i is nevertheless chosen. In such a case, c_{i+1} would have a value of $c_i - (N-n)/(NM)$. c_{i+1} can therefore have a negative value but this negative value is always greater than $-(N-n)/(NM)$. In fact, if one of the c_i is already negative, the unit i is not selected and therefore c_{i+1} has a value greater than c_i.

Let us now assume that before the $(N - M)$-th step of the algorithm, one c_i is very slightly smaller than 1 and that nevertheless unit i is not selected. In such a case, c_{i+1} would have a value of $c_i + n/(NM)$. c_{i+1} can therefore take on a value greater than 1 but this value greater than 1 is nevertheless always smaller than $1 + n/(NM)$. In fact, if one of the c_i is already greater than 1, the unit i is always selected and therefore c_{i+1} has a value smaller than c_i.

We obtain

$$\Pr\left[-\frac{N-n}{NM} < c_i < 1 + \frac{n}{NM}\right] = 1, i = 0, \ldots, N - M. \tag{13}$$

The design is however of fixed size, a result that follows the following proposition:

Proposition 2 If $b_i = \min(M, N - i)$, $(N/n < M < N)$, $0 = 1, \ldots, N - 1$, then the design is of fixed size.

The demonstration is provided in the appendix.

Since the c_i are not always within the interval $[0,1]$, we carried out 50 simulations of the moving stratum algorithm for various sample and population sizes. The selected N population sizes were 100, 500, 2500, 12500, 62500, 312500. The reciprocals of sampling rates (N/n) were 2, 4, 8, 16, 32, 64, 128, 256, 512, 1024, 2048, 4096. We carried out several simulations by varying the size of the moving stratum as follows: $M = N/n, 2N/n, 3N/n, \ldots$. The simulations seem to indicate that the greater the value for M, the smaller the probability that a c_i will fall outside of $[0,1]$. As soon as $M \geq 10N/n$, for all the simulations that we carried out, the problem was no longer raised. This first result does not imply that the probability that at least one of the c_i will fall outside of $[0,1]$ is zero when $M \geq 10N/n$. However, it may be said that such a probability would then be very small.

4.3 Estimating the Mean and Bias

In examining the results yielded by expression (2) and proposition 1, we get, as a first approximation, a value of about $\pi_i \approx n/N$ for first order inclusion probabilities. This approximation of inclusion probabilities makes it possible to construct an estimator.

$$\hat{y}_{sm} = \frac{1}{n} \sum_{i \in s} y_i.$$

This estimator is slightly biased since the c_i are not all exactly within the interval $[0,1]$. This bias is

$$B[\hat{y}_{sm}] = \frac{1}{N} \sum_{i \in U} \alpha_i y_i$$

where $\alpha_i = \pi_i N/n - 1$. Since the design is of fixed size, $\sum_{i \in U} \alpha_i = 0$. We can therefore write the bias in the form of a covariance: $B[\hat{y}_{sm}] = \sigma_{y\alpha}$ where

$$\sigma_{y\alpha} = \frac{1}{N} \sum_{i \in U} \alpha_i (y_i - \bar{y}). \tag{14}$$

Since the absolute value of a covariance is always equal to or smaller than the product of the two standard deviations, we obtain an upper bound for the absolute value of the bias

$$| B[\hat{y}_{sm}] | \leq \sigma_y \sigma_\alpha$$

where σ_y is defined by (10) and

$$\sigma_\alpha^2 = \frac{1}{N} \sum_{i \in U} \alpha_i^2.$$

The variance of the estimator is of a magnitude that is comparable (for N and fixed n) to the variance of the estimator of the mean in the simple design without replacement. We can therefore write

$$| B[\hat{y}_{sm}] | \leq C_\alpha \sqrt{\text{Var}[\hat{y}_{srs}]}$$

where Var $[\hat{y}_{srs}]$ is defined by (9) and

$$C_\alpha = \sigma_\alpha \sqrt{\frac{n(N-1)}{(N-n)}}.$$

We will assume that the bias is negligible when the upper bound of the bias of the estimator $\hat{y}_{sm}$ is negligible with respect to $\text{Var}[\hat{y}_{srs}]^{1/2}$, *i.e.*, when C_α is small.

Recursively we can calculate the exact value of the $\Pr[n_i = j]$ since we have

$$\Pr[I_i = 1 \mid n_i] = \tilde{c}_i, i = 1, \ldots, N - M$$

where $\tilde{c}_i$ has a value of 0 if $c_i < 0$, c_i if $0 \leq c_i \leq 1$ and 1 if $c_i > 1$. From this result we can derive the exact value of first order inclusion probabilities.

We have calculated (Appendix, Table 1) the values of C_α for various sample and population (100 - 312500) sizes. The values of C_α are provided for sizes of moving strata M equal to N/n, $2N/n$, $3N/n$, $4N/n$ and $5N/n$. It can be seen that as soon as the value of the moving stratum is $2N/n$, C_α never exceeds 0.07. When $M = 3N/n$, the coefficient C_α is expressed in thousandths. According to Cochran (1977, pp. 13-14), the bias is then negligible. The table therefore shows that if $M \geq 3N/n$, the bias of the estimator will be negligible at least for the specified sample and population sizes.

However, these results do not imply that the bias of the estimator is large when M is very small (for example $M = N/n$). The C_α are bias upper bounds. From expression (14), we see that the bias will be all the greater as the variable of interest correlates with the exact inclusion probabilities. We have shown (Figure 1) the exact inclusion probabilities (y axis) for N individuals (x axis) obtained by using the moving stratification algorithm with the

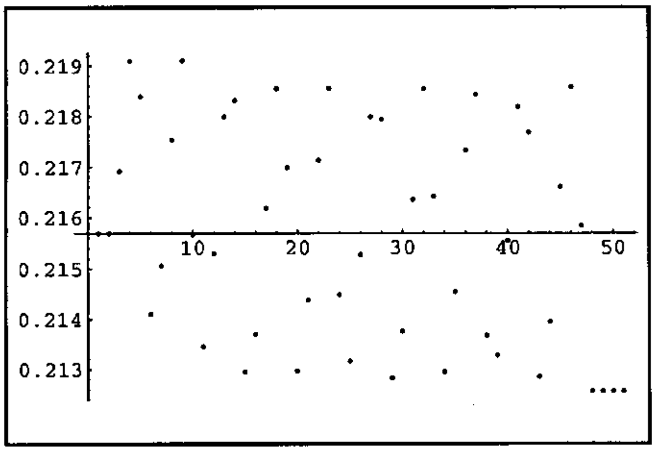

Figure 1. Inclusion probabilities.

parameters $N = 51$, $n = 11$, $M = N/n$. This case is obviously very unfavourable. The result is interesting. In this case, $n/N = 0.215686$. The inclusion probabilities are distributed on both sides of n/N with no marked tendency associated with the ordering of the file. In practical terms, the probability can be considered very small that there will be a variable of interest that strongly correlates with the exact inclusion probabilities; as a result, the bias will most often be clearly smaller than the given upper bound.

We could, of course, use the exact inclusion probabilities to establish an estimate. We feel that this is not worthwhile, for two reasons:

- first, because calculating the exact inclusion probabilities requires a significant amount of time,
- second, because the exact first order inclusion probabilities are such that

$$\mathrm{Var}\left[\sum_{i\in s}\frac{1}{\pi_i}\right] \neq 0.$$

In this case, we have a random Horvitz-Thompson estimator of a constant variable ($y_k = C$). To overcome this problem, an estimate of the mean is usually carried out using Hájek's (1971) ratio. This estimator is also biased.

4.4 Estimating the Variance of the Estimator

Assuming that $\Pr(0 \leq c_i \leq 1) \approx 1$, we can also build an approximation of second order inclusion probabilities using corollary 1. Given that b_i has a value of M if $i \leq N - M$ and $N - i$ otherwise, we obtain the following approximation:

$$\pi_{ik} \approx \frac{n^2}{N^2}(1 - \theta_{ik})$$

where

$$\theta_{ik} = \frac{N-n}{2n}\frac{1}{M-1}\left\{1 + \left(\frac{M-2}{M}\right)^{\min(i-1,N-M)}\right\}$$

$$\times \left(\frac{M-1}{M}\right)^{\max(0,\min(N-M-i+1,k-i))} \qquad k > i.$$

Assuming that the first order inclusion probabilities have a value of n/N, an approximation of the variance of $\hat{\bar{y}}_{sm}$ can be obtained:

$$\mathrm{Var}_{app}[\hat{\bar{y}}_{sm}] = \frac{1}{2N^2}\sum_{i\in U}\sum_{\substack{k\in U\\ k\neq i}}(y_i - y_k)^2\theta_{ik}. \qquad (15)$$

From (15), an estimator of the variance of the estimator of the mean can be obtained:

$$\widehat{\mathrm{Var}}_{app}[\hat{\bar{y}}_{sm}] = \frac{1}{2N^2}\sum_{i\in s}\sum_{\substack{k\in s\\ k\neq i}}(y_i - y_k)^2\frac{\theta_{ik}}{1-\theta_{ik}}. \qquad (16)$$

Again, this estimator is biased. In order to assess the magnitude of the bias, we carried out a series of simulations. The results are given in Table 2 in the appendix. We generated populations of size $N = 400$. The values assumed by the two variables x and y were generated by means of pseudo-random numbers having a bivariate normal distribution with a fixed coefficient of correlation ρ. The populations were then sorted in terms of the variable x. The objective was to estimate $\bar{y}$.

In these populations, samples of size 64 were selected using the moving stratum method (*sm*), a stratified design with proportional allocation in which the sizes of the strata were all equal (*strat*), as well as a simple design without replacement (*srs*). These three methods are particular cases of the general algorithm and they were implemented using the same random numbers. Simulations were carried out for different values of the moving stratum M (case: *sm*) and for different numbers of strata H (case: *strat*). An explanation is provided below for the choices of M and H. For each simulation, 200,000 samples were selected.

For each of the simulations, three results are given:

- The means for the simulations of the estimators of the variance of the estimator of the mean, which are expressed as $E_{sim}\widehat{\mathrm{Var}}(\hat{\bar{y}})$. These variance estimators are given by expressions (11) (*srs*) and (16) (*sm*).
- The mean-square errors for the simulations of the estimators of the mean. These quantities are expressed as $EQM_{sim}(\hat{\bar{y}}) = E_{sim}(\hat{\bar{y}} - \bar{y})^2$.
- The variances of the estimators of the mean. These variances are given by expressions (9) (*srs*) and (15) (*sm*). In the case of the moving stratification, this is of course the proposed approximation.

A careful reading of the results seems to indicate that the variance estimator proposed for the moving stratum algorithm is not affected by a systematic bias no matter what the value for the coefficient of correlation between x and y. The results also seem to indicate that the approximate expression given for the variance of the estimator of the mean for the moving stratification is a valid approximation.

4.5 Interest of the Algorithm

Within the class of algorithms defined by the general algorithm, we call the mean horizon of an algorithm the quantity

$$\bar{b} = \frac{1}{N} \sum_{i=0}^{N-1} b_i.$$

For the simple design, we get $\bar{b}_{srs} = (N + 1)/2$. For the algorithm of the moving stratum, we have

$$\bar{b}_{sm} = \frac{1}{N} \left\{ \sum_{i=0}^{N-M-1} M + \sum_{i=N-M}^{N-1} (N - i) \right\}$$

$$= \frac{M}{N} \left\{ N - \frac{M-1}{2} \right\}.$$

Let us now assume that, as described in section 3.2, we select a sample using a design with proportional allocation in which all the strata are of the same size and in which the sizes of H strata are all equal. In such a design, the mean horizon has a value of

$$\bar{b}_{strat} = \frac{1}{2} \left(\frac{N}{H} + 1 \right).$$

A change in the mean horizon does not fundamentally affect the first order inclusion probabilities. The second order inclusion probabilities, on the other hand, are strongly affected by a change of horizon. In fact, it can easily be seen that the smaller the mean horizon, the smaller the probability of selecting two close individuals. (Two individuals are said to be close if the absolute value of the difference of their serial numbers in the data file is small.) Intuitively, we can expect the moving stratum algorithm to have a stratification effect similar to that of a stratified design with proportional allocation having the same mean horizon, *i.e.*, when

$$\bar{b}_{strat} = \bar{b}_{sm},$$

or in other words, when

$$M = N + \frac{1}{2} - \sqrt{\frac{1}{4} + N^2 \frac{H-1}{H}}. \qquad (17)$$

When N is large in relation to M, we have approximately

$$M \approx \frac{2N}{H}.$$

For each series of simulations presented in the Appendix (Table 2), the sizes of the moving strata (case: *sm*) were fixed in terms of the number of strata (case: *strat*) in such a way that the mean horizons of the two designs were identical in terms of expression (17). It is observed that, in such a case, the increased precision (compared to that of the simple design) derived from the moving stratum algorithm is of the same order of magnitude as that derived by means of stratification.

5. COMMENTS

The simulations that were carried out clearly show that the moving stratification algorithm yields a stratification effect of the same type as classical stratification with proportional allocation. This algorithm makes it possible to study the delicate problem of subdividing a continuous variable into strata. The estimators of the mean that are proposed are slightly biased. However, as long as $M \geq 10N/n$, simulations show that it is extremely rare for at least one of the c_i to fall outside of $[0,1]$. Moreover, we have shown that even when that probability is not zero, the bias of the estimator that we propose is negligible as long as $M \geq 3N/n$.

ACKNOWLEDGEMENTS

The author wishes to thank Pierre Lavallée for comments made on previous drafts of this paper. The author also wishes to thank an associate editor and a referee who provided numerous constructive comments that significantly improved this paper.

APPENDIX 1

Demonstration of the Lemmas and Propositions

Demonstration of Lemma 3

$$\mathrm{Var}[n_{i+1}]$$

$$= \mathrm{Var}[n_i] + \mathrm{Var}[I_{i+1}]$$

$$+ 2E\left(E\left\{\left(n_i - i\frac{n}{N}\right)E\left[I_{i+1} - \frac{n}{N} \mid n_i\right]\right\}\right).$$

Since

$$2E\left[E\left\{\left(n_i - i\frac{n}{N}\right)E\left[\left(I_{i+1} - \frac{n}{N}\right) \mid n_i\right]\right\}\right]$$

$$= 2E\left[\left(n_i - i\frac{n}{N}\right)\left(\frac{(b_i + i)n/N - n_i}{b_i} - \frac{n}{N}\right)\right]$$

$$= \frac{-2}{b_i}\mathrm{Var}[n_i],$$

we obtain

$$\mathrm{Var}[n_{i+1}] = \mathrm{Var}[n_i]\frac{b_i - 2}{b_i} + \frac{n}{N}\frac{N-n}{N},$$

$$i = 1, \ldots, N-1. \qquad (18)$$

We then show that (3) verifies the recursion equation (18) and the initial condition given by

$$\mathrm{Var}(n_1) = \frac{n}{N}\frac{N-n}{N}.$$

Demonstration of Proposition 1

Case 1: $i = 0$. From lemma 2 we immediately get:

$$E[I_k I_1] = E[E[I_k \mid n_1]n_1]$$

$$= \frac{n^2}{N^2} - \frac{n}{N}\frac{N-n}{N}\frac{1}{b_{k-1}}\prod_{\ell=1}^{k-2}\frac{b_\ell - 1}{b_\ell}.$$

Case 2: $i > 0$. Using lemma 2, we obtain:

$$E[I_{i+k}I_{i+1} \mid n_i = t]$$

$$= E[I_{i+k} \mid n_{i+1} = t+1]E[I_{i+1} \mid n_i = t]$$

$$= \left\{\frac{n}{N} - \left((t+1) - (i+1)\frac{n}{N}\right)\frac{1}{b_{i+k-1}}\prod_{\ell=i+1}^{i+k-2}\frac{b_\ell - 1}{b_\ell}\right\}$$

$$\times \left\{\frac{n}{N} - \left(t - i\frac{n}{N}\right)\frac{1}{b_i}\right\}.$$

Which means that

$$E[E[I_{i+k}I_{i+1} \mid n_i]]$$

$$= E\left\{\frac{n}{N} - \left((n_i+1) - (i+1)\frac{n}{N}\right)\frac{1}{b_{i+k-1}}\prod_{\ell=i+1}^{i+k-2}\frac{b_\ell - 1}{b_\ell}\right\}$$

$$\times \left\{\frac{n}{N} - \left(n_i - i\frac{n}{N}\right)\frac{1}{b_i}\right\}$$

$$= \frac{n^2}{N^2} - \frac{1}{b_{i+k-1}}\left\{\frac{n}{N}\frac{N-n}{N} - \frac{\mathrm{Var}[n_i]}{b_i}\right\}\prod_{\ell=i+1}^{i+k-2}\frac{b_\ell - 1}{b_\ell}.$$

Lemma 3 thus gives us $\mathrm{Var}[n_i]$. We immediately obtain (4).

Demonstration of Proposition 2

Using (13), we have

$$\Pr\left[n - M - \frac{n}{N} < n_{N-M} < \frac{N-n}{N} + n\right] = 1.$$

Therefore,

$$\Pr[0 \leq n - n_{N-M} \leq M] = 1.$$

Beginning with step $N - M$, the algorithm is a selection-rejection algorithm of the type described in section 3.1. This algorithm yields a sample of exactly $n - n_{N-M}$ observation units during the final M steps. Since $n - n_{N-M} \leq M$, this operation raises no difficulty and the algorithm is therefore of fixed size n.

APPENDIX 2

Tables, Bias Upper Bounds and Simulations

Table 1

Value of the Bias Upper Bounds C_α

N	n	Value of the Coefficient C_α				
		$M = \frac{N}{n}$	$M = \frac{2N}{n}$	$M = \frac{3N}{n}$	$M = \frac{4N}{n}$	$M = \frac{5N}{n}$
100	50	0.000000	0.000000	0.000000	0.000000	0.000000
	25	0.057326	0.002610	0.000185	0.000015	0.000001
	12	0.041716	0.002604	0.000235	0.000023	0.000002
	6	0.032227	0.002029	0.000134	0.000005	0.000000
	3	0.023515	0.000645	0.000000		
500	250	0.000000	0.000000	0.000000	0.000000	0.000000
	125	0.129091	0.006002	0.000437	0.000038	0.000004
	62	0.090863	0.005664	0.000534	0.000059	0.000007
	31	0.066891	0.004666	0.000484	0.000059	0.000008
	15	0.048544	0.003586	0.000384	0.000046	0.000006
	7	0.035508	0.002552	0.000215	0.000015	0.000001
	3	0.024046	0.000699	0.000000		
2,500	1,250	0.000000	0.000000	0.000000	0.000000	0.000000
	625	0.289060	0.013495	0.000987	0.000086	0.000008
	312	0.202458	0.012607	0.001190	0.000133	0.000016
	156	0.147113	0.010234	0.001064	0.000130	0.000017
	78	0.105662	0.007742	0.000841	0.000107	0.000015
	39	0.075975	0.005719	0.000634	0.000082	0.000012
	19	0.054525	0.004174	0.000466	0.000060	0.000008
	9	0.039560	0.003014	0.000301	0.000029	0.000002
	4	0.028388	0.001451	0.000034	0.000000	
12,500	3,125	0.646539	0.030208	0.002211	0.000193	0.000018
	1,562	0.452450	0.028177	0.002661	0.000297	0.000036
	781	0.327879	0.022798	0.002371	0.000290	0.000039
	390	0.234114	0.017131	0.001863	0.000238	0.000033
	195	0.166626	0.012500	0.001388	0.000181	0.000026
	97	0.118357	0.008995	0.001009	0.000133	0.000019
	48	0.084217	0.006452	0.000727	0.000096	0.000014
	24	0.060797	0.004689	0.000529	0.000069	0.000010
	12	0.044677	0.003461	0.000377	0.000044	0.000005
	6	0.033727	0.002356	0.000173	0.000008	0.000000
	3	0.024172	0.000712	0.000000		
62,500	3 906	0.732684	0.050942	0.005299	0.000649	0.000087
	1,953	0.522918	0.038250	0.004159	0.000531	0.000074
	976	0.371301	0.027833	0.003092	0.000403	0.000057
	488	0.263300	0.019979	0.002243	0.000295	0.000042
	244	0.186736	0.014259	0.001609	0.000213	0.000031
	122	0.132653	0.010168	0.001150	0.000152	0.000022
	61	0.094601	0.007273	0.000823	0.000109	0.000016
	30	0.067467	0.005207	0.000590	0.000078	0.000011
	15	0.049227	0.003820	0.000427	0.000054	0.000007
	7	0.035847	0.002637	0.000227	0.000016	0.000001
	3	0.024176	0.000713	0.000000		
312,500	4,882	0.829762	0.062191	0.006909	0.000901	0.000128
	2,441	0.587909	0.044596	0.005006	0.000659	0.000095
	1,220	0.416165	0.031758	0.003583	0.000474	0.000068
	610	0.294647	0.022555	0.002551	0.000339	0.000049
	305	0.208743	0.016008	0.001813	0.000241	0.000035
	152	0.147877	0.011356	0.001287	0.000171	0.000025
	76	0.105272	0.008098	0.000918	0.000122	0.000018
	38	0.075422	0.005817	0.000659	0.000087	0.000013
	19	0.054695	0.004238	0.000479	0.000062	0.000009
	9	0.039644	0.003038	0.000305	0.000030	0.000002
	4	0.028427	0.001457	0.000034	0.000000	

Table 2

Results of the Simulations, Simple Design, Stratification and Moving Stratification

ρ^2	Plan	Parameters	$E_{sim}\widehat{\mathrm{Var}}\,\hat{\bar{y}}$	$\mathrm{Var}\hat{\bar{y}}$	$EQM_{sim}\hat{\bar{y}}$
0.0	*sm*	$M = 18.83N/n$	0.01318	0.01317	0.01301
	srs		0.01317	0.01316	0.01296
	strat	$H = 2$	0.01319	0.01319	0.01318
0.2	*sm*	$M = 18.83N/n$	0.01210	0.01210	0.01187
	srs		0.01316	0.01316	0.01287
	strat	$H = 2$	0.01172	0.01188	0.01164
0.4	*sm*	$M = 18.83N/n$	0.01073	0.01073	0.01080
	srs		0.01316	0.01316	0.01320
	strat	$H = 2$	0.00943	0.00929	0.00946
0.6	*sm*	$M = 18.83N/n$	0.00957	0.00957	0.00954
	srs		0.01315	0.01316	0.01301
	strat	$H = 2$	0.00783	0.00778	0.00774
0.8	*sm*	$M = 18.83N/n$	0.00839	0.00839	0.00839
	srs		0.01315	0.01316	0.01322
	strat	$H = 2$	0.00630	0.00624	0.00622
1.0	*sm*	$M = 18.83N/n$	0.00757	0.00757	0.00760
	srs		0.01314	0.01316	0.01319
	strat	$H = 2$	0.00514	0.00508	0.00513
0.0	*sm*	$M = 8.65N/n$	0.01319	0.01319	0.01317
	srs		0.01317	0.01316	0.01296
	strat	$H = 4$	0.01320	0.01318	0.01316
0.2	*sm*	$M = 8.65N/n$	0.01107	0.01107	0.01084
	srs		0.01316	0.01316	0.01287
	strat	$H = 4$	0.01080	0.01076	0.01054
0.4	*sm*	$M = 8.65N/n$	0.00876	0.00876	0.00882
	srs		0.01316	0.01316	0.01320
	strat	$H = 4$	0.00811	0.00793	0.00796
0.6	*sm*	$M = 8.65N/n$	0.00695	0.00694	0.00688
	srs		0.01315	0.01316	0.01301
	strat	$H = 4$	0.00637	0.00639	0.00632
0.8	*sm*	$M = 8.65N/n$	0.00484	0.00484	0.00485
	srs		0.01315	0.01316	0.01322
	strat	$H = 4$	0.00402	0.00391	0.00390
1.0	*sm*	$M = 8.65N/n$	0.00312	0.00312	0.00313
	srs		0.01314	0.01316	0.01319
	strat	$H = 4$	0.00206	0.00197	0.00197
0.0	*sm*	$M = 4.21N/n$	0.01317	0.01317	0.01316
	srs		0.01317	0.01316	0.01296
	strat	$H = 8$	0.01321	0.01324	0.01325
0.2	*sm*	$M = 4.21N/n$	0.01067	0.01067	0.01046
	srs		0.01316	0.01316	0.01287
	strat	$H = 8$	0.01055	0.01047	0.01025
0.4	*sm*	$M = 4.21N/n$	0.00810	0.00809	0.00808
	srs		0.01316	0.01316	0.01320
	strat	$H = 8$	0.00794	0.00789	0.00789
0.6	*sm*	$M = 4.21N/n$	0.00592	0.00592	0.00588
	srs		0.01315	0.01316	0.01301
	strat	$H = 8$	0.00575	0.00564	0.00561
0.8	*sm*	$M = 4.21N/n$	0.00344	0.00344	0.00345
	srs		0.01315	0.01316	0.01322
	strat	$H = 8$	0.00315	0.00311	0.00308
1.0	*sm*	$M = 4.21N/n$	0.00124	0.00124	0.00125
	srs		0.01314	0.01316	0.01319
	strat	$H = 8$	0.00085	0.00079	0.00080

Table 2

Results of the Simulations, Simple Design, Stratification and Moving Stratification - end

ρ^2	Plan	Parameters	$E_{sim}\widehat{\mathrm{Var}}\,\hat{y}$	$\mathrm{Var}\,\hat{y}$	$EQM_{sim}\hat{y}$
0.0	*sm*	$M = 2.11N/n$	0.01319	0.01319	0.01328
	srs		0.01315	0.01316	0.01332
	strat	$H = 16$	0.01315	0.01308	0.01331
0.2	*sm*	$M = 2.11N/n$	0.01038	0.01036	0.01021
	srs		0.01317	0.01316	0.01334
	strat	$H = 16$	0.01034	0.01034	0.01025
0.4	*sm*	$M = 2.11N/n$	0.00796	0.00796	0.00792
	srs		0.01316	0.01316	0.01323
	strat	$H = 16$	0.00790	0.00801	0.00794
0.6	*sm*	$M = 2.11N/n$	0.00572	0.00573	0.00561
	srs		0.01315	0.01316	0.01299
	strat	$H = 16$	0.00568	0.00572	0.00563
0.8	*sm*	$M = 2.11N/n$	0.00295	0.00294	0.00290
	srs		0.01317	0.01316	0.01325
	strat	$H = 16$	0.00287	0.00288	0.00285
1.0	*sm*	$M = 2.11N/n$	0.00048	0.00048	0.00048
	srs		0.01317	0.01316	0.01335
	strat	$H = 16$	0.00037	0.00034	0.00034
0.0	*sm*	$M = 1.09N/n$	0.01325	0.01316	0.01310
	srs		0.01313	0.01316	0.01317
	strat	$H = 32$	0.01201	0.01239	0.01302
0.2	*sm*	$M = 1.09N/n$	0.01070	0.01062	0.01064
	srs		0.01313	0.01316	0.01316
	strat	$H = 32$	0.00972	0.01018	0.01083
0.4	*sm*	$M = 1.09N/n$	0.00807	0.00803	0.00811
	srs		0.01315	0.01316	0.01309
	strat	$H = 32$	0.00732	0.00751	0.00803
0.6	*sm*	$M = 1.09N/n$	0.00538	0.00534	0.00536
	srs		0.01315	0.01316	0.01310
	strat	$H = 32$	0.00484	0.00484	0.00543
0.8	*sm*	$M = 1.09N/n$	0.00283	0.00281	0.00276
	srs		0.01317	0.01316	0.01283
	strat	$H = 32$	0.00255	0.00276	0.00280
1.0	*sm*	$M = 1.09N/n$	0.00016	0.00016	0.00017
	srs		0.01317	0.01316	0.01304
	strat	$H = 32$	0.00012	0.00007	0.00011

REFERENCES

BEBBINGTON, A.C. (1975). A simple method of drawing a sample without replacement. *Applied Statistics*, 24, 136.

COCHRAN, W.G. (1977). *Sampling Techniques*. New York: Wiley.

DEVILLE, J.-C., and GROSBRAS, J.-M. (1987). Algorithmes de tirage. In *Les sondages*. Droesbeke, J.-J., Fichet, B., and Tassi, P. (Eds.). Paris: Economica, 209-233.

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

HÁJEK, J. (1971). Comment on an essay of D. Basu. In *Foundations of Statistical Inference*. Godambe V.P., and Sprott, D.A. (Eds). Toronto: Holt, Rinehart and Winston.

McLEOD, A.I., and BELLHOUSE, D.R. (1983). A convenient algorithm for drawing a simple random sampling. *Applied Statistics*, 32, 182-184.

SUNTER, A.B. (1977). List sequential sampling with equal or unequal probabilities without replacement. *Applied Statistics*, 26, 261-268.

SUNTER, A.B. (1986). Solutions to the problem of unequal probability sampling without replacement. *International Statistical Revue*, 54, 33-50.

YATES, F., and GRUNDY, P.M. (1953). Selection without replacement from within strata with probability proportional to size. *Journal of the Royal Statistical Society*, B, 15, 235-261.

Survey Methodology, June 1996
Vol. 22, No. 1, pp. 95–103
Statistics Canada

A View on Statistical Disclosure Control for Microdata

A.G. de WAAL and L.C.R.J. WILLENBORG[1]

ABSTRACT

Problems arising from statistical disclosure control, which aims to prevent that information about individual respondents is disclosed by users of data, have come to the fore rapidly in recent years. The main reason for this is the growing demand for detailed data provided by statistical offices caused by the still increasing use of computers. In former days tables with relatively little information were published. Nowadays the users of data demand much more detailed tables and, moreover, microdata to analyze by themselves. Because of this increase in information content statistical disclosure control has become much more difficult. In this paper the authors give their view on the problems which one encounters when trying to protect microdata against disclosure. This view is based on their experience with statistical disclosure control acquired at Statistics Netherlands.

KEY WORDS: Statistical disclosure control; Microdata; Uniqueness.

1. INTRODUCTION

Statistical disclosure control (SDC) is becoming increasingly important as a result of the growing demand for information provided by statistical offices. The information released by these statistical offices can be divided into two major parts: tabular data and microdata. Whereas tables have been released traditionally by statistical offices, microdata sets are released only since fairly recently. In the past the users of data usually did not have the tools to analyze these microdata sets properly themselves. Nowadays every serious researcher is in possession of a powerful personal computer. Analyzing microdata is therefore no longer a privilege of the statistical office. The users of data can and want to analyze these microdata themselves. This creates non-trivial SDC-problems.

A key problem in the theory of SDC for microdata is the determination of the probability that a record in a released microdata set is re-identified. In order to estimate this probability a number of different approaches have been attempted. The aim of these attempts differ considerably. In some publications the aim was to gain a qualitative insight into the probability of re-identification of an unspecified record from a microdata set. In other publications the aim was set much higher, namely to obtain the probability that a specific record is re-identified. These are, of course, extreme cases. The former case is comparatively easy to solve, although still difficult. The latter case is more difficult and may be impossible to solve.

In this paper we give an overview of the problems for which Statistics Netherlands has attempted to provide a solution and problems of which the suggested solution has attracted our attention. We consider the problems and their outline of the solutions, while technical points are skipped. The choice of the problems and the possible solutions we consider is heavily influenced by the experiences of Statistics Netherlands in the field of SDC.

The rest of this paper is organized as follows. Basic concepts are defined in Section 2. Preliminaries on SDC for microdata are the subject of Section 3. Our basic philosophy of SDC for microdata is discussed in Section 4. In Section 5 we describe the ideal situation for microdata: in this case we would have a probability for each record that this specific record can be re-identified. A somewhat less ideal situation is described in Section 6: in this case we have a probability for a data set that an unspecified record can be re-identified. In Section 7 we have to face reality: at the moment we do not have a good disclosure risk model and we have to be satisfied with heuristic arguments. In Section 8 we summarize our conclusions and suggest some possibilities for future research.

2. BASIC CONCEPTS

In this section a number of basic concepts are defined. We will assume that the statistical office wants to release a microdata set containing records of a sample of the population. Each record contains information about an individual entity. Such an entity could be a person, a household or a business enterprise. In the rest of this paper we will usually consider the individual entity to be a person, although this is not essential.

The two most important concepts in the field of SDC are re-identification and disclosure. Re-identification is said to occur if an attacker establishes a one-to-one relationship between a microdata record and a target individual with a sufficient degree of confidence. Following

[1] A.G. de Waal and L.C.R.J. Willenborg, Statistics Netherlands, Division of Research and Development, Department of Statistical Methods, P.O. Box 4000, 2270 JM Voorburg, The Netherlands (E-mail: TWAL@CBS.NL and LWLG@CBS.NL).

Skinner (1992) we distinguish between two kinds of disclosure. Re-identification disclosure occurs if the attacker is able to deduce the value of a sensitive variable for the target individual after this individual has been re-identified. Prediction disclosure (or attribute disclosure) occurs if the microdata enable the attacker to predict the value of a sensitive variable for some target individual with a sufficient degree of confidence. For prediction disclosure it is not necessary that re-identification has taken place. Most research so far has concentrated on re-identification disclosure. In this paper we will use the term disclosure to indicate re-identification disclosure unless stated otherwise.

Now, let us define what is meant by an identifying variable. A variable is called identifying if it can serve, alone or in combination with other variables, to re-identify some respondents by some user of the data. Examples of identifying variables are residence, sex, nationality, age, occupation and education. A subset of the set of identifying variables is the set of direct (or formal) identifiers. Examples of direct identifiers are name, address and public identification numbers. Direct identifiers must have been removed from a microdata set before it is released for else re-identification is very easy. Other identifiers in most cases do not have to be removed from the microdata set. A combination of identifying variables is called a key. The identifying variables that together constitute a key are also called key variables. A key value is a combination of scores on the identifying variables that together constitute the key.

In practice, determining whether or not a variable is identifying is a problem that can only be solved by sound judgment. No limitative list of intrinsically identifying variables exists, nor, for that matter, an unambiguous and well-defined set of rules to determine such variables. Selecting a set of identifying variables, and therefore of keys, is generally based on subjective assumptions about the population. Statistics Netherlands applies some criteria, like the visibility of the categories of a variable, to determine whether or not a variable is identifying, but these criteria do not provide a definite answer to this problem for all variables. Whether or not a variable is considered identifying is essentially a matter of judgment. In the remainder of this paper we will assume however that a set of keys has been determined.

The counterparts of identifying variables are the sensitive (or confidential) variables. A variable is called sensitive (or confidential) if some of the values represent characteristics a respondent would not like to be revealed about him. In principle, Statistics Netherlands considers all variables sensitive, but in practice some variables are considered more sensitive than others. Like in the case of identifying variables, determining whether or not a variable is sensitive can be solved only by sound judgment in practice. The variables sexual behavior and criminal past are generally considered sensitive, but for other variables this may depend on, for instance, cultural background. Keller and Bethlehem (1992) give as an example the variable income. In the Netherlands income is considered sensitive, whereas in Sweden it is not. Moreover, there are variables which should be considered both identifying and sensitive. An example of such a variable is ethnic membership. However, in the literature it is usually assumed that the identifying and sensitive variables can be divided into disjoint sets. In the remainder of this paper we will also assume that a set of sensitive variables has been determined which is disjoint from the set of identifying variables.

By using information about the identifying variables a potential attacker can try to disclose information about sensitive variables. Note that this way of disclosure is only possible in case the link between the values of the identifying variables and the values of the sensitive variables has not been perturbed by noise in the data or by a technique like data-swapping.

To end this section, we give a definition of SDC. Statistical disclosure control aims to reduce the risk that sensitive information of individual persons can be disclosed to an acceptable level. What is acceptable depends on the policy of the data releaser. In order to reduce the risk of disclosure an estimate for the risk of disclosure would be very helpful although it is not a necessary requisite (*cf.* Section 7). Some research has been devoted to defining and estimating this risk of disclosure.

3. PRELIMINARIES ON SDC FOR MICRODATA

As a customer of a statistical office, the user of a microdata set should be satisfied with its quality. The user is usually not interested in individual records, but only in statistical results which can be drawn from the total set of records. For instance, he wants to examine tables he has produced himself from the microdata set.

Because a microdata set is meant for statistical analysis it is not necessary that each record in the set is correct. The statistical office has the possibility to perturb records, *e.g.*, by adding noise or by swapping parts of records between different records, in order to reduce the risk of re-identification. By perturbing records the risk of re-identification is reduced because even when a correct re-identification takes place the information which is disclosed may be incorrect. In any case the attacker cannot be sure that the disclosed information is correct. The statistical office 'only' has to guarantee that the statistical quality of, for instance, the tables the user wants to examine is high enough. This may be quite complicated to achieve in practice, however.

Although data perturbation methods may prove to be useful, for the time being Statistics Netherlands does not use them. To protect its microdata sets Statistics Netherlands applies local suppression and global recoding only.

When local suppression is applied some values of variables in some records are set to 'missing', *i.e.*, deleted from the microdata set. When global recoding is applied some variables are given a coarser categorization. In a first step, we try to protect a microdata set by means of global recoding. However, when protecting a microdata set entirely by means of global recodings would result in a considerable information loss, we apply local suppressions as well. In this way we try to avoid that too much information will be lost. It should be clear that local suppressions are only applied parsimoniously.

An advantage of local suppression and global recoding is that these techniques preserve the integrity of the data. A disadvantage of local suppression is that it introduces a bias, because extreme values will be locally suppressed. However, when local suppressions are only applied parsimoniously, this bias will be small.

From the SDC point of view a user of the data should also be looked upon as a potential attacker. Hence, it is useful to consider the ways in which disclosure can take place. An attacker tries to match records from the microdata set with records from an identification file or with individuals from his circle of acquaintances. An identification file is a file containing records with values on direct identifiers and values on some other identifiers of the microdata set. The latter identifiers may be used to match records from the released microdata set with records from the identification file. After matching the direct identifiers in the identification file can be used to determine whose record has been matched, and the sensitive variables in the released microdata set can be used to disclose information about this person. A circle of acquaintances is the set of persons in the population for which the attacker knows the values on a certain key from the microdata set. So, a circle of acquaintances could actually be an identification file, and vice versa. In the rest of this paper we will therefore use the terms 'identification file' and 'circle of acquaintances' interchangeably.

In order for re-identification of a record of an individual to occur the following conditions have to be satisfied:

C_1. The individual is unique on a particular key value K.
C_2. The individual belongs to an identification file or a circle of acquaintances of the attacker.
C_3. The individual is an element of the sample.
C_4. The attacker knows that the record is unique in the population on the key K.
C_5. The attacker comes across the record in the microdata set.
C_6. The attacker recognizes the record of the individual.

Whenever one of the conditions C_1 to C_6 does not hold, re-identification cannot be accomplished with absolute certainty. If either condition C_1 or C_4 does not hold, then a matching can be made but the attacker cannot be sure that this leads to a correct re-identification.

It is clear from the conditions C_1 to C_6 that a 'good' model for the risk of re-identification should incorporate aspects of both the data set and the user. When a Dutch microdata set is used by someone in, say, China who is essentially unfamiliar with the Dutch population, then the risk of re-identification is negligible. In order to re-identify someone in a microdata set it is necessary to acquire sufficient knowledge about the population. The amount of work that should be done to acquire this knowledge is proportional to the safety of the microdata set.

4. A PHILOSOPHY OF SDC

It seems likely that the attention of a potential attacker is drawn by combinations of identifying variables that are rare in the sample or in the population. Combinations that occur quite often are less likely to trigger his curiosity. If he tries to match records deliberately then he will probably try to do this for key values that occur only a few times. If the user does not try to match records deliberately, but he knows an acquaintance with a rare key value then a record with that particular key value may trigger him to consider the possibility that this record belongs to this acquaintance. Moreover, the probability of a correct match is higher in case the number of persons that score on the matching key value is smaller. Finally, it is also very likely that among the persons that score on a rare key value there are many uniques if the key is augmented with an additional variable. Records that score on such rare combinations of identifying variables are therefore more likely to be re-identified.

In particular key values which occur only once in the population, *i.e.*, uniques in the population, can lead to re-identification. In the past emphasis was placed almost exclusively on uniqueness. It should be noted, however, that uniqueness is neither sufficient nor necessary for re-identification. If a person is unique in the population on certain key variables, but nobody realizes this, then this person may never be re-identified. If on the other hand this person is not unique in the population, but there is only one other person in the population with the same key, then this other person is, in principle, able to re-identify him. Furthermore, suppose a person is not unique, but belongs to a small group of people. Suppose also that the attacker happens to know information about him which is not considered to be identifying by the statistical office, but which is contained in the released microdata set, then it is very well possible that he is unique on the key combined with the new information. So, it is possible that a person is re-identified although he is not unique on the keys of identifying variables in the population. Finally, prediction disclosure may occur. That is, if a person is not unique in the population, but belongs to a group of people with (almost) the same score on a particular sensitive variable, then sensitive information can be disclosed about this individual without actual re-identification. Prediction disclosure is not discussed further in this paper. For more information on prediction disclosure we refer to Skinner (1992), US Department of Commerce (1978), Duncan and Lambert (1986), and Cox (1986).

SDC should concentrate on key values that are rare in the population. A probability that information from a particular respondent, whose data are included in a microdata set, is disclosed should reflect the 'rareness' of the key value of this respondent's record. A probability for the event that information from an arbitrary respondent is disclosed should reflect the 'overall rareness' of the records in the data set. If there are many records in a microdata set of which the key value is rare, then the probability of disclosure for this data set should be high. In the next sections we will examine some attempts to incorporate these ideas within a mathematical framework.

5. RE-IDENTIFICATION RISK PER RECORD

In an ideal world (as far as SDC is concerned) a releaser of microdata would be able to determine a risk of re-identification for each record, *i.e.*, a probability that the respondent of this record can be re-identified. Such a risk per record would enable us to adopt the following strategy. First, order the records according to their risk of re-identification with respect to a single key. Second, select a maximum risk the statistical office is willing to accept. Finally, modify all the records for which the risk of re-identification with respect to the key chosen is too high. Repeat this procedure for each key in case there are more keys.

Unfortunately, we do not live in such an ideal world at the moment. However, steps towards the ideal situation have been made by Paass and Wauschkuhn (1985), and Fuller (1993). In Paass and Wauschkuhn (1985) it is assumed that a potential attacker has both a microdata file, released by a statistical office, and an identification file at his disposal. Between both files there may be many data incompatibilities. These data incompatibilities may be caused by *e.g.*, coding errors, by different definitions of categories or by 'noise' in the data. By assuming a probability distribution for these data incompatibilities and a disclosure scenario Paass and Wauschkuhn develop a sophisticated model to estimate the probability that a specific record from the microdata file is re-identified. The type of distribution of the errors that caused the data incompatibilities was assumed to be known to the attacker. The variance of the errors was assumed unknown to him. A potential attacker had to estimate this variance, on the basis of the (assumed) knowledge of the statistical production process. The model of Paass and Wauschkuhn is essentially based on discriminant analysis and cluster analysis.

Paass and Wauschkuhn distinguish between six different scenarios. Each scenario corresponds to a special kind of attacker. The number of records in the identification file and the information content of the identification file depend on the chosen scenario. An example of such a scenario is the journalist scenario, where a journalist selects records with extreme attribute combinations in order to re-identify respondents with the aim of showing that the statistical office fails to secure the privacy of its respondents.

Paass and Wauschkuhn apply their method to match records from the identification file with records from the microdata file. If the probability that a specific record from the identification file belongs to a specific record from the microdata set is high enough, then these two records are matched. This probability is the probability of re-identification per record, conditional on a particular disclosure scenario.

Müller, Blien, Knoche, Wirth *et al.* (1991) and Blien, Wirth and Müller (1992) applied the method recommended in Paass and Wauschkuhn (1985) to real data. When compared to simple matching, *i.e.*, a record is considered re-identified by an attacker if he succeeds in finding a unique value set in the microdata file which is identical to a value set in the identification file, the method suggested by Paass and Wauschkuhn turned out to be not superior. Apparently, the number of correctly matched records when applying the method by Paass and Wauschkuhn was in disagreement with the probability of re-identification per record.

In the context of masking procedures, *i.e.*, procedures for microdata disclosure limitation by adding noise to the microdata, Fuller (1993) obtained an expression for the probability that a specific record in the released microdata set is the same as a specific target record from an identification file. That is, an expression for the re-identification probability per record is derived. To derive this expression several assumptions are made. It is assumed that the data, the noise and errors in the data are normally distributed. Moreover, it is assumed that the covariance matrices of both the noise and the errors in the data are known to an attacker. Finally, it is assumed that the data have been obtained by simple random sampling. These assumptions allow Fuller (1993) to derive his expression for the re-identification probability by means of probability theoretical considerations. Unfortunately, the approach by Fuller has not been tested on real data yet. Hence, it is hard judge the applicability of this approach. For a comment on the approach by Fuller see Willenborg (1993).

Paass and Wauschkuhn (1985), and Fuller (1993) are mainly interested in the effects of noise that has (unintentionally and intentionally, respectively) been added to the data on the disclosure risk. A weak point of their respective approaches is the, implicit, assumption that the key is a high-dimensional one. Assuming a high-dimensional key implies that (almost) everyone in the population is unique. The probability that a combination or key value occurs more than once in the population is negligible. This makes the computation of the probability of re-identification per record considerably easier. On the other hand, in case of low-dimensional keys it is not unlikely that certain key values occur many times in the population. Therefore, deriving a probability of re-identification per record for low-dimensional keys is much harder than for high-dimensional keys, because for high-dimensional keys the probability of statistical twins in the population is almost zero.

A good model for the re-identification risk per record does not appear to exist at the moment. In Section 6 we therefore consider less ambitious models, namely models for the re-identification risk per file.

6. RE-IDENTIFICATION RISK PER FILE

In a somewhat less ideal world a releaser of microdata would not be able to determine the risk of re-identification for each record, but he would be able to determine the risk that an unspecified record from the microdata set is re-identified. In this case, the statistical office should decide on the maximal risk it is willing to take when releasing a microdata set. If the actual risk is less than the maximal risk, then the microdata set can be released. If the actual risk is higher than the maximal risk, then the microdata set has to be modified. Determining which records have to be modified remains a problem, however.

A basic model to determine the probability that an arbitrary record from a microdata set is re-identified has been proposed by Mokken, Pannekoek and Willenborg (1989) and Mokken, Kooiman, Pannekoek and Willenborg (1992). In Mokken *et al.* (1989) only the case where there is a single researcher, an unstratified population and a single key is considered. It has been extended to include the cases of subpopulations, multiple researchers and multiple keys (*cf.* Willenborg 1990a; Willenborg 1990b; Mokken *et al.* 1992). The model of Mokken *et al.* (1992) takes three probabilities into account. The first probability, f, is equal to the sampling fraction. In other words, f, is the probability that a randomly chosen person from the population has been selected in the sample. The second probability, f_a, is the probability that a specific researcher who has access to the microdata knows the values of a randomly chosen person from the population on a particular key. The third probability, f_u, is the probability that a randomly chosen person from the population is unique in the population on a particular key. Combining these three probabilities, f, f_a and f_u, the probability that a record from a microdata set is re-identified can be evaluated.

For each sample element a number of variables is measured. The values obtained by these measurements (scores) are collected in records, one for each sample element. It is assumed that the variables in the key are either categorical variables or variables for which the measurements fall into a finite number of categories. Together, the records constitute a data set S that will be made available to an researcher R. We recall that whenever we use the term disclosure in fact re-identification disclosure is meant. The model of Mokken *et al.* (1989, 1992) does not take prediction disclosure into account.

In terms of the Paass and Wauschkuhn (1985) set-up f_a and f_u together reflect the *Informationsgehalt der Überschneidungsmerkmale*, *i.e.*, the information content of the matching values. The various scenarios they consider differ in terms of f_a and f_u. In particular, f_u is influenced by the number of variables and the information content of these variables, *i.e.*, their categorization, an attacker has at his disposal to re-identify a record. The parameter f_a is determined by the number of records that are contained in the information file.

With respect to researcher R and key K there is a circle of acquaintances A. Obviously, A and its size $|A|$ will depend on the particular researcher R as well as on the key K and the variables as registered and coded in the data set.

It is assumed that if conditions C_1, C_2 and C_3 of the conditions for re-identification given in Section 3 hold, then conditions C_4, C_5 and C_6 hold too. Condition C_4 is a rather exacting one, but it can be introduced as an assumption for the sake of convenience in formulating a disclosure risk model. Note that it then yields a worst-case situation, in the sense that fallible perception and memory or other sources of ignorance, confusion and uncertainty for a potential discloser are excluded. Taken as an assumption together with C_5 and C_6 the implication is that the occurrence of any unique acquaintance E of R in data set S is equivalent to re-identification by R. It is assumed that re-identification of a record implies disclosure of confidential information. Thus re-identification can be treated as equivalent to disclosure. Implicitly, it is assumed that the link between the identifying variables and the sensitive variables has not been disturbed by a technique such as data-swapping.

Furthermore it is assumed that both the identifying and the confidential information are free of error or noise to researcher R, contrary to *e.g.*, Paass and Wauschkuhn (1985), and Fuller (1993). Clearly, this assumption is unrealistic for most microdata sets.

The disclosure risk D_R for a certain microdata set S with respect to a certain researcher R and a certain key K, is defined to be the probability that the researcher makes at least one disclosure of a record in S on the basis of K. In order to apply a criterion based on the disclosure risk, the value of this quantity for a given data set has to be determined. An expression for this quantity can be derived on the basis of a set of assumptions.

In the model of Mokken *et al.* the following assumptions are made in addition to $C_1 - C_6$:

A_1. The circle of acquaintances A can be considered as a random sample from the population.

A_2. The data set S is a random sample from the population.

Assumption A_1 serves to imply that the probability that a randomly chosen element from the population is an acquaintance of R is $f_a = \mid A \mid /N$, where N is the size of the population. As a consequence the expected number of unique elements in A, $\mid U_a \mid$, is equal to $f_a \mid U \mid = \mid A \mid f_u$, where U is the set of unique persons in the population and $\mid U \mid$ its size. Obviously assumption A_2 implies that the probability that a specific unique element E is selected in the sample is f. These assumptions allow one to obtain a very simple expression for the disclosure risk D_R in terms of f, f_a and f_u, namely

$$D_R = 1 - \exp(-Nff_af_u). \qquad (1)$$

Two of the parameters in the model of Mokken *et al.* (1989, 1992), f_a and f_u, are unknown. The parameter f_a can be 'guestimated', *i.e.*, obtained by inspired guesswork, by assuming different scenarios an attacker may follow. A number of such scenarios has been described in Paass and Wauschkuhn (1985) and Paass (1988). Evaluating f_a seems difficult, however. In order to estimate the other parameter, f_u, a number of models has been proposed in the literature. Models to estimate the number of uniques in the population, and hence f_u, that have been proposed include the Poisson-gamma model (Bethlehem, Keller and Pannekoek 1989; Mokken *et al.* 1989; Willenborg, Mokken and Pannekoek 1990; De Jonge 1990), the negative binomial superpopulation model (Skinner, Marsh, Openshaw and Wymer 1990), the Poisson-lognormal model (Skinner and Holmes 1992; Hoogland 1994), models based on equivalence classes (Greenberg and Zayatz 1992) and models based on modified negative binomial-gamma functions (Crescenzi 1992; Coccia 1992). As we have remarked in Section 4 not only the number of population uniques is important, but the numbers of cells with two, three, *etc.* persons are important as well. The Poisson-gamma model, the Poisson-lognormal model and the negative binomial superpopulation model can be applied to estimate the number of cells with two, three, *etc.* persons as well. It seems that the other models mentioned above can be extended in order to estimate these numbers. A major drawback is that the results are not very reliable in many cases.

From the model by Mokken *et al.* (1989, 1992) it is clear that the statistical office that disseminates the data is able to influence the risk of re-identification. The statistical office basically has two ways to do this. First of all, the size of the data set can be reduced, *i.e.*, the sampling fraction f can be reduced. A reduction of f implies a reduction of the risk. However, lowering f is generally undesirable, because usually f has to be reduced substantially to be effective. This implies that only a small part of the data available can be released. The second way in which the statistical office can influence the re-identification risk is by reducing the number of population uniques, *i.e.*, by reducing f_u. The fraction f_u depends on the information provided by the key variables. The less information the key variables provide the less uniques there are in the population. In order words, f_u can be reduced by collapsing categories (global recoding) and by replacing values by missings (local suppression). Collapsing categories is a global action, because it generally affects many records; replacing values by missings is a local action because it affects only a few individual records. Usually, the loss in information when reducing f_u is considerably less than the loss in information when reducing f. Therefore, a statistical office will usually choose to control the re-identification risk by reducing f_u rather then reducing f. The third possibility of controlling the re-identification risk, *i.e.*, by reducing f_a, is not applied in practice, because f_a is difficult to model.

Although the model by Mokken *et al.* (1989, 1992) provides some insight in how to reduce the disclosure risk it can hardly be used as a basis for the protection of microdata sets. The reason for this is that the two parameters of the model, f_u and f_a, are often difficult to evaluate. Usually there is insufficient data available to estimate f_u and f_a accurately. We conclude that even a model for a re-identification risk for an entire microdata set is difficult to apply in practice. In Section 7 we therefore face reality in which we have no satisfactory model for either the re-identification risk per record or re-identification risk for an entire microdata set.

7. INTUITIVE RE-IDENTIFICATION RISK

In reality we are, unfortunately, forced to base SDC on heuristic arguments rather than on a solid theoretical basis. The SDC rules mentioned in this section all reduce the re-identification risk. It is, however, not possible to evaluate this reduction of the re-identification risk. At Statistics Netherlands, rules for SDC of microdata are based on testing whether scores on certain keys occur frequently enough in the population. A few problems arising here are the determination of the keys that have to be examined, the way to estimate the number of persons in the population that score on a certain key, to make operational the meaning of the phrase 'frequently enough' by determining *e.g.*, (a) threshold value(s), and how to determine appropriate SDC-measures.

Statistics Netherlands distinguishes between two kinds of microdata sets. The first kind is a so-called public use file. A public use file can be obtained by everybody. The keys that have to be examined for a public use file are all combinations of two identifying variables. The number of identifying variables is limited, and certain identifying variables, such as place of residence are not included in a public use file. Moreover, sampling weights have to be examined before they can be included in a public use file, because there are many situations in which weights can give additional information (*cf.* De Waal and Willenborg 1995a).

For instance, when a certain subpopulation is oversampled then this subpopulation can be recognized by the low weights associated with its members in the sample. Weights may only be published when they do not provide additional information that can be used for disclosure purposes. In case sampling weights are not considered suited for publication SDC measures should be taken, such as subsampling the units with a low weight in order to get a subsample in which all units have approximately the same weight. Because the weights are approximately equal assuming that they are exactly equal would introduce only a small error. The second kind of microdata set is a so-called microdata set for research. A microdata set for research can only be obtained by well-respected (statistical) research offices. The information content of a microdata set for research is much higher than that of a public use file. The number of identifying variables is not limited and an identifying variable such as place of residence may be included in a microdata set for research. Because of the high information content of a microdata set for research, researchers have to sign a declaration stating that they will protect any information about an individual respondent that might be disclosed by them. The keys that have to be examined for a microdata set for research consist of three-way combinations of variables describing a region with variables describing the sex, ethnic group or nationality of a respondent with an ordinary identifying variable.

The rules Statistics Netherlands applies for SDC are based on the following idea: a key value, *i.e.*, a combination of scores on the identifying variables that together constitute the key, is considered safe for release if the frequency that this key value occurs in the population is more than a certain threshold value d_0. This value d_0 was chosen after a careful and extensive search considering many different values and comparing the records which have to be modified for each value of d_0. The value that leads to the 'most likely' set of records which have to be modified has been chosen to be the value of d_0. Which records are considered to be the 'most likely' ones to be modified is a matter of personal judgment.

When applying one of the above rules we are generally posed with the problem that we do not know the number of times that a key value occurs in the population. We only have the sample available to us. The population frequency of a key value has to be estimated based upon the sample. For large regions it is possible to use an interval estimator to test whether or not a key value occurs often enough in a region. This interval estimator is based on the assumption that the number of times that a key value occurs in the population is Poisson distributed (*cf.* Pannekoek 1995). However, for relatively small regions the number of respondents is low, which causes the estimator to have a high variance which in turn causes a lot of records to be modified. To estimate the number of times that a key value occurs in a small region we therefore suggest to apply a point estimator. We will now discuss some possibilities for such an estimator.

A simple point estimator for the number of times that a certain key value occurs in a region is the direct point estimator. The fraction of a key value in a region i is estimated by the sample frequency of this key value in region i divided by the number of respondents in region i. The population frequency is then estimated by this estimated fraction multiplied by the number of inhabitants in region i. When the number of respondents in region i is low, which is often the case, the direct estimator is unreliable. Another point estimator is based on the assumption that the persons who score on a certain key value are distributed homogeneously over the population. In this case the fraction of a key value in region i can be estimated by the fraction in the entire sample. The advantage of this, so-called, synthetic, estimator is that the variance is much smaller than the variance of the direct estimator. Unfortunately, the homogeneity assumption is usually not satisfied which causes the estimator to be biased. However, a combined estimator can be constructed with both an acceptable variance and an acceptable bias by using a convex combination of the direct estimator and the synthetic estimator. Such a combined estimator has been tested in Pannekoek and de Waal (1995).

Another practical problem that deserves attention is top-coding of extreme values of continuous (sensitive) variables. These extreme values may lead to re-identification because these values are rare in the population. At the moment Statistics Netherlands uses an interval estimator to test whether there is a sufficient number of individuals in the population who score on a 'comparable' value of the continuous variable (*cf.* Pannekoek 1992). If this is the case, then the extreme value may be published, otherwise the extreme value must be suppressed. In order to apply this method in practice it remains to specify what is meant by 'sufficient' and by 'comparable'.

Some important practical problems occur when determining which protection measures should be taken when a microdata set appears to be unsafe. In that case the original data set must be modified in such a way that the information loss due to SDC-measures is as low as possible while the resultant data set is considered safe. In De Waal and Willenborg (1994a) and De Waal and Willenborg (1995b) a model for determining the optimal local suppressions is presented. Determining the optimal global recodings is much more difficult. Comparing the information loss due to global recodings to the information loss to local suppressions is already a problem. In De Waal and Willenborg (1995c) this latter problem is solved by using the entropy.

Currently a general purpose software package for SDC of microdata is being developed at Statistics Netherlands (*cf.* De Jong 1992; De Waal and Willenborg 1994b; Van Gelderen 1995; Pieters and De Waal 1995; De Waal and

Pieters 1995). The package, ARGUS, should enable the statistical office to analyze the data and to carry out suitable protection measures. It will consist of two separate parts: μ-ARGUS for SDC of microdata and τ-ARGUS for SDC of tabular data. The structure of the package is such that it will be possible to specify different disclosure control rules. This implies that ARGUS will be suited for other statistical offices too. Moreover, it will be possible to incorporate changes in the rules fairly easily in the package.

8. CONCLUSIONS

There is one important conclusion one can draw from this paper: SDC still offers a lot of possibilities for future research, despite the considerable amount of research that has been carried out to date. The theory of SDC for microdata has a number of gaps. Among the technical problems that remain to be solved are the following. When we want to release data for small regions we need an acceptable estimator for the number of times that a key value occurs in these regions. Such an estimator is difficult to construct, although the preliminary results obtained at Statistics Netherlands seem encouraging. An important practical problem is the determination of appropriate global recodings and local suppressions. Yet another one is the determination of the number of uniques, or more generally the number of rare frequencies, in the population. Some of the models proposed in Section 6 appear to be acceptable, but can probably be improved upon. An alternative approach is to determine which elements in the sample are unique in the population. In Verboon (1994), and Verboon and Willenborg (1995) this approach is examined. An extension of the model by Mokken *et al.* (1989, 1992) to estimate the risk of re-identification of a file is yet another problem to be solved. This extension should take into account that measurement errors have been made and that population uniqueness is not necessary in order to disclose information. Finally, a model to estimate the re-identification risk per record would be very welcome. In fact, it would yield a sound criterion to judge the safety of a microdata set. This criterion can guide one in producing safe microdata sets by applying SDC-measures such as global recoding and local suppression.

Apart from technical problems there are also some policy problems. Based on the policy that a statistical office wants to pursue the following decisions should be made. The combinations of variables that should be examined should be specified. Suitable threshold values should be selected.

More and better software must be developed in order to deal with time-consuming calculations. For microdata, software must be developed to indicate which records and variables must be modified, and how they should be modified, when applying a particular disclosure rule. At the time of writing an international project on SDC is about to start. The participating institutions in this project are the Eindhoven University of Technology, the University of Manchester, the University of Leeds, the Office of Population Censuses and Surveys (OPCS), the Istituto Nazionale di Statistica (ISTAT), the Consortio Padova Ricerche (CPR), and Statistics Netherlands. One of the major aims of the project is to develop software for the SDC of both microdata (μ-ARGUS) and tabular data (τ-ARGUS).

Finally, some very practical problems remain to be solved. An example of such a problem is the determination of a set of rules for selecting identifying variables. Such a set of rules would be a very valuable asset. Without these rules identifying variables are selected by making subjective choices. Developing such a set of rules is another goal of the above mentioned SDC-project.

REFERENCES

BETHLEHEM, J.A., KELLER, W.J., and PANNEKOEK, J. (1989). Disclosure control of microdata. *Journal of the American Statistical Association*, 85, 38-45.

BLIEN, U., WIRTH, H., and MÜLLER, M. (1992). Disclosure risk for microdata stemming from official statistics. *Statistica Neerlandica*, 46, 69-82.

COCCIA, G. (1992). Disclosure risk in Italian current population surveys. International Seminar on Statistical Confidentiality, Dublin.

COX, L.H. (1986). Comment on Duncan and Lambert (1986). 19-21.

CRESCENZI, F. (1992). Estimating population uniques; methodological proposals and applications on Italian census data. International Seminar on Statistical Confidentiality, Dublin.

De JONG, W.A.M. (1992). ARGUS: An integrated system for data protection. International Seminar on Statistical Confidentiality, Dublin.

De JONGE, G. (1990). The estimation of population unicity from microdata files (in Dutch), Internal note, Statistics Netherlands, Voorburg.

De WAAL, A.G., and PIETERS, A.J. (1995). ARGUS user's guide. Report, Statistics Netherlands, Voorburg.

De WAAL, A.G., and WILLENBORG, L.C.R.J. (1994a). Minimizing the number of local suppressions in a microdata set. Report, Statistics Netherlands, Voorburg.

De WAAL, A.G., and WILLENBORG, L.C.R.J. (1994b). Development of ARGUS: past, present, future. Report, Statistics Netherlands, Voorburg.

De WAAL, A.G., and WILLENBORG, L.C.R.J. (1995a). Statistical disclosure control and sampling weights. Report, Statistics Netherlands, Voorburg.

De WAAL, A.G., and WILLENBORG, L.C.R.J. (1995b). Local suppression in statistical disclosure control and data editing. Report, Statistics Netherlands, Voorburg.

De WAAL, A.G., and WILLENBORG, L.C.R.J. (1995c). Optimal global recoding and local suppression. Report, Statistics Netherlands, Voorburg.

DUNCAN, G.T., and LAMBERT, D. (1986). Disclosure-limited data dissemination. *Journal of the American Statistical Association*, 81, 10-28.

FULLER, W.A. (1993). Masking procedures for microdata disclosure limitation. *Journal of Official Statistics*, 9, 383-406.

GREENBERG, B.V., and ZAYATZ, L.V. (1992). Strategies for measuring risk in public use microdata files. *Statistica Neerlandica*, 46, 33-48.

HOOGLAND, J. (1994). Protecting microdata sets against statistical disclosure by means of compound Poisson distributions (in Dutch). Report. Statistics Netherlands, Voorburg.

KELLER, W.J., and BETHLEHEM, J.A. (1992). Disclosure protection of microdata: problems and solutions. *Statistica Neerlandica*, 46, 5-19.

MOKKEN, R.J., PANNEKOEK, J., and WILLENBORG, L.C.R.J. (1989). Microdata and disclosure risks, CBS Select 5, Statistical Essays, Staatsuitgeverij (The Hague), 181-200.

MOKKEN, R.J., KOOIMAN, P., PANNEKOEK, J., and WILLENBORG, L.C.R.J. (1992). Disclosure risks for microdata. *Statistica Neerlandica*, 46, 49-67.

MÜLLER, W., BLIEN, U., KNOCHE, P., WIRTH, H. *et al.* (1991). *The Factual Anonymity of Microdata* (in German). Stuttgart: Metzler-Poeschel Verlag.

PAASS G., and WAUSCHKUHN, U. (1985). Data access, data protection and anonymization - analysis potential and identifiability of anonymized individual data (in German). Gesellschaft für Mathematik und Datenverarbeitung, Oldenbourg-Verlag, Munich.

PAASS, G. (1988). Disclosure risk and disclosure avoidance for microdata. *Journal of Business and Economic Studies*, 6, 487-500.

PANNEKOEK, J. (1992). Disclosure control of extreme values of continuous identifiers (in Dutch). Report, Statistics Netherlands, Voorburg.

PANNNEKOEK, J. (1995). Statistical methods for some simple disclosure limitation rules. Report, Statistics Netherlands, Voorburg.

PANNEKOEK, J., and de WAAL, A.G. (1995). Synthetic and combined estimators in statistical disclosure control. Report, Statistics Netherlands, Voorburg.

PIETERS, A.J., and De WAAL, A.G. (1995). A demonstration of ARGUS. Report, Statistics Netherlands, Voorburg.

SKINNER, S., MARSH, C., OPENSHAW, S., and WYMER, C. (1990). Disclosure avoidance for census microdata in Great Britain. *Proceedings of the 1990 Annual Research Conference*, U.S. Bureau of the Census, Washington, DC, 131-143.

SKINNER, C.J. (1992). On identification disclosure and prediction disclosure for microdata. *Statistica Neerlandica*, 46, 21-32.

SKINNER, C.J., and HOLMES, D.J. (1992). Modelling population uniqueness. International Seminar on Statistical Confidentiality, Dublin.

US DEPARTMENT OF COMMERCE (1978). Report on statistical disclosure and disclosure avoidance techniques. Statistical Policy Working Paper 2, Washington DC.

Van GELDEREN, R. (1995). ARGUS: Statistical disclosure control of survey data. Report, Statistics Netherlands, Voorburg.

VERBOON, P. (1994). Some ideas for a masking measure for statistical disclosure control. Report, Statistics Netherlands, Voorburg.

VERBOON, P., and WILLENBORG, L.C.R.J. (1995). Comparing two methods for recovering population uniques in a sample. Report. Statistics Netherlands, Voorburg.

WILLENBORG, L.C.R.J. (1990a). Remarks on disclosure control of microdata. Report, Statistics Netherlands, Voorburg.

WILLENBORG, L.C.R.J. (1990b). Disclosure risks for microdata sets: stratified populations and multiple investigators. Report, Statistics Netherlands, Voorburg.

WILLENBORG, L.C.R.J. (1993). Discussion statistical disclosure limitation. *Journal of Official Statistics*, 9, 469-474.

WILLENBORG, L.C.R.J., MOKKEN, R.J., and PANNEKOEK, J. (1990). Microdata and disclosure risks. *Proceedings of the 1990 Annual Research Conference*, U.S. Bureau of the Census, Washington DC, 167-180.

CONTENTS TABLE DES MATIÈRES

Volume 24, No. 1, March/mars 1996

Volume 24, No. 2, June/juin 1996

JOURNAL OF OFFICIAL STATISTICS

An International Quarterly 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 11, Number 4, 1995

Lars Lyberg, U/LEDN, Statistics Sweden, S-115 81 Stockholm, Sweden

GUIDELINES FOR MANUSCRIPTS

Before having a manuscript typed for submission, please examine a recent issue (Vol. 19, No. 1 and onward) of *Survey Methodology* as a guide and note particularly the following points:

1. Layout

1.1 Manuscripts should be typed on white bond paper of standard size (8½ × 11 inch), one side only, entirely double spaced with margins of at least 1½ inches on all sides.
1.2 The manuscripts should be divided into numbered sections with suitable verbal titles.
1.3 The name and address of each author should be given as a footnote on the first page of the manuscript.
1.4 Acknowledgements should appear at the end of the text.
1.5 Any appendix should be placed after the acknowledgements but before the list of references.

2. Abstract

The manuscript should begin with an abstract consisting of one paragraph followed by three to six key words. Avoid mathematical expressions in the abstract.

3. Style

3.1 Avoid footnotes, abbreviations, and acronyms.
3.2 Mathematical symbols will be italicized unless specified otherwise except for functional symbols such as "exp(·)" and "log(·)", etc.
3.3 Short formulae should be left in the text but everything in the text should fit in single spacing. Long and important equations should be separated from the text and numbered consecutively with arabic numerals on the right if they are to be referred to later.
3.4 Write fractions in the text using a solidus.
3.5 Distinguish between ambiguous characters, (e.g., w, ω; o, O, 0; l, 1).
3.6 Italics are used for emphasis. Indicate italics by underlining on the manuscript.

4. Figures and Tables

4.1 All figures and tables should be numbered consecutively with arabic numerals, with titles which are as nearly self explanatory as possible, at the bottom for figures and at the top for tables.
4.2 They should be put on separate pages with an indication of their appropriate placement in the text. (Normally they should appear near where they are first referred to).

5. References

5.1 References in the text should be cited with authors' names and the date of publication. If part of a reference is cited, indicate after the reference, e.g., Cochran (1977, p. 164).
5.2 The list of references at the end of the manuscript should be arranged alphabetically and for the same author chronologically. Distinguish publications of the same author in the same year by attaching a, b, c to the year of publication. Journal titles should not be abbreviated. Follow the same format used in recent issues.