## SURVEY <br> METHODOLOGY

JUNE 1999

VOLUME 25

NUMBER 1

# SURVEY METHODOLOGY 

# A JOURMAL PUBLISHED BY STATISTICS CANADA 

JUNE 1999 • VOLUME 25 • NUMBER 1

Published by authority of the Minister responsible for Statistics Canada

- Minister of Industry, 1999

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

Marketing Division, Statistics Canada,
Ottawa, Ontario, Canada K1A OT6.
September 1999
Catalogue no. 12-001-XPB
Frequency: Semi-annual
ISSN 0714-0045
Ottawa

## SURVEY METHODOLOGY

## A Journal Published by Statistics Canada

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

## MANAGEMENT BOARD

Chairman G.J. Brackstone
Members
R. Platek (Past Chairman)

D. Binder<br>G.J.C. Hole<br>F. Mayda (Production Manager)<br>C. Patrick

D. Roy
M.P. Singh

EDITORIAL BOARD
Editor M.P. Singh, Statistics Canada

## Associate Editors

D.R. Bellhouse, University of Western Ontario
D. Binder, Statistics Canada
J.-C. Deville, INSEE
J.D. Drew, Statistics Canada
J. Eltinge, Texas A\&M University
W.A. Fuller, Iowa State University
R.M. Groves, University of Maryland
M.A. Hidiroglou, Statistics Canada
D. Holt, Central Statistical Office, U.K.
G. Kalton, Westat, Inc.
R. Lachapelle, Statistics Canada
P. Lahiri, University of Nebraska-Lincoln
S. Linacre, Australian Bureau of Statistics
G. Nathan, Central Bureau of Statistics, Israel

D. Pfeffermann, Hebrew University<br>J.N.K. Rao, Carleton University<br>L.-P. Rivest, Université Laval<br>I. Sande, Bell Communications Research, U.S.A.<br>F.J. Scheuren, Ernst and Young, LLP<br>J. Sedransk, Case Western Reserve University<br>R. Sitter, Simon Fraser University<br>C.J. Skinner, University of Southampton<br>R. Valliant, Westat, Inc.<br>V.K. Verma, University of Essex<br>P.J. Waite, U.S. Bureau of the Census<br>J. Waksberg, Westat, Inc.<br>K.M. Wolter, National Opinion Research Center<br>A. Zaslavsky, Harvard University

Assistant Editors $\quad$ P. Dick, H. Mantel, B. Quenneville and D. Stukel, Statistics Canada

## EDITORIAL POLICY

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

## Submission of Manuscripts

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

## Subscription Rates

The price of Survey Methodology (Catalogue no. 12-001-XPB) is $\$ 47$ per year in Canada and US $\$ 47$ per year Outside Canada. Subscription order should be sent to Statistics Canada, Operations and Integration Division, Circulation Management, 120 Parkdale Avenue, Ottawa, Ontario, Canada K1A 0T6 or by dialling (613) $951-7277$ or $1800700-1033$, by fax (613) $951-1584$ or $1800889-9734$ or by Internet: order@statcan.ca. A reduced price is available to members of the American Statistical Association, the International Association of Survey Statisticians, the American Association for Public Opinion Research, and the Statistical Society of Canada.

## SURVEY METHODOLOGY

## A Journal Published by Statistics Canada

Volume 25, Number 1, June 1999

## CONTENTS

In This Issue ..... 1
H. KRÖGER, C.-E. SÄRNDAL and I. TEIKARI
Poisson Mixture Sampling: A Family of Designs for Coordinated Selection Using Permanent Random Numbers ..... 3
W. R. BELL and M. KRAMER
Toward Variances for X-11 Seasonal Adjustments ..... 13
J. DE HAAN, E. OPPERDOES and C. M. SCHUT Item Selection in the Consumer Price Index: Cut-off Versus Probability Sampling ..... 31
P. DUCHESNE
Robust Calibration Estimators ..... 43
Y. TILLÉ
Estimation in Surveys Using Conditional Inclusion Probabilities: Complex Design ..... 57
N.G.N. PRASAD and J.N.K. RAO
On Robust Small Area Estimation Using a Simple Random Effects Model ..... 67
F.A.S. MOURA and D. HOLT
Small Area Estimation Using Multilevel Models ..... 73
M. CHATTOPADHYAY, P. LAHIRI, M. LARSEN and J. REIMNITZ
Composite Estimation of Drug Prevalences for Sub-State Areas ..... 81
S. RUBIN BLEUER and M. KOVAČEVIĆ
Some Issues in the Estimation of Income Dynamics ..... 87
P. F. TATE
Utilising Longitudinally Linked Data from the British Labour Force Survey ..... 99
S. GABLER, S. HAEDER and P. LAHIRI
A Model Based Justification of Kish's Formula for Design Effects for Weighting and Clustering ..... 105

# In This Issue 

## Dear Readers,

I would like to share with you good news on two fronts. First, the upcoming December issue will mark the $25^{\text {th }}$ anniversary of Survey Methodology. This issue of the journal will be slightly larger than usual and will contain papers from some very prominent statisticians of our time. Second, we are looking into producing an electronic version of the Journal. Our current plan is to make the December 1999 issue available on a special Web site. All current subscribers will be able to download the Journal free of charge. Based on the response to this trial we will see if it is feasible to offer the Journal in that medium instead of or in addition to the current paper version. Watch for further information in the next issue. As usual your comments and suggestions are always welcome.

This issue covers a variety of topics - three papers on small area estimation, four papers on general estimation issues and two each on new sampling designs and data analysis.

Kröger, Särndal and Teikari they introduce a new family of sampling designs, called Poisson Mixture Sampling, which comprises of a weighted mixture of Poisson and Bernoulli sampling. Through a Monte Carlo study using Finnish data, they empirically show that, for a variety of point estimators, Poisson Mixture Sampling is more efficient than the usual Poisson sampling.

Bell and Kramer deal with the long standing problem of estimating the variance of X-11 estimators. Each month, statistical bureaus throughout the world publish the raw estimates of variables along with a corresponding measure of error, usually a standard error or a coefficient of variation. However, the corresponding seasonally adjusted or trend estimates, obtained by application of the X-11 method, do not have such an associated measure of error. Bell and Kramer present an interesting approach that offers a practical solution to this problem. They calculate two sources of error: one resulting from the sampling error and the other resulting from the use of ARIMA extrapolations at the two ends of the series.

De Haan, Opperdoes and Schut discuss sampling the items in a commodity group for input to the Consumer Price Index using scanner data. While most statistical offices currently use a judgmental selection procedure, this naturally leads to biased estimates. The authors address the question of whether probability sampling would lead to better results in terms of mean square error, with interesting results.

Pierre Duchesne considers a new class of robust calibration estimators used to obtain constrained weights at given intervals. The process involves changing carefully selected robust default weights into calibrated weights. In a brief empirical study, the new estimators are illustrated and compared to estimators which have already been proposed.

Tillé investigates a repeated sampling approach which takes into account auxiliary information. First he generalizes the use of conditional inclusion probabilities for use with any sampling design. He then constructs estimators that can be viewed as optimal linear estimators, and compares them with the GREG-estimator. He contrasts all of the estimators via a set of simulations. Finally he discusses the problem of interaction between the design and the auxiliary variables.

Prasad and Rao consider the problem of small area estimation through the use of a random effects model. While traditional methods rely on model-based methods to obtain estimates of small area means, Prasad and Rao obtain design-based (model-assisted) estimates by integrating survey weights. Corresponding model-based estimators of the mean squared errors (MSE) of the small area estimates are also derived. Through simulation results, they show that their MSE estimator has low bias and is quite stable.

In their paper on small area estimation, Moura and Holt focus on multilevel models, which make use of auxiliary information at both the unit and the small area levels, and allow small area random effects for both the intercepts and the regression slopes. The fixed and random effects parameters are estimated using restricted iterative generalized least squares. The mean square error is approximated. Simulations show that the model can lead to better small area estimators than those based on simpler models, that overspecification of the model does not lead to a serious loss of efficiency, and that the MSE approximation and associated MSE estimator work well.

Chattopadhyay, Lahiri, Larsen and Reimnitz consider estimation of proportions for rare events in small areas. Their method is illustrated and compared to other approaches using data from a telephone survey of alcohol and drug use. Their proposed estimator combines
census-based demographic estimates of population within age/sex/county groups with survey-based empirical Bayes estimates of proportions within those groups. A jackknife estimator of mean square error is proposed which captures variability due to estimation of model parameters.

The problem of estimating longitudinal low income proportions from a longitudinal survey having a complex design is studied in Rubin-Bleuer and Kovacevic. Two design-based estimators are considered: one based on both the longitudinal and cross-sectional sample, called the "mixed estimator", and one based entirely on the longitudinal sample. Through simulation, the two estimators are compared in the presence of attrition using models of compensation that assume "missing at random" and "completely missing at random" underlying mechanisms. The results are illustrated using data from two longitudinal sueveys.

Tate considers linking data on the same individuals from subsequent quarters of the British Labour Force Survey, a rotating panel survey in which one fifth of the sample is renewed at each occasion. She analyzes the various factors which can introduce bias into analyses derived from such linked data. In particular, she studies the possible effects of sample attrition, respondent errors and proxy respondents. She also considers various approaches to adjusting for these biases.

Finally, in a short note, Gabler, Haeder and Lahiri present a model-based justification for Kish's well known formula for design effects. They show that the result is actually a conservative value for the actual design effect.

The Editor

# Poisson Mixture Sampling: A Family of Designs for Coordinated Selection Using Permanent Random Numbers 

HANNU KRÖGER, CARL-ERIK SÄRNDAL and ISMO TEIKARI ${ }^{1}$


#### Abstract

This paper introduces Poisson Mixture sampling, a family of sampling designs so named because each member of the family is a mixture of two Poisson sampling designs, Poisson $\pi \mathrm{ps}$ sampling and Bernoulli sampling. These two designs are at opposite ends of a continuous spectrum, indexed by a continuous parameter. Poisson Mixture sampling is conceived for use with the highly skewed populations often arising in business surveys. It gives the statistician a range of different options for the extent of the sample coordination and the control of response burden. Some Poisson Mixture sampling designs give considerably more precise estimates than the usual Poisson $\pi \mathrm{ps}$ sampling. This result is noteworthy, because Poisson $\pi \mathrm{ps}$ is in itself highly efficient, assuming it is based on a strong measure of size.


KEY WORDS: Business surveys; Skewed populations; Response burden; Regression estimators.

## 1. THE OBJECTIVES OF POISSON MIXTURE SAMPLING

Poisson Mixture (Pomix) sampling is a family of sampling designs suitable for business surveys with its often highly skewed populations. The Pomix family contains the traditional Bemoulli sampling and Poisson $\pi \mathrm{ps}$ sampling designs as two special cases, situated at the two extremes of a range of possibilities indexed by a continuous parameter. This parameter, called the Bernoulli width and denoted $B$, satisfies $0 \leq B \leq f_{R}$, where $f_{R}$ is the predetermined expected sampling fraction in the "take-some" portion of the population, that is, the portion where randomized selection is applied.

Random numbers, in the form of independent realizations of the Unif $(0,1)$ random variable, are commonly used in modern computerized sample selection. Fan, Muller and Rezucha (1962) introduced several sequential (unit by unit) drawing mechanisms based on random numbers. Now, Pomix sampling is based on the Permanent Random Number (PRN) technique, which calls for assigning at birth a random number to each unit in the frame (the business register, in the case of a business survey). The random number is permanent in the sense of remaining attached to the unit during its entire lifetime. The PRN technique makes it easy to achieve coordination of samples and control of response burden. Early references to sampling with the aid of PRN's are Brewer, Early and Joyce (1972) and Atmer, Thulin and Bäcklund (1975). A recent review of different PRN techniques, and important extensions, are given in Ohlsson (1995).

Poisson $\pi p s$ sampling has the desirable feature of selecting large units with relatively greater probability than small units, whose contribution to estimated population totals will in any case be relatively minor. Coordination of

Poisson $\pi p s$ samples with the aid of PRN's was introduced by Brewer et al., (1972) and is discussed subsequently by several authors, including Sunter (1977) and Ohlsson (1995).

Similarly as in Poisson $\pi p s$, Pomix sampling allows control of the response burden, as explained in the next section. Larger enterprises will be selected relatively more often than smaller ones. The selection is controlled through rotation so as to distribute the response burden. Another objective of Pomix sampling is for all (or a substantial portion) of the population units to be included in sample (therefore observed, so that their basic data can be updated) with regularity over a period of time. The objective can be, for example, that every enterprise should be in sample at least once during a ten or twelve year period.

## 2. THE SELECTION PROCEDURE UNDER POISSON MIXTURE SAMPLING

Denote the finite population as $U=\{1, \ldots, k, \ldots, N\}$, where the integer $k$ represents the $k$-th population unit. Denote by $y$ the variable of interest and by $y_{k}$ its value for unit $k ; y_{k}$ is unknown before sample selection and observation. With the unit $k \in U$ is also associated a known positive size measure $x_{k}$. Its role in Pomix sampling is to bring about a more frequent selection of the larger units; in addition, the size variable should be used as an auxiliary variable at the estimation stage.

A sample, $s$, is realized from the population $U$. The size of $s$ may be random; its expected size, denoted $n$, is a number fixed in advance. We allow $s$ to consist of two nonoverlapping parts, $s=s_{C} \cup s_{R}$, where $s_{C}$ is called the certainty part of $s$ and $s_{R}$ the randomization part of $s$. The part $s_{C}$, consisting of very large units selected with

[^0]probability one, is designated in a preliminary step, with the aid of the known size measures $x_{k}$. One procedure for this is given in section 3. Depending on the population characteristics, it could happen that no certainty part is designated, so that $s_{C}$ is the empty set, but this eventuality is rather exceptional with the highly skewed populations usually occurring in business surveys.

A frequently used synonymous term for the certainty part is take-all stratum. If the take-all stratum is denoted $U_{C}$, a probabilistic description is to say that $s_{C}$ is drawn from the take-all stratum $U_{C}$ so that $s_{C}=U_{C}$ with probability one. We denote the size of $s_{C}=U_{C}$ by $n_{C}$.

Next, the randomization sample, $s_{R}$, is selected from the rest of the population, $U_{R}=U-U_{C}$, of size $N_{R}=N-n_{C}$. It consists of units with inclusion probability $\pi_{k}$ strictly less than unity. In this paper, $s_{R}$ is drawn by Pomix sampling (thus it uses the PRN technique). The size of $s_{R}$ is random; its expected size, denoted $n_{R}$, is fixed by the equation $n_{R}=n-n_{C}$.

In this paper, we use the term Poisson sampling for selection corresponding to independent unit by unit Bernoulli trials with any inclusion probabilities $\pi_{k}$. More specifically, by Poisson $\pi p s$ we mean Poisson sampling with $\pi_{k}$ directly proportional to a measure of size. Bernoulli sampling is the special case of Poisson sampling where all $\pi_{k}$ are equal.

For Pomix sampling, we need some more notation. For unit $k \in U_{R}$, define the relative size measure

$$
\begin{equation*}
A_{k}=n_{R} x_{k} / \sum_{U_{R}} x_{k} \tag{2.1}
\end{equation*}
$$

We can from now on assume that $A_{k}<1$ for all $k \in U_{R}$, because if $A_{k}<1$ had not been true for certain units $k \in U_{R}$, then the procedure in section 3 for constructing the certainty part of the sample would in effect have assigned those units to the certainty part $s_{C}$.

We now define Pomix sampling with the aid of a twodimensional diagram. On the horizontal axis, a unit's PRN is plotted. On the vertical axis, a size-related measure, $Q_{k}$, is plotted. At each survey occasion, a new sampling selection region is designated by rotation in this diagram, and sample coordination is realized in the manner that we now describe.

Pomix sampling is characterized by two parameters, $B$ and $f_{R}$, where $f_{R}$ such that $0<f_{R}=n_{R} / N_{R}<1$ denotes the fixed expected sampling rate in $U_{R}=U-U_{C}$. The parameter $B$, called the Bemoulli width, is such that $0 \leq B \leq f_{R}$. For every unit $k \in U_{R}$, define

$$
\begin{equation*}
Q_{k}=\frac{f_{R}-B}{1-B} \frac{x_{k}}{\bar{x}_{U_{R}}}=\frac{1-B / f_{R}}{1-B} A_{k} \tag{2.2}
\end{equation*}
$$

where $\bar{x}_{U_{R}}=\sum_{U_{R}} x_{k} / N_{R}$. For $B=0$, we have $Q_{k}=A_{k}$, which is the size measure for the usual Poisson $\pi$ ps sampling. At the other extreme, $B=f_{R}$, we have $Q_{k}=0$ for all $k \in U_{R}$; in this case, size will play no role in the selection
from $U_{R}$, which will be seen to reduce to Bernoulli sampling. The measures $Q_{k}$ are used in Pomix selection of coordinated samples, as we now describe.

Start with a plot of the points $\left(r_{k}, Q_{k}\right)$ for $k \in U_{R}$, where $r_{k}$ denotes the PRN attached at birth to unit $k$, and $Q_{k}$ is given by (2.2). With reference to Figure 1, Pomix sampling is defined as follows: Include in the randomization sample, $s_{R}$, all units having PRN's $r_{k}$ falling in the ( $0, B$ ] interval, and also include some units having PRN's $r_{k}$ in the ( $\left.B, 1\right]$ interval, namely, those for which $Q_{k}$ is at least equal to a threshold value situated on the line joining the points ( $B, 0$ ) and $(1,1)$. The selection area is thus the shaded part of Figure 1. Note that since $A_{k}<1$ for all $k \in U_{R}$, we have by (2.2) that $Q_{k}<\left(1-B / f_{R}\right) /(1-B) \leq 1$ for all $k \in U_{R}$, when $0 \leq B<f_{R}$.

For Poisson $\pi$ ps sampling, the inclusion probability of unit $k$ is $\pi_{k}=A_{k}$ given by (2.1). Coordination of PRN-based Poisson samples was introduced by Brewer et al., (1972) using a graphical representation corresponding to $B=0$ in Figures 1 and 2. At each occasion, the selection area is then a triangle; the unit's PRN on the horizontal axis is plotted against the unit's size measure, $A_{k}$, on the vertical axis. Coordination is obtained by "moving the selection area over" to the right. Coordination of Pomix samples is realized in a similar fashion.


Fig 1. Sampling at time 1


Fig 2. Sampling at time 2

Figures 1 and 2 illustrate how coordinated Pomix sampling from $U_{R}=U-U_{C}$ can be carried out at two consecutive survey occasions. In each of the two figures, the sample is defined as the set of units for which the point $\left(r_{k}, Q_{k}\right)$ falls in the shaded area. The "starting point" on the PRN axis is the point to the right of which we start to count units for inclusion in the sample. At time 1 (Figure 1), the starting point is 0 ; at time 2 (Figure 2), the starting point is $D$. (In general, the starting point can be a randomly selected point in the unit interval; in other words, the sample identified in Figure 2 is also the one that would be selected at time $t=1$ if at that time the randomly selected starting point on the PRN axis had been equal to $D$.)

A convenient way to achieve sample rotation is through the constant shift method, which implies that the starting point is moved over to the right by a fixed amount at every
new occasion of sampling; see Ohlsson (1995). The constant $D$ is called the constant shift. The starting point at time 3 would thus be $2 D$, and so on.

In the following we examine Pomix sampling and estimation at a single occasion, and we can concentrate on Figure 1 (time 1), with starting point 0 on the PRN axis. The algorithm for Pomix sampling with parameters $B$ and $f_{R}$, and starting point 0 , is thus as follows: From Figure 1, unit $k$ is included in the randomization part, $s_{R}$, (i) if $0<r_{k} \leq B$, or (ii) if $B<r_{k} \leq 1$ and $Q_{k} \geq\left(r_{k}-B\right) /(1-B)$.

Consequently, $k$ is included in $s_{R}$ if

$$
0<r_{k} \leq B+Q_{k}(1-B) .
$$

Because $r_{k} \sim \operatorname{Unif}(0,1)$, the first order inclusion probabilities under Pomix sampling are

$$
\pi_{k}= \begin{cases}B+Q_{k}(1-B) & \text { for } k \in U_{R}  \tag{2.3}\\ 1 & \text { for } k \in U_{C} .\end{cases}
$$

It is easy to see that the inclusion probabilities satisfy the necessary requirement that their sum must equal the expected sample size fixed in advance:

$$
\sum_{U} \pi_{k}=\sum_{U_{C}} 1+\sum_{U_{R}}\left\{B+Q_{k}(1-B)\right\}=n_{C}+n_{R}=n .
$$

We now note two extreme cases of the family of Pomix sampling schemes: Bernoulli sampling is obtained if $B=f_{R}$ in the Pomix algorithm, because then $Q_{k}=0$ for all $k \in U_{R}$, and the algorithm becomes: Include unit $k \in U_{R}$ in $s_{R}$ if $0<r_{k} \leq f_{R}$, which is Bernoulli sampling. Poisson $\pi p s$ sampling is obtained if $B=0$ in the Pomix algorithm, because then the algorithm becomes: Include unit $k \in U_{R}$ in $s_{R}$ if $0<r_{k} \leq A_{k}$, where $A_{k}$ is given by (2.1). But this is Poisson $\pi$ ps sampling from $U_{R}$, the inclusion probability being $\pi_{k}=A_{k}$, that is, directly proportional to the size measure $x_{k}$.

Pomix sampling is a mixture of Poisson $\pi \mathrm{ps}$ and Bernoulli in that the Pomix inclusion probability, $\pi_{k}=B+Q_{k}(1-B)$, equals a linear combination of the inclusion probabilities that apply under the two extreme designs, weighted by the relative Bernoulli width, $\lambda=B / f_{R}$, such that $0 \leq \lambda \leq 1$. We have $\pi_{k}=\lambda \pi_{k}^{\mathrm{BE}}+(1-\lambda) \pi_{k}^{\pi \mathrm{Ps}}$, where $\pi_{k}^{\mathrm{BE}}=f_{R}$ for all $k$ (Bernoulli) and $\pi_{k}^{\pi_{k} s}=A_{k}$ (Poisson $\pi \mathrm{ps}$ ).

The character of Pomix sampling is determined by its two parameters, $B$ and $f_{R}$. To illustrate, we note from (2.1) and (2.3) that the inclusion probability of unit $k \in U_{R}$ is $\pi_{k}=B+\left(1-B / f_{R}\right) A_{k}$. Thus, for a unit $k$ that is large (but not large enough to qualify for $s_{C}=U_{C}$ ), so that $A_{k}$ is near unity, we have, to close approximation, $\pi_{k}=B+1-B / f_{R}$. By contrast, for a unit that is small, so that its value $A_{k}$ is very near zero, we have, to close approximation, $\pi_{k}=B$, independently of the size. For example, with $f_{R}=10 \%$, the following Table 1 shows how the inclusion probability $\pi_{k}$ varies with $B$, where $B=0$ is Poisson $\pi \mathrm{ps}$, and $B=f_{R}=0.10$ is Bernoulli.

Table 1
Values of the Inclusion Probability $\pi_{k}$ as a Function of the Parameter $B$ and the Relative Size $A_{k}$ When the Fixed Expected Sampling Rate, $f_{R}$, is 0.1

| Values of $\pi_{k}$ |  |  |  |  |  |
| :--- | :--- | :--- | :--- | :--- | :--- |
| Value of $B$ | 0 | 0.03 | 0.05 | 0.07 | 0.10 |
| $A_{k}=0$ (small unit) | 0 | 0.03 | 0.05 | 0.07 | 0.10 |
| $A_{k}=1$ (large unit) | 1 | 0.73 | 0.55 | 0.37 | 0.10 |

This illustrates that for a Pomix sampling design close to Bernoulli ( $B$ near 0.10), the inclusion probabilities of large and small units alike lie near the fixed expected sampling rate, 0.10 . By contrast, in a Pomix design close to Poisson $\pi \mathrm{ps}$ ( $B$ near 0 ), a small unit is practically certain not to be in sample, and a large unit is practically sure to be in sample. The table also illustrates how Pomix sampling with an intermediate value of $B$ will modify the inclusion probabilities: a small unit's chances to appear in the sample are decreased somewhat compared to Bernoulli, and the selection of a large unit becomes less probable than under Poisson $\pi$ ps.

The implications for response burden are: The total response burden on the population rests the same for all values of $B$; an expected total of $n=n_{R}+n_{C}$ units are always asked to report. Compared to Poisson $\pi \mathrm{ps}(B=0)$, the fixing of a value $B$ in the interior of the interval $\left[0, f_{R}\right]$ will have the effect of shifting some of the response burden from larger onto smaller units; at the same time the precision of the estimates is increased in many cases (see sections 5 and 6).

Finally, we need to mention the second order inclusion probabilities under Pomix sampling, because they are required for the design-based variance calculation. They are simple. If $\pi_{k l}$ denotes the probability that units $k$ and $l$ are both in the sample, then

$$
\begin{equation*}
\pi_{k l}=\pi_{k} \pi_{\ell} \tag{2.4}
\end{equation*}
$$

for $k \neq l \in U=U_{R} \cup U_{C}$, because the PRN's $r_{k}$ are independent realizations of the $\operatorname{Unif}(0,1)$ random variable. For $k=l$, we have $\pi_{k l}=\pi_{k k}=\pi_{k}$. The multiplicative feature (2.4) of the $\pi_{k l}$ greatly simplifies the design-based variance calculation. We get a simple, single-sum variance estimator, as in (4.2) below.

## 3. DETERMINING THE CERTAINTY PART OF THE SAMPLE

If the population is highly skewed, a set of units (the certainty part of the sample, $s_{C}$ ) will be sampled with probability one, and Pomix sampling can be used for randomized selection in the remaining part of the population, $U_{R}$. Several procedures could be considered for the construction of the certainty set; here we give one that is reasonable (though not necessarily optimal) and used in the

Monte Carlo simulation reported later in section 5. The certainty set is designated with the aid of the known positive size measures $x_{k}$ through the following procedure in one or more steps. An expected sample size, $n$, is fixed for the whole sample, $s=s_{R} \cup s_{C}$. In step one, compute the relative size measures $A_{k(1)}=n x_{k} / \sum_{U} x_{k}$ for $k \in U$. Those units $k$, if any, for which $A_{k(1)} \geq 1$ are assigned to the certainty part. They form a set denoted $U_{C(1)}$; let its size be $n_{C(1)}$. The procedure is then repeated to see if additional units should be assigned to the certainty part. In step two, calculate the relative size measures $A_{k(2)}=\left(n-n_{C(1)}\right)$ $x_{k} / \sum x_{k}$ where the summation extends over the set $U-U_{C(1)}$. If $A_{k(2)}<1$ for all $k \in U-U_{C(1)}$, the procedure stops, and the final certainty part is $s_{C}=U_{C(1)}$. But if $A_{k(2)} \geq 1$ for some units, then these are also assigned to the certainty part, and so on until a step $i$ is reached where all intermediate relative size measures $A_{k(i)}$ are less than unity. The ultimate certainty part $s_{C}$ will contain, say, $n_{C}$ units, and we then have $A_{k}<1$ for all $k \in U-s_{c}$, where $A_{k}=n_{R} x_{k} / \sum x_{k}$ with $n_{R}=n-n_{C}$, and the sum extends over $U^{k}-s_{C}$.

## 4. ESTIMATION FOLLOWING POMIX SAMPLING

Although the auxiliary variable serves a useful purpose at the sampling stage, we advocate using it also at the estimation stage. To estimate the population total, $Y=\sum_{U} y_{k}$, consider the generalized regression (GREG) estimator

$$
\begin{equation*}
\hat{Y}_{\text {GREG }}=\sum_{s} a_{k} g_{k} y_{k} \tag{4.1}
\end{equation*}
$$

where $a_{k}=1 / \pi_{k}$ is the sampling weight and the second weight, the $g$-weight, is given by

$$
g_{k}=1+(X-\hat{X})^{\prime} \boldsymbol{T}_{s}^{-1} x_{k} / c_{k}
$$

where $\boldsymbol{x}_{k}$ denotes the auxiliary vector value for unit $k$, $\hat{\boldsymbol{X}}=\sum_{s} a_{k} x_{k}$, and $\boldsymbol{T}_{s}=\sum_{s} a_{k} x_{k} x_{k}^{\prime} / c_{k}$. The auxiliary information requirement is that the vector total $X=\sum_{U} x_{k}$ must be known from a reliable source. The unidimensional size variable $x_{k}$ used for computing the $Q_{k}$ in the Pomix sampling scheme can be one of the components of $\boldsymbol{x}_{k}$ or it can a linear combination of the components of $x_{\boldsymbol{k}}$. In the empirical study reported in section $5, x_{k}$ is unidimensional, and $x_{k}=x_{k}$. The constants $c_{k}$, specified by the user, provide a means of weighing the data, in addition to the survey weights $a_{k}$. An often used choice is $c_{k}=1$ for all $k$.

A commonly used estimator of the variance of the GREG estimator (see Sämdal, Swensson and Wretman 1992, Ch. 6) is given as a quadratic form in $g_{k} e_{k k}$ where $e_{k}$ is the regression residual $e_{k}=y_{k}-\boldsymbol{x}_{k}^{\prime} \boldsymbol{b}$, where $\hat{b}=\boldsymbol{T}_{s}^{-1} \sum_{s} a_{k} x_{k} y_{k} / c_{k}$. One of the advantages of Pomix sampling is that the corresponding variance estimation is simple. This is because the PRN's $r_{k}$ are independent realizations from the $\operatorname{Unif}(0,1)$ distribution. Hence equation
(2.4) applies, and all product terms of the quadratic form are zero. With only the squared terms left, the variance estimator becomes simply $\hat{V}=\sum_{s} a_{k}\left(a_{k}-1\right) g_{k}^{2} e_{k}^{2}$. Finally, because $a_{k}-1=0$ for all $k \in s_{C}$, we get the variance estimator used in the Monte Carlo study reported later in section 5 , namely,

$$
\begin{equation*}
\hat{V}=\sum_{s_{R}} a_{k}\left(a_{k}-1\right) g_{k}^{2} e_{k}^{2} \tag{4.2}
\end{equation*}
$$

## 5. A MONTE CARLO STUDY OF POMIX SAMPLING USING FINNISH DATA

To illustrate various aspects of Pomix sampling, we conducted a Monte Carlo study involving four different estimators of the population total $Y$. The experiment involved repeated draws of samples as well as repeated assignments of the set of $N$ PRN's to the population units. Note that since every assignment of the $N$ PRN's to the population units is a random outcome, a proper Monte Carlo study also requires repetitions of the PRN assignments. Therefore, after the first assignment of the $N$ PRN's, we selected 100 Pomix samples, using a fixed value of the Bernoulli width $B$. (Each sample was realized using a new, randomly selected starting point on the PRN axis). Then a new set of $N$ PRN's were assigned, 100 additional samples were drawn, and so on, until we had reached $100 \times 100=10,000 \mathrm{PRN} /$ sample pairs, for the given value of $B$. For each of the 10,000 pairs, we computed the four point estimators, the corresponding four variance estimators, and the corresponding four confidence intervals. With 10,000 repetitions, we expect the Monte Carlo error to be rather small.

The four estimators used in the Monte Carlo study have the following expressions, where $a_{k}=1 / \pi_{k}$ is the sampling weight of unit $k$, and $\pi_{k}$ is given by (2.3):

1. The Horvitz-Thompson estimator,

$$
\hat{Y}_{1}=\sum_{s} a_{k} y_{k}=\sum_{U_{c}} y_{k}+\sum_{s_{k}} a_{k} y_{k}
$$

2. The (combined) ratio estimator,

$$
\hat{Y}_{2}=X \hat{b}_{2}
$$

where $X=\sum_{U} x_{k}$ and $\hat{b}_{2}=\sum_{s} a_{k} y_{k} / \sum_{s} a_{k} x_{k}$. It is a special case of (4.1) such that $x_{k}=x_{k}=c_{k}$.
3. The GREG estimator

$$
\hat{Y}_{3}=\sum_{U_{c}} y_{k}+\sum_{s_{k}} a_{k} y_{k}+\hat{b}_{3}\left(X_{R}-\sum_{s_{k}} a_{k} x_{k}\right)
$$

where $X_{R}=\sum_{U_{R}} x_{k}$ and $\hat{b}_{3}=\sum_{s_{R}} a_{k}\left(a_{k}-1\right) y_{k} x_{k} / \sum_{S_{R}}$ $a_{k}\left(a_{k}-1\right) x_{k}^{2}$. It ${ }^{R}$ is a special case of (4.1) such that $x_{k}=x_{k}$ and $c_{k}=\left(a_{k}-1\right)^{-1}$.
4. The (separate) ratio estimator,

$$
\hat{Y}_{4}=\sum_{U_{c}} y_{k}+X_{R} \hat{b}_{4}
$$

where $X_{R}=\sum_{U_{R}} x_{k}$ and $\hat{b}_{4}=\sum_{s_{R}} a_{k} y_{k} / \sum_{s_{R}} a_{k} x_{k}$.
For Poisson $\pi$ ps sampling ( $B=0$ ), we have $\hat{b}_{4}=$ ( $\sum_{s_{R}} y_{k} / x_{k}$ )/ $n_{s_{R}}$, where $n_{s_{s_{R}}}$ is the random size of $s_{R}$; the corresponding estimator $\hat{Y}_{4_{k}}$ was considered by Brewer et al., (1972). Now $\hat{Y}_{2}$ and $\hat{Y}_{4}$ differ in that the regression slope is calculated in $\hat{Y}_{2}$ on the pooled sample $s$, but in $\hat{Y}_{4}$ the slope is calculated separately for the randomization sample $s_{R}$. Finally, $\hat{Y}_{3}$ differs from $\hat{Y}_{2}$ and $\hat{Y}_{4}$ in that it uses the weighting $a_{k}\left(a_{k}-1\right)$, instead of just $a_{k}$. Note that all of $\hat{Y}_{2}$, and $\hat{Y}_{3}$ and $\hat{Y}_{4}$ are members of the GREG family of estimators given by (4.1). By equating $\hat{Y}_{2}, \hat{Y}_{3}$ and $\hat{Y}_{4}$ to (4.1), we find the $g$-weights implied by each of the three estimators. These weights are required for the variance estimation. We can expect the simulation to show that $\hat{Y}_{2}, \hat{Y}_{3}$ and $\hat{Y}_{4}$, which use the auxiliary variable both at the design stage and at the estimation stage, will improve on (have smaller variance than) the HT estimator $\hat{Y}_{1}$, which uses the auxiliary information only at the sampling stage, but the extent of the improvement is unpredictable and interesting to observe.

We used a real data population for the Monte Carlo simulation. This population consists of $N=1,000$ Finnish enterprises. For enterprise $k, k=1, \ldots, 1,000, y_{k}$ is number of employees (full time equivalents) $\times 10$, and $x_{k}$ is the wages paid by the enterprise to its employees, in thousands of FIM (Finnish Marks). The auxiliary information (wages paid) comes from the Finnish tax authority's VAT register. The employment variable is the one requiring estimation. The 1,000 units were selected (in an essentially random manner) from an original larger population of Finnish enterprises. Units with a value of zero either on $y_{k}$ or on $x_{k}$ were eliminated so that the simulation results would not be disturbed by extraneous factors. Consequently, as for the values $y_{k}$ and $x_{k}$, the population used in the simulation is a natural one, but because of the elimination of units, its features (mean, standard deviation, skewness, etc.) differ from those of the original larger population.

The population $y$-total to be estimated is $Y=\sum_{U} y_{k}=$ 169,168 . We fixed the expected sample size for the total sample, $s=s_{C} \cup s_{R}$, as $n=100$. The procedure described in section 3 was used to determine the certainty part $s_{C}$ of the sample. This resulted in a certainty part $s_{C}$ consisting of the largest $29=n_{c}$ units. The rest of the population, $U_{R}=U-s_{C}$ has the following descriptive characteristics: Its size is $N_{R}=1,000-29=971$; the total of $y$ is $Y_{R}=$ 46,138 (which equals $27 \%$ of the entire population total $Y$ $=169,168$ ); the coefficient of variation (standard deviation
divided by mean) is 1.78 for the variable $y$ and 1.94 for the variable $x$; the coefficient of correlation between $x$ and $y$ is 0.965 . The randomization part $s_{R}$, of expected size $n_{R}$ $=100-29=71$ units, is realized, in the simulation, by repeated Pomix sample selection from $U_{R}$. A plot of ( $x_{k}, y_{k}$ ) for the units $k$ in $U_{R}$ is shown in the Appendix.

To see the effect of the Bernoulli width, we carried out the simulation for a range of different $B$-values: $B=0,0.01$, ..., 0.07, and, in addition, $B=f_{R}=n_{R} / N_{R}=71 / 971=0.073$ (which gives Bernoulli). For each value of $B, 100 \times 100=$ $10,000 \mathrm{PRN} /$ sample pairs were realized, and the results were used to calculate, for each of the four point estimators, five Monte Carlo summary statistics. These are as follows, if $\hat{Y}$ denotes one of the four point estimators, $\hat{V}$ the corresponding variance estimator obtained from (4.2), and $\left(\hat{Y}-z_{1-\alpha / 2} \sqrt{ } \hat{V}, \hat{Y}+z_{1-\alpha / 2} \sqrt{ } \hat{V}\right)$ the corresponding confidence interval for $Y$ at the nominal confidence level 1- $\alpha$, where $z_{1-\alpha / 2}$ is the standard normal score, $z_{1-\mathrm{a} / 2}=1.960$ for $\alpha=5 \%$, and $z_{1-\alpha / 2}=1.645$ for $\alpha=10 \%$ :
(1) $\operatorname{MCE} \hat{Y}=$ the Monte Carlo expectation of the point estimator $\hat{Y}$, that is, the arithmetic mean of the 10,000 point estimates;
(2) $\operatorname{MCV} \hat{Y}=$ the Monte Carlo variance of the point estimator $\hat{Y}$, that is, the variance of the 10,000 point estimates;
(3) $\operatorname{MCE} \hat{V}=$ the Monte Carlo expectation of the variance estimator $\hat{V}$, that is, the arithmetic mean of the 10,000 variance estimates;
(4) MCRTE95 = Monte Carlo coverage rate for nominal $95 \%$ confidence intervals, that is, the number of times that the target parameter $Y$ is contained in the confidence interval, divided by 10,000 , and expressed in per cent;
(5) MCRTE90 = Monte Carlo coverage rate for nominal $90 \%$ confidence intervals; its definition is analogous to that of MCRTE95.

The simulation results are shown in Table 2 (Average sample size, Monte Carlo variance; Monte Carlo expectation of variance estimator) and in Table 3 (Monte Carlo coverage rates). The tables do not show MCE $\hat{Y}$, because in all cases this quantity was very close to the target parameter value $Y=169,168$, confirming that all estimators are essentially unbiased. The deviation of MCE $\hat{Y}$ from $Y$ was in all cases less than $0.14 \%$, in most cases considerably less. The average sample size over the 10,000 repetitions is seen to be very close to $n=100$, as it should.

Table 2
Results of Simulation Study for Different Bernoulli Widths B: Average Sample Size, MCV $\hat{Y}$ and MCE $\hat{V}$; est. $j$ Refers to Estimator $\hat{Y}_{j} ; j=1, \ldots, 4$. (Values in the Last Eight Columns to be Multiplied by $10^{6}$.)

| Bernoulli | Average | MCV $\hat{y}$ |  |  |  |  |  |  |  |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| width $B$ | sample size | est. 1 | est. 2 | est. 3 | est. 4 | est. 1 | est. 2 | est. 3 | est. 4 |
| 0.000 | 99.95 | 24.56 | 3.63 | 3.43 | 3.46 | 24.92 | 3.67 | 3.43 | 3.46 |
| 0.010 | 100.05 | 22.74 | 1.96 | 1.84 | 1.85 | 23.53 | 1.98 | 1.86 | 1.87 |
| 0.020 | 100.04 | 24.75 | 1.82 | 1.77 | 1.78 | 25.37 | 1.85 | 1.77 | 1.78 |
| 0.025 | 100.09 | 25.51 | 1.82 | 1.78 | 1.79 | 26.86 | 1.87 | 1.81 | 1.82 |
| 0.030 | 100.06 | 28.03 | 1.83 | 1.80 | 1.81 | 28.58 | 1.91 | 1.87 | 1.88 |
| 0.040 | 99.86 | 35.17 | 2.01 | 2.03 | 2.06 | 33.54 | 2.11 | 2.11 | 2.15 |
| 0.050 | 100.02 | 42.25 | 2.56 | 2.64 | 2.67 | 41.42 | 2.48 | 2.51 | 2.59 |
| 0.060 | 99.99 | 56.08 | 3.42 | 3.65 | 3.67 | 55.70 | 3.20 | 3.24 | 3.44 |
| 0.070 | 100.05 | 90.73 | 4.80 | 5.47 | 5.59 | 91.28 | 4.89 | 4.72 | 5.37 |
| 0.073 | 100.02 | 119.13 | 6.06 | 7.09 | 7.43 | 116.27 | 6.00 | 5.34 | 6.49 |

Table 3
Results of Simulation Study for Different Bernoulli Widths B: Coverage Rates in \% of Nominal $95 \%$ and $90 \%$ Confidence Intervals, MCRTE95 and MCRTE90; est. $j$ Refers to Estimator $\hat{Y}_{j} ; j=1, \ldots, 4$

| Bernoulli | nominal 95\% confidence level |  |  | nominal 90\% confidence level |  |  |  |  |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| width $B$ | est. 1 | est. 2 | est. 3 | est. 4 | est. 1 | est. 2 | est. 3 | est. 4 |
| 0.000 | 94.50 | 92.03 | 92.75 | 92.48 | 89.70 | 86.36 | 87.35 | 86.74 |
| 0.010 | 95.20 | 93.13 | 93.47 | 93.52 | 90.43 | 88.06 | 87.93 | 88.02 |
| 0.020 | 95.06 | 94.23 | 93.88 | 93.88 | 90.36 | 89.36 | 88.49 | 88.55 |
| 0.025 | 95.06 | 93.73 | 94.56 | 94.70 | 90.64 | 89.15 | 89.73 | 89.72 |
| 0.030 | 94.63 | 94.12 | 94.09 | 94.19 | 89.85 | 89.06 | 88.70 | 88.86 |
| 0.040 | 93.84 | 94.44 | 94.47 | 94.64 | 88.77 | 89.60 | 89.41 | 89.60 |
| 0.050 | 93.97 | 94.38 | 93.76 | 93.82 | 88.67 | 88.67 | 88.08 | 88.53 |
| 0.060 | 93.54 | 93.57 | 92.12 | 92.69 | 89.10 | 87.57 | 85.99 | 87.27 |
| 0.070 | 92.93 | 94.99 | 90.67 | 92.03 | 88.40 | 88.62 | 84.27 | 86.11 |
| 0.073 | 91.03 | 95.02 | 88.03 | 90.46 | 86.53 | 88.62 | 81.26 | 83.86 |

Tables 2 and 3 generate these comments:

1. Let us begin the examination of Table 2 by a comparison of Monte Carlo variances across estimators, for a fixed Bemoulli width $B$. This shows that, for every value of $B$, there is little to choose between $\hat{Y}_{2}, \hat{Y}_{3}$ and $\hat{Y}_{4}$, in terms of variance. By contrast, the HT estimator $\hat{Y}_{1}$ has considerably greater variance. To illustrate, the ratio $\operatorname{MCV} \hat{Y}_{3} / \mathrm{MCV} \hat{Y}_{1}$ equals $3.43 / 24.56=0.140$ for $B=0$ (Poisson $\pi \mathrm{ps}$ ), $1.80 / 28.03=0.064$ for $B=0.03$, and 7.09/119.13 $=0.060$ for $B=0.073$ (Bemoulli). This confirms that the HT estimator is a poor choice compared to an altemative that uses the strongly correlated auxiliary variable. This is true, not surprisingly, for Bernoulli, but also for $B$-values near the lower end of the $\left[0, f_{R}\right]$ interval, which shows that the sampling design alone does not extract all the power of the auxiliary variable, even though with $B$ near zero, we are close to a strict $\pi$ ps selection (thus supposedly highly efficient). Part of the reason that the HT estimator has a
comparatively large variance is that the randomness of the sample size under Pomix sampling penalizes the HT estimator (but not the GREG estimators). Since the HT estimator is inefficient, we do not further discuss it.
2. Examining the small differences between $\hat{Y}_{2}, \hat{Y}_{3}$ and $\hat{Y}_{4}$, we note in Table 2: As measured by the Monte Carlo variance, $\hat{Y}_{3}$ is better than $\hat{Y}_{4}$ for all Bernoulli widths $B$, but only marginally so. Also, $\hat{Y}_{3}$ and $\hat{Y}_{4}$ are better than $\hat{Y}_{2}$ at the lower end of the range of $B$-values, possibly because of the fact that in $\widehat{Y}_{2}$ we allow the certainty part the sample to contribute to the slope estimate, somewhat inappropriately, since there is only an estimation problem for the randomization part. But at the upper end, the relation is reversed and for the upper extreme $B=0.073$ (Bernoulli), $\hat{Y}_{2}$ is clearly better than $\widehat{Y}_{3}$. That the differences between $\hat{Y}_{2}, \hat{Y}_{3}$ and $\widehat{Y}_{4}$ are so small is not surprising, because all are varieties of the GREG estimator (4.1) using essentially the same auxiliary information.
3. Table 2 confirms that the proposed variance estimator $\hat{V}$ works well, as we would expect; MCE $\hat{V}$ is with few exceptions very close to the target that $\hat{V}$ aims at estimating, that is, the variance of $\hat{Y}$, measured here by MCV $\hat{Y}$. This holds for all estimators and all values of $B$, with a few notable exceptions, namely, in the case of $\hat{Y}_{3}$ and $\boldsymbol{Y}_{4}$ when $B$ is close to the upper extreme (Bernoulli). Then the variance estimator underestimates the variance.
4. The most interesting result in Table 2 we consider to be the fact that the variance of $\hat{Y}_{2}$ or $\hat{Y}_{3}$ or $\hat{Y}_{4}$, when viewed as a function of the Bernoulli width $B$, does not attain its minimum at $B=0$ (Poisson $\pi \mathrm{ps}$ ), as one might have initially guessed, but rather for a value of $B$ somewhere between 0.02 and 0.03 . Moreover, the improvement of the case $B=0.02$ over the case $B=0$ is substantial for all of $\hat{Y}_{2}, \hat{Y}_{3}$ and $\hat{Y}_{4}$. Measuring this improvement by $\operatorname{MCV}(\hat{Y} \mid B=0.02)$ divided by $\operatorname{MCV}(\mathcal{Y} \mid B=0)$, we find that this ratio is only about $50 \%$ for all of $\hat{Y}_{2}, \hat{Y}_{3}$ and $\hat{Y}_{4}$. More precisely, for $\hat{Y}_{2}$ the ratio is $1.82 / 3.63=0.501$, for $\hat{Y}_{3}$ it is $1.77 / 3.43=0.516$, and for $\hat{Y}_{4}$ it is $1.78 / 3.46=0.514$. In view of these results, we added a simulation for $B=0.025$, a value not examined in the original round of simulations. The results, also displayed in Table 1, confirm that a minimum variance is obtained, for all three estimators, $\hat{Y}_{2}, \hat{Y}_{3}$ and $\hat{Y}_{4}$, at a point in the vicinity of $B=0.025$. One possible explanation of why it is considerably better to take $B$ to be a value distinctly greater than $B=0$ (which gives Poisson $\pi \mathrm{ps}$ ) is the following: When $B$ is (or very near 0 ), the units with the smallest $x$-values, when selected, will have unduly large weights, which induces high variability. This is avoided by choosing $B$ clearly away from zero.
5. The Monte Carlo results in Table 3 concerning the coverage rates show that the variance estimation and the confidence interval procedure function to satisfaction. As theory leads us to expect, MCRTE95 and MCRTE90 are close, for all four estimators, to their theoretical values, $95 \%$ and $90 \%$, respectively. Only for and $\hat{Y}_{3}$ and $\hat{Y}_{4}$, when $B$ gets close to the upper extreme (Bernoulli), do we notice any marked tendency for the MCRTE to drop below the nominal value, resulting in part from the underestimation of variance mentioned earlier.

## 6. FURTHER EVIDENCE THAT POMIX SAMPLING IS MORE EFFICIENT THAN POISSON $\pi$ ps SAMPLING

Initially we had no strong reason to believe that Pomix sampling combined with a GREG estimator would be more efficient for some Bernoulli widths $B$ in the interior of $\left[0, f_{R}\right]$ than for Poisson $\pi \mathrm{ps}(B=0)$. The strong improvement - a variance reduction of around $50 \%$ for our
particular population - was rather surprising. For other populations, the variance reduction can be more or less than the $50 \%$ we found. Because our finding is data dependent, it is desirable to provide some more general evidence in support of the proposition that Pomix sampling with a $B$-value well into the interior of $\left[0, f_{R}\right]$ is better than Poisson $\pi$ ps sampling ( $B=0$ ). We now present some evidence of this kind.

We examined the Taylor variance of $\hat{Y}$ (that is, the variance of the Taylor linearized statistic). It is given by (see Särndal et al., 1992, Ch. 6) $V_{\text {TAY }}=\sum_{U_{R}}\left(a_{k}-1\right) E_{k}^{2}$, where $E_{k}$ is the population analogue of the sample based residual $e_{k}$ used in the variance estimator (4.2). For example, for the estimator $\hat{Y}_{3}$, the residual in question is $E_{k}=y_{k}-b_{3} x_{k}$ with $b_{3}=\sum_{U_{R}}\left(a_{k}-1\right) y_{k} x_{k} / \sum_{U_{R}}\left(a_{k}-1\right) x_{k}^{2}$; for $\hat{Y}_{4}, b_{4}=\sum_{U_{R}} y_{k} / \sum_{U_{R}} x_{k}$ replaces $b_{3}$.

It is reasonable to model the squared residual as $E_{k}^{2}=\sigma^{2} x_{k}^{p}\left(1+\delta_{k}\right)$, where $p$ satisfies $0 \leq p \leq 2$, and $\delta_{k}$ is near zero. This corresponds to assuming a superpopulation model $y_{k}=x_{k}{ }^{\prime} \beta+\varepsilon_{k}$, where the $\varepsilon_{k}$ are independent errors with model expected value zero for every $k$, and $\varepsilon_{k}$ has model variance $\sigma^{2} x_{k}^{p}$. Using the approximation $E_{k}^{2^{k}} \approx \sigma^{2}$ $x_{k}^{p}$, and that $a_{k}=1 / \pi_{k}$ with $\pi_{k}$ given by (2.3), we have

$$
V_{\mathrm{TAY}} \approx \sigma^{2} \sum_{U_{R}}\left(a_{k}-1\right) x_{k}^{p}=\sigma^{2}\{H(B, p)-T(p)\}
$$

where $H(B, p)=\bar{x}_{U_{R}} \sum_{U_{N}} x_{k}^{p}\left[B \bar{x}_{U_{R}}+\left(f_{R}-B\right) x_{k}\right]^{-1}$ and $T(p)=\sum_{U_{R}} x_{k}^{p}$. Now consider a fixed value of $p$ such that $0 \leq p \leq 2$. ${ }^{\text {K }}$ We want to find out if $H(B, p)$ has a smaller value for some $B$ in the interior of the interval $\left[0, f_{R}\right]$, compared to its value at $B=0$, which is $H(0, p)$. To this end, let us examine if the derivative $H^{\prime}(B, p)=$ $\partial H(B, p) / \partial B$ is negative at $B=0$. We find

$$
H^{\prime}(B, p)=\bar{x}_{U_{R}} \sum_{U_{R}} x_{k}^{p}\left(x_{k}-\bar{x}_{U_{R}}\right)\left[B \bar{x}_{U_{R}}+\left(f_{R}-B\right) x_{k}\right]^{-2}
$$

Its value at $B=0$ is $H^{\prime}(0, p)=\left(\bar{x}_{U_{R}} / f_{R}^{2}\right) \sum_{U_{R}} x_{k}^{p-2}\left(x_{k}-\bar{x}_{U_{R}}\right)$. The sign of $H^{\prime}(0, p)$ is the same as that of $\sum_{U_{R}} x_{k}^{p-2}\left(x_{k}-\bar{x}_{U_{\mathcal{R}}}\right)$. But this quantity equals, apart from the factor $1 /\left(N_{R}-1\right)$, the covariance in $U_{R}$ between $x_{k}^{p-2}$ and $x_{k}-\bar{x}_{U_{R}}$ (note that $x_{k}-\bar{x}_{U_{R}}$ has zero mean). When $p$ satisfies $0 \leq p<2$, this covariance is negative: when $x_{k}$ increases, $x_{k}-\bar{x}_{U_{R}}$ increases steadily, and $x_{k}^{p-2}$ decreases steadily (and remains always positive). The sign of $H^{\prime}(0, p)$ is therefore negative; consequently, it is not at $B=0$ that $H(B, p)$ attains its minimum value, but for some $B$ in the interior of $\left[0, f_{R}\right]$. Now for $p=2, H^{\prime}(0, p)=0$, and $H(B, p)$ has a minimum at $B=0$.

These considerations raise the question whether the population used for our simulation in section 5 corresponds to a value of $p \in[0,2]$, but distinctly less than 2 , so that we can expect significant gains from Pomix sampling. To obtain an answer, we estimated $p$ by fitting the logarithmic version of the model $E_{k}^{2}=\sigma^{2} x_{k}^{p}\left(1+\delta_{k}\right)$ to the data available for $U_{R}=U-U_{C}$. That is, we fitted $w_{k}=a+p z_{k}$,
where $w_{k}=\log E_{k}^{2} ; E_{k}=y_{k}-b_{4} x_{k}$ with $b_{4}=\sum_{U_{R}} y_{k} / \sum_{U_{R}}$ $x_{k} ; z_{k}=\log x_{k}$, and a is an intercept term. We obtained the value $p=1.45$, by treating $p$ as a linear regression slope estimated as $p=\sum_{U_{R}}\left(w_{k}-\bar{w}_{U_{R}}\right)\left(z_{k}-\bar{z}_{U_{R}}\right) / \sum_{U_{R}}\left(z_{k}-\bar{z}_{U_{R}}\right)^{2}$. Since this $p$-value is considerably less than $2^{R}$, our Monte Carlo population is indeed one where one can expect significant gains from the use of Pomix sampling with a value of $B$ in the interior of $\left[0, f_{R}\right]$.

## 7. CONCLUDING DISCUSSION AND TENTATIVE RECOMMENDATIONS

The survey sampler will ask: If I consider using Pomix sampling for my survey, combined with a GREG estimator, what is an appropriate choice for $B$ ? Recall that in this paper we found, for one particular population, that a large efficiency gain (roughly $50 \%$ variance reduction compared to Poisson $\pi \mathrm{ps}$ ) is realized by fixing the Pomix parameter $B$ at around $30 \%$ of $f_{R}$. We were led to suspect that the variance gain is related to residual error characteristics, and this was confirmed in Section 6 which presented evidence that when the squared residual pattern conforms to $E_{k}^{2}=\sigma^{2} x_{k}^{p}$, where $p$ satisfies $0 \leq \mathrm{p}<2$, as is the case in many business survey populations, then Pomix sampling with $B$ in the interior of the interval $\left[0, f_{R}\right]$ may well be
advantageous. However, the present paper does not address the question of the optimal choice of $B$. A difficulty is that in practice the value of $B$ must be fixed at the design stage, and that the optimal $B$ depends on unknown population characteristics. Prior knowledge of the population, notably about its residual variance structure, can guide the choice of $B$.

Our tentative recommendations based on this paper are: If prior information suggests a squared residual pattern conforming to $\sigma^{2} x_{k}^{p}$ with $p<2$, then use Pomix sampling with $B=0.3 f_{R}$. On the other hand, if in reality the unknown $p$ is such that $p>2$, then, although the best choice in this case might be $B=0$ (Poisson $\pi \mathrm{ps}$ ), little harm would probably be done to use $B=0.3 f_{R}$, because the variance viewed as a function of $B$ is likely to increase at a gentle rate. Therefore, $B=0.3 f_{R}$, seems a reasonable all-purpose suggestion. These recommendations are tentative; the question merits a further study that lies beyond the scope of this paper.

## ACKNOWLEDGEMENTS

The authors gratefully acknowledge the work of a referee, whose suggestions led to valuable improvements in the original manuscript.

## APPENDIX



Figure 3. Scatter plot of $x$ (yearly wages) against $y$ (employment) for the portion $U_{R}$ of the Monte Carlo population of 1,000 Finnish enterprises.

## REFERENCES

ATMER, J., THULIN, G., and BÄCKLUND, S. (1975). Coordination of samples with the JALES technique. Statistisk Tidskrift, 13, 443-450.
BREWER, K.R.W., EARLY, L.J., and JOYCE, S.F. (1972). Selecting several samples from a single population. Australian Journal of Statistics, 14, 231-239.

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

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

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

SUNTER, A.B. (1977). Response burden, sample rotation, and classification renewal in economic surveys. International Statistical Review, 45, 209-222.

# Toward Variances for X-11 Seasonal Adjustments 

WILLIAM R. BELL and MATTHEW KRAMER ${ }^{1}$


#### Abstract

We develop an approach to estimating variances for X-11 seasonal adjustments that recognizes the effects of sampling error and errors from forecast extension. In our approach, seasonal adjustment error in the central values of a sufficiently long series results only from the effect of the X-11 filtering on the sampling errors. Towards either end of the series, we also recognize the contribution to seasonal adjustment error from forecast and backcast errors. We extend the approach to produce variances of errors in X-11 trend estimates, and to recognize error in estimation of regression coefficients used to model, e.g., calendar effects. In empirical results, the contribution of sampling error often dominated the seasonal adjustment variances. Trend estimate variances, however, showed large increases at the ends of series due to the effects of fore/backcast error. Nonstationarities in the sampling errors produced striking patterns in the seasonal adjustment and trend estimate variances.


KEY WORDS: Sampling error; Forecast error; Trading-day; ARIMA model.

## 1. INTRODUCTION

The problem of how to obtain variances for seasonally adjusted data is long-standing (President's Committee to Appraise Employment and Unemployment Statistics 1962). Model-based methods of seasonal adjustment (see Bell and Hillmer 1984, for a discussion) use results from signal extraction theory to produce estimates and associated error variances of the seasonal and nonseasonal components. Most official seasonal adjustments, however, are made using empirical methods, most notably X-11 (Shishkin, Young and Musgrave 1967) or X-11-ARIMA (Dagum 1975). These methods are based on fixed filters, not models, and so it is not obvious how to calculate variances of the seasonal adjustment errors. Various approaches for obtaining variances for $\mathrm{X}-11$ seasonal adjustments have been proposed, as summarized below.

Wolter and Monsour (1981) suggested two approaches. They recognized that many time series that are seasonally adjusted are estimates from repeated sample surveys, and thus are subject to sampling error. Their first approach accounts only for the effect of sampling error on the variance associated with seasonal adjustments. Their second approach tries to also reflect uncertainty due to stochastic time series variation in the seasonal adjustment variances. However, this second approach assumes that, apart from regression terms, the time series is stationary. This type of model is now seldom used for seasonal time series. Also, their second approach contains a conceptual error: it produces the variance of the seasonally adjusted estimate, instead of the desired variance of the error in the seasonally adjusted estimate.

Burridge and Wallis (1985) investigated use of the steady-state Kalman filter for calculation of model-based seasonal adjustment variances, and applied this approach to
a model they obtained previously (Burridge and Wallis 1984) for approximating the X-11 filters. They suggested that this approach could be used to, "provide measures of the variability of the $\mathrm{X}-11$ method when it is applied to data for which it is optimal," (p. 551), but cautioned against doing this when the $\mathrm{X}-11$ filter would be suboptimal (i.e., very different from the optimal model-based filter). Hausman and Watson (1985) suggested an approach to estimating the mean squared error for $\mathrm{X}-11$ when it is used in suboptimal situations. Bell and Hillmer (1984, section 4.3.4) pointed out a problem with the use of model-based approximations to $\mathrm{X}-11$ for calculating seasonal adjustment variances. The problem is that X-11 filters (or any seasonal adjustment filter, for that matter) are not sufficient to uniquely determine models for the observed series and its components.

Pfeffermann (1994) developed an approach that recognizes the contributions of sampling error and irregular variation (time series variation in the irregular component) to $\mathrm{X}-11$ seasonal adjustment variances. The properties of the combined error (sampling error plus irregular) are estimated using the X-11 estimated irregular. These properties are then used to estimate two types of seasonal adjustment variances. A drawback to this approach is that it relies on an assumption that the X-11 adjustment filter annihilates the seasonal component and reproduces the trend component. (Note Pfeffermann (1994, p. 90), discussion surrounding equation (2.7).) Violations of this assumption in practice compromise the approach to an extent which appears difficult to assess. Thus, this assumption seems to us highly questionable and also, in any particular case, uncheckable. A second drawback is that one of the variance types proposed by Pfeffermann assumes that the X-11 seasonally adjusted series, rather than the trend estimate, is taken as an estimate of the trend. Breidt

[^1](1992) and Pfeffermann, Morry and Wong (1993) further develop Pfeffermann's general approach.

The goal of this paper is the development and application of an approach to obtaining variances for $\mathrm{X}-11$ seasonal adjustments accounting for two sources of error. The first error source is sampling error. The second is error that arises from the need to extend the time series with forecasts and backcasts before applying the symmetric X-11filters. These latter errors lead to seasonal adjustment revisions (Pierce 1980). Note that revisions eventually vanish as sufficient data beyond the time point being adjusted become available. Also note that a seasonally adjusted series will not contain sampling error if the corresponding unadjusted series does not. This is the case for certain economic time series, e.g., export and import statistics for most countries.

Our approach assumes that the $\mathrm{X}-11$ seasonal adjustment target (what we assume application of X-11 is intended to estimate) is what would result from application of the symmetric linear X-11 filter (with no forecast and backcast extension required) if the series contained no sampling error. While this definition of target might be criticized for ignoring time series variation in the underlying seasonal and nonseasonal components, we think this may be appropriate for typical users of X-11 seasonally adjusted data. Such users are most likely to be concerned about uncertainty reflected in differences between initial adjustments and final adjustments, i.e., in revisions. Some of these users will also be aware that the unadjusted series consists of sample-based estimates of the true underlying population quantities, and will realize that the effects of sampling error on adjustments should also be reflected in seasonal adjustment variances.

Our development is based on use of the symmetric linear $\mathrm{X}-11$ filters. We assume that the symmetric filters are applied to the series extended with minimum mean squared error forecasts and backcasts. In practice, the forecasts and backcasts are obtained from a fitted time series model. This is in the spirit of the X-11-ARIMA method of Dagum (1975), but with full forecast and backcast extension, as recommended by Geweke (1978), Pierce (1980), and Bobbitt and Otto (1990). Our results apply directly to the use of additive or log-additive X-11 (with forecast and backcast extension), and the log-additive results are assumed to apply approximately (Young 1968) to multiplicative $\mathrm{X}-11$.

Section 2 of this paper develops our approach, which builds on the first approach of Wolter and Monsour (1981). The differences between the two approaches are discussed in section 2.4. Section 3 then discusses three extensions to the results of section 2 . The first is to note that our approach works equally well with seasonal, trend, or irregular estimates, and that more generality is easily accommodated by allowing different filter choices for different months. The second extension produces variances of estimates of month-to-month or year-to-year change. Finally, when seasonal adjustment involves estimation of regression
effects (e.g., for trading-day or holiday variation), the results are extended to allow for additional variance due to error in estimating the regression parameters.

Section 4 then presents several examples illustrating the basic approach and the extensions given in section 3 . One thing evident from the examples is that for time series with sampling error, our seasonal adjustment variances will often be dominated by the contribution of the sampling error. In the center of the series, our results effectively reduce to the first approach results of Wolter and Monsour. Our results do differ from those of Wolter and Monsour near the end of the series. This is important since the most recent seasonally adjusted values receive the most scrutiny. Also, the contribution of forecast and backcast error to trend estimate variances can be very large at the ends of a time series. Other results of particular interest are the effects of certain nonstationarities in the sampling errors. The examples of section 4 show that nonstationarities such as sampling error variances that change over time, or periodic independent redrawings of the sample, can yield striking changes in the pattern of the variances of seasonally adjusted data or trend estimates over time.

Section 5 provides concluding remarks.

## 2. METHODOLOGY

Define the observed unadjusted time series as $y_{t}$ for $t=1, \ldots, n$. Time series that are seasonally adjusted are often estimates obtained from repeated (monthly or quarterly) sample surveys, and thus can be viewed as composed of a true underlying time series $Y_{i}$, and a series of sampling errors $e_{t}$ assumed uncorrelated with $Y_{t}$. (See Bell and Hillmer 1990.) In vector notation, $\mathbf{y}_{o}=\mathbf{Y}_{o}+\boldsymbol{e}_{o}$, where the subscript $o$ indicates that the time span of these vectors is the set of observed time points $1, \ldots, n$. In certain cases $y_{t}$ may arise from repeated censuses (as is typically the case for national export and import statistics, for example), in which case there is no sampling error, i.e., $e_{t}=0$.

The development that follows assumes that both $Y_{t}$ and $e_{t}$ follow known time series models. The model for $Y_{t}$ will generally involve differencing, as in ARIMA (auto-regressive-integrated-moving average) and ARIMA component (structural) models. The model for $Y_{t}$ may be extended to include regression terms. (This will be considered in section 3.3.) The series $e_{t}$ is assumed to not require differencing, but it may nonetheless exhibit certain nonstationarities, such as variances that change over time. Any such nonstationarities are assumed to be accounted for in the model for $e_{t}$. In practice, the models will be developed from observed data, as is discussed by, e.g., Bell and Hillmer (1990, submitted), Binder and Dick (1989, 1990), and Tiller (1992).

In applying a symmetric X-11 filter of length $2 m+1$ for seasonal adjustment with full forecast and backcast extension, the vector $y_{o}$ needs to be augmented by $m$ backcasts and $m$ forecasts. The vector holding the $m$ values
of $y_{t}$ prior to the observed data, and the corresponding $m \times 1$ vectors for $Y_{t}$, and $e_{t}$, are denoted $\mathbf{y}_{b}, \mathbf{Y}_{b}$, and $\mathbf{e}_{b}$. The analogous vectors of the $m$ future values of $y_{t}, Y_{t}$, and $e_{t}$ are denoted $\mathbf{y}_{f}, \mathbf{Y}_{f}$, and $\mathbf{e}_{f}$. Thus,

$$
\left(\begin{array}{l}
\mathbf{y}_{b}  \tag{2.1}\\
\mathbf{y}_{o} \\
\mathbf{y}_{f}
\end{array}\right)=\left(\begin{array}{l}
\mathbf{Y}_{b} \\
\mathbf{Y}_{o} \\
\mathbf{Y}_{f}
\end{array}\right)+\left(\begin{array}{c}
\mathbf{e}_{b} \\
\mathbf{e}_{o} \\
\mathbf{e}_{f}
\end{array}\right)
$$

The full vectors in (2.1), hereafter denoted as $\mathbf{y}, \mathbf{Y}$, and $\mathbf{e}$, have length $n+2 m$.

The backcasts and forecasts used to augment $\mathbf{y}_{o}$ are assumed to be minimum mean squared error (MMSE) linear predictions of $\mathbf{y}_{b}$ and $\mathbf{y}_{f}$ (using $\mathbf{y}_{o}$ ) obtained from the known time series model. (In practice, the model will be fitted to the data $y_{o}$.) Under normality, the backcasts and forecasts are $E\left(\mathbf{y}_{b} \mid \mathbf{y}_{o}\right)$ and $E\left(\mathbf{y}_{f} \mid \mathbf{y}_{o}\right)$. The vector of observed data augmented with the backcasts and forecasts is denoted $\hat{\mathbf{y}}=\left(\hat{\mathbf{y}}_{b}^{\prime}, \mathbf{y}_{o}^{\prime}, \hat{\mathbf{y}}_{f}{ }^{\prime}\right)^{\prime}$, where $\hat{\mathbf{y}}_{b}=E\left(\mathbf{y}_{b} \mid \mathbf{y}_{o}\right)$ and $\hat{\mathbf{y}}_{f}=E\left(\mathbf{y}_{f} \mid \mathbf{y}_{o}\right)$. To simplify notation, from now on we will take expressions such as ( $\hat{\mathbf{y}}_{b}, \mathbf{y}_{o}, \hat{\mathbf{y}}_{f}$ ) to mean the column vector $\left(\hat{\mathbf{y}}_{b}^{\prime}, \mathbf{y}_{o}^{\prime}, \hat{\mathbf{y}}_{f}^{\prime}\right)^{\prime}$.

Let the linear symmetric $\mathrm{X}-11$ seasonal adjustment filter be written $\omega(B)=\sum_{-m}^{m} \omega_{j} B^{j}$, where $B$ is the backshift operator and the $\omega_{j}$ are the filter weights ( $\omega_{j}=\omega_{-j}$ ). Calculation of the $\omega_{j}$ is discussed by Young (1968) and Wallis (1982). Results of Bell and Monsell (1992) were used here. Application of $\omega(B)$ to the forecast and backcast extended series can be written as $\boldsymbol{\Omega} \hat{\mathbf{y}}$, where $\boldsymbol{\Omega}$ is a matrix of dimension $n \times(n+2 m)$. Each row of $\boldsymbol{\Omega}$ contains the filter weights ( $\omega_{-m}, \ldots, \omega_{0}, \ldots, \omega_{m}$ ), preceded and followed by the appropriate number of zeroes such that the center weight of the X-11 filter ( $\omega_{0}$ ) multiplies the observation being adjusted. Thus, in the first row of $\boldsymbol{\Omega}$ there are no preceding zeroes and $n-1$ trailing zeroes, in the second row there is one preceding zero and $n-2$ trailing zeroes, etc. For the default X-11 filter, $m=84$. Choice of alternative seasonal or trend moving averages in $\mathrm{X}-11$ changes the value of $m$ from a low of 70 to a high of 149 .

The question arises as to what $\boldsymbol{\Omega} \hat{\mathbf{y}}$ is estimating. As noted in the introduction, we define the "target" of the seasonal adjustment as the adjusted series that would result if there were no sampling error and there were sufficient data before and after all time points of interest for the symmetric filter to be applied. The target is thus $\omega(B) Y_{t}$, or in vector notation $\boldsymbol{\Omega} \mathbf{Y}$, and the seasonal adjustment error vector is $\mathbf{v}=\boldsymbol{\Omega}(\mathbf{Y}-\hat{\mathbf{y}})$. We are interested in the variancecovariance matrix $\operatorname{var}(\mathbf{v})=\boldsymbol{\Omega} \operatorname{var}(\mathbf{Y}-\hat{\mathbf{y}}) \boldsymbol{\Omega}^{\prime}$. This can be easily computed once $\operatorname{var}(\mathbf{Y}-\hat{\mathbf{y}})$ is obtained. From here through section 2.3 we discuss the calculation of $\operatorname{var}(\mathbf{Y}-\hat{\mathbf{y}})$.

We start by writing $\mathbf{Y}-\hat{\mathbf{y}}=(\mathbf{y}-\mathbf{e})-\hat{\mathbf{y}}=(\mathbf{b}, \mathbf{0}, \mathbf{f})-\mathbf{e}$, where $\mathbf{b}=\mathbf{y}_{b}-\hat{\mathbf{y}}_{b}$ is the $m \times 1$ vector of backcast errors, and $\mathbf{f}=\mathbf{y}_{f}-\hat{\mathbf{y}}_{f}$ is the $m \times 1$ vector of forecast errors. Given the models for $Y_{t}$ and $e_{t}$, we calculate $\operatorname{var}(\mathbf{Y}-\hat{\mathbf{y}})$ by separately computing $\operatorname{var}(\mathbf{e}), \operatorname{var}(\mathbf{b}, \mathbf{0}, \mathbf{f})$, and
$\operatorname{cov}[(\mathbf{b}, \mathbf{0}, \mathbf{f}), \mathbf{e}]$, as discussed in sections 2.1 to 2.3. Then, $\operatorname{var}(\mathbf{Y}-\hat{\mathbf{y}})$ easily follows as $\operatorname{var}(\mathbf{b}, \mathbf{0}, \mathbf{f})+\operatorname{var}(\mathbf{e})-\operatorname{cov}$ $[(\mathbf{b}, \mathbf{0}, \mathbf{f}), \mathbf{e}]-\operatorname{cov}[(\mathbf{b}, \mathbf{0}, \mathbf{f}), \mathbf{e}]^{\prime}$. Thus,

$$
\begin{gathered}
\operatorname{var}(\mathbf{v})=\boldsymbol{\Omega}(\operatorname{var}(\mathbf{b}, \mathbf{0}, \mathbf{f})+\operatorname{var}(\mathbf{e})-\operatorname{cov}[(\mathbf{b}, \mathbf{0}, \mathbf{f}), \mathbf{e}]- \\
\left.\operatorname{cov}[(\mathbf{b}, \mathbf{0}, \mathbf{f}), \mathbf{e}]^{\prime}\right\} \mathbf{\Omega}^{\prime} .
\end{gathered}
$$

Example-U.S. 5+ Unit Housing Starts. As the computations for each piece of $\operatorname{var}(\mathbf{Y}-\hat{\mathbf{y}})$ are explained, we illustrate the results graphically for an example series: housing starts in the U.S. for buildings of five or more units from January 1975 through November 1988 ( 167 observations). The original series, seasonally adjusted series, and estimated trend are shown in Figure 1. In practice, seasonal adjustment at the Census Bureau of this series uses a multiplicative decomposition with a $3 \times 9$ seasonal moving average and a 13 -term Henderson trend filter. The following model for this series was developed in Bell and Hillmer (submitted):

$$
y_{t}=Y_{t}+e_{t}
$$

$$
(1-B)\left(1-B^{12}\right) Y_{t}=\left(1-0.67 B+0.36 B^{2}\right)\left(1-.8753 B^{12}\right) a_{t}
$$

$$
\sigma_{a}^{2}=0.0191
$$

$$
\begin{equation*}
e_{t}=\left(1-0.11 B-0.10 B^{2}\right) b_{t}, \sigma_{b}^{2}=0.00714 \tag{2.2}
\end{equation*}
$$



Figure 1. U.S. Housing Starts with Five or more Units. The top panel gives the original series from January 1975 through November 1988. The strong seasonality of the series is apparent from the yearly dips that typically occur during the winter months. The bottom panel gives the X-11 seasonally adjusted series (solid line) and trend estimate (dotted line) for the same period. The seasonal adjustment is multiplicative using a $3 \times 9$ seasonal moving average and 13 -term Henderson trend filter.

Here, $y_{t}$ denotes the logarithms of the original time series $\left(e^{y_{1}}\right)$, so that (2.2) implies a multiplicative decomposition for the original series ( $e^{y_{t}}=e^{Y_{t}} e^{e_{t}}$ ).

### 2.1 Computation of $\operatorname{Var}(\mathrm{e})$

If $e_{l}$ follows a stationary ARMA model, then var(e) can be computed from standard results, e.g., McLeod (1975, 1977), Wilson (1979). If $\operatorname{var}\left(e_{t}\right)$ changes over time, we write $e_{t}=h_{t} \bar{e}_{t}$, where $h_{t}^{2}=\operatorname{var}\left(e_{t}\right)$, and $\tilde{e}_{t}$ has variance one and the same autocorrelation function as $e_{t}$. (See Bell and Hillmer submitted.) Then, writing $\mathbf{e}=\mathbf{H e}$, where $\mathbf{H}=$ $\operatorname{diag}\left(h_{1-m}, \ldots, h_{n+m}\right)$, we have $\operatorname{var}(\mathbf{e})=\mathbf{H} \operatorname{var}(\tilde{\mathbf{e}}) \mathbf{H}^{\prime}$. $\operatorname{Var}(\tilde{\mathbf{e}})$ is the autocorrelation matrix of $\mathbf{e}$, and it can be computed as just noted using the model for $\bar{e}_{t}$.

If the sample is independently redrawn at certain times, then var (e) will be block diagonal, with blocks corresponding to the time points when each distinct sample is in effect. Each diagonal block of var (e) be computed as just discussed. These two types of nonstationarities in $e_{i}^{-}$ variance changing over time and "covariance breaks" due to independent redrawings of the sample - are those that arise in the examples of section 4.
Example-U.S. 5+ Unit Housing Starts (continued). Autocovariances for the MA(2) model for $e_{t}$ given in (2.2) are easily computed. The resulting var (e) is a band matrix, with $\operatorname{var}\left(e_{\mathrm{f}}\right)=0.007298$ on the diagonal, $\operatorname{cov}\left(e_{\mathrm{r}}, e_{\mathrm{f}-1}\right)=$ -0.000707 on the first sub- and super-diagonals, and $\operatorname{cov}\left(e_{t}, e_{t-2}\right)=-0.000714$ on the second sub- and superdiagonals. The rest of var (e) is zero. Following pre- and post multiplication by the seasonal adjustment filter matrices $\boldsymbol{\Omega}$ and $\boldsymbol{\Omega}^{\prime}$ the contribution of the sampling error to the variance of the seasonally adjusted series is constant for each observation (Figure 2). This occurs because the result of a time invariant linear filter applied to a stationary series $\left(\omega(B) e_{t}\right)$ is a stationary series, which has a constant variance.

### 2.2 Computation of Var ( $\mathbf{b}, \mathbf{0}, \mathrm{f}$ )

The central $n$ rows and $n$ columns of $\operatorname{var}(\mathbf{b}, \mathbf{0}, \mathbf{f})$ are all zeroes. We require computation of $\operatorname{var}(\mathbf{b})$, $\operatorname{var}(\mathbf{f})$, and $\operatorname{cov}(\mathbf{b}, \mathbf{f})$ for the comer blocks of $\operatorname{var}(\mathbf{b}, \mathbf{0}, \mathbf{f})$. Although computation of variances of forecast (or backcast) errors for given models is standard in time series analysis, it is complicated here by the component representation of $y_{t}$ as $Y_{t}+e_{t}$, and by differencing in the model for $Y_{t}$. Although computations for such models are often handled by the Kalman filter (Bell and Hillmer submitted; Binder and Dick 1989, 1990; Tiller 1992), this is inconvenient here since we require covariances of all distinct pairs of random variables from among the $m$ forecast and $m$ backcast errors. We instead use a direct matrix approach due to Bell and Hillmer (1988).

Assume that the differencing operator required to render $Y_{t}$ stationary is $\delta(B)$, which is of degree $d$. Since $e_{t}$ is assumed not to require differencing, $\delta(B)$ is also the differencing operator required by $y_{t}$. Define $\delta(B) y_{t}=w_{t}$,
thus $w_{t}=\delta(B) Y_{t}+\delta(B) e_{t}$. We introduce the matrix $\Delta$, corresponding to $\delta(B)$, defined such that $\Delta \mathbf{y}=\mathbf{w}$ is the vector of differenced $\mathbf{y}$. The vector $\mathbf{w}=\left(\mathbf{w}_{b}, \mathbf{w}_{o}, \mathbf{w}_{f}\right)$, which is of length $n+2 m-d$, is partitioned so that $\mathbf{w}_{b}$ and $\mathbf{w}_{f}$ are $m \times 1$ vectors, and $\mathbf{w}_{o}$ is the $n-d$ vector of differenced observed data. Thus, $\boldsymbol{\Delta}$ has dimensions $(n+2 m-d) \times(n+2 m)$. Note that, because $d$ observations are lost in differencing, $\mathbf{w}_{b}$ and $\mathbf{w}_{o}$ start $d$ time points later than $\mathbf{y}_{b}$ and $\mathbf{y}_{o}$, respectively. That is, $\mathbf{y}_{b}$ and $\mathbf{y}_{o}$ start at time points $1-m$ and 1 , but $\mathbf{w}_{b}$ and $\mathbf{w}_{o}$ start at time points $1-m+d$ and $d+1$.


[^2]Define $\mathbf{u}=\left(u_{1-m+d}, \ldots, u_{n+m}\right)^{\prime}=\mathbf{\Delta Y}$. The time series $u_{t}$ is stationary. Since $\mathbf{w}=\mathbf{u}+\Delta \mathbf{e}$, with $\mathbf{u}$ and $\mathbf{e}$ uncorrelated with each other, $\operatorname{var}(\mathbf{w})=\operatorname{var}(\mathbf{u})+\Delta \operatorname{var}(\mathbf{e}) \Delta^{\prime}$. We partition var (w) as

$$
\operatorname{var}(\mathbf{w})=\left(\begin{array}{lll}
\Sigma_{11} & \Sigma_{12} & \Sigma_{13} \\
\Sigma_{21} & \Sigma_{22} & \Sigma_{23} \\
\Sigma_{31} & \Sigma_{32} & \Sigma_{33}
\end{array}\right),
$$

where $\boldsymbol{\Sigma}_{11}$ is $\operatorname{var}\left(\mathbf{w}_{b}\right), \boldsymbol{\Sigma}_{12}$ is $\operatorname{cov}\left(\mathbf{w}_{b}, \mathbf{w}_{o}\right)$, etc.
Since $\mathbf{y}$, when differenced to $\mathbf{w}$ using $\delta(B)$, has lost $d$ data values, $\mathbf{y}$ cannot be obtained from $\mathbf{w}$ without also knowing a sequence of $d$ "starting values". Consider obtaining $\mathbf{y}_{f}$ from $\mathbf{w}_{f}$ and starting values $\mathbf{y}_{\mathbf{s}}=$ $\left(y_{n+1-d}, \ldots, y_{n}\right)^{\prime}$. Theorem 1 in Bell (1984a) can be used to show that

$$
\begin{equation*}
\mathbf{y}_{f}=\mathbf{A y} \mathbf{y}_{.}+\mathbf{C w}{ }_{f} \tag{2.3}
\end{equation*}
$$

for matrices $\mathbf{A}$ and $\mathbf{C}$ determined by $\delta(B)$. The rows of the $m \times m$ matrix $C$ consist of the coefficients of $\xi(B)=$ $1+\xi_{1} B+\xi_{2} B^{2}+\ldots=\delta(B)^{-1}$ in the form

$$
C=\left(\begin{array}{ccccc}
1 & 0 & \cdots & 0 & 0 \\
\xi_{1} & 1 & & 0 & 0 \\
\xi_{2} & \xi_{1} & \ddots & 0 & 0 \\
\vdots & \vdots & \ddots & 1 & 0 \\
\xi_{m-1} & \xi_{m-2} & \cdots & \xi_{1} & 1
\end{array}\right)
$$

A is an $m \times d$ matrix which accounts for the effect of the $d$ starting values in $\mathbf{y}$. on $\mathbf{y}_{f}$. The exact form of $\mathbf{A}$ is given in Bell (1984a) and, since it will exactly cancel in our application, it will not be given here. $\ln (2.3) \mathbf{y}_{\text {, }}$ is known since it is part of $y_{o}$, the observed data. Thus, from (2.3) the MMSE forecast of $\mathbf{y}_{f}$ is $\hat{\mathbf{y}}_{f}=\mathbf{A y},+\mathbf{C} \hat{\mathbf{w}}_{f}$, where $\hat{\mathbf{w}}_{f}$ is the MMSE forecast of $\mathbf{w}_{f}$. Therefore, $\mathbf{f}=\mathbf{y}_{f}-\hat{\mathbf{y}}_{f}=\mathbf{A y} .+\mathbf{C w}_{f}$ $-\left(\mathbf{A y} .+\mathbf{C} \hat{\mathbf{w}}_{f}\right)=\mathbf{C}\left(\mathbf{w}_{f}-\hat{\mathbf{w}}_{f}\right)$, and $\operatorname{var}(\mathbf{f})=\mathbf{C v a r}\left(\mathbf{w}_{f}-\hat{\mathbf{w}}_{f}\right) \mathbf{C}^{\prime}$.

Under Assumption A of Bell (1984a), which leads to the standard results for forecasting nonstationary series (as in, e.g., Box and Jenkins 1976, Chapter 5), $\hat{\mathbf{w}}_{f}=\Sigma_{32} \Sigma_{22}^{-1} \mathbf{w}_{\rho}$. Note that this uses only the differenced data $\mathbf{w}_{o}$ in forecasting $\mathbf{w}_{f}$ Then, from standard results on linear prediction, $\operatorname{var}\left(\mathbf{w}_{f_{1}}-\hat{\mathbf{w}}_{f}\right)=\boldsymbol{\Sigma}_{33}-\boldsymbol{\Sigma}_{32} \boldsymbol{\Sigma}_{22}^{-1} \boldsymbol{\Sigma}_{32}^{\prime}$. Thus, $\operatorname{var}(\mathbf{f})=$ $\mathbf{C}\left(\boldsymbol{\Sigma}_{33}-\boldsymbol{\Sigma}_{32} \boldsymbol{\Sigma}_{22}^{\boldsymbol{L}^{\prime}} \boldsymbol{\Sigma}_{32}^{\prime}\right) \mathbf{C}^{\prime}$.

To obtain var (b) and $\operatorname{cov}(\mathbf{b}, \mathbf{f})$ we note that results obtained by Bell (1984a, p. 651) imply similar calculations hold for the backcast errors $\mathbf{b}$. In fact, it can be shown that $\mathbf{b}=(-1)^{r} \mathbf{C}^{\prime}\left(\mathbf{w}_{b}-\hat{\mathbf{w}}_{b}\right)$, where $\hat{\mathbf{w}}_{b}$ is the MMSE backcast of $\mathrm{w}_{b}$, and $r$ is the number of times $(1-B)$ appears in the polynomial $\delta(B)$. (The appearance of $\mathbf{C}^{\prime}$ in this expression instead of $\mathbf{C}$ stems from the indexing of $\mathbf{w}_{b}$ and $\hat{\mathbf{w}}_{b}$ forward through time although the backcasting process proceeds backwards through time.) Thus, $\operatorname{var}(\mathbf{b})=\mathbf{C}^{\prime} \operatorname{var}\left(\mathbf{w}_{b}-\hat{\mathbf{w}}_{b}\right)$ $\mathbf{C}=\mathbf{C}^{\prime}\left(\boldsymbol{\Sigma}_{11}-\boldsymbol{\Sigma}_{11} \boldsymbol{\Sigma}_{22}^{-1} \boldsymbol{\Sigma}_{12}^{\prime}\right) \mathbf{C}$. Similarly $\operatorname{cov}(\mathbf{f}, \mathbf{b})=(-1)^{r}$ $\mathbf{C}\left(\Sigma_{31}-\Sigma_{32} \Sigma_{22}^{-1} \Sigma_{12}\right) \mathbf{C}$. In practice, to avoid inverting $\Sigma_{22}$,
$\operatorname{var}(\mathbf{f}), \operatorname{var}(\mathbf{b})$, and $\operatorname{cov}(\mathbf{f}, \mathbf{b})$ can be computed using the Cholesky decomposition of $\boldsymbol{\Sigma}_{22}$. (See Appendix A.)
Example-U.S. 5+ Unit Housing Starts (continued). The contribution to seasonal adjustment variance from $\operatorname{var}(\mathbf{b}, \mathbf{0}$, $\mathrm{f})$ is shown in Figure 2. This is zero or essentially zero for observations in the middle of the series, where no or few fore/backcasts need be made to apply the symmetric adjustment filter. Towards the ends of the series, the contribution of fore/backcast error becomes more substantial since an increasing number of observations need to be fore/backcast to apply the filter. The jumps in the graph occur when an additional fore/backcasted observation is multiplied by a weight in the adjustment filter that is a multiple of the seasonal period, since these weights have the greatest magnitude (Bell and Monsell 1992). Note that the contributions from $\operatorname{var}(\mathbf{b}, \mathbf{0}, \mathbf{f})$ at the very ends of the series are smaller than the contributions from var(e), but are not negligible.

### 2.3 Computation of $\operatorname{Cov}[(b \mathbf{f}), \mathrm{e}]$

To compute $\operatorname{cov}(\mathbf{f}, \mathbf{e})$, we first note from results of the preceding section that $\mathbf{f}=\mathbf{y}_{f}-\hat{\mathbf{y}}_{f}=\mathbf{C}\left(\mathbf{w}_{f}-\hat{\mathbf{w}}_{f}\right)=\mathbf{C}$ $\left(\mathbf{w}_{f}-\boldsymbol{\Sigma}_{32} \boldsymbol{\Sigma}_{22}^{-1} \mathbf{w}_{o}\right)=\mathbf{C}\left[0\left|-\boldsymbol{\Sigma}_{32} \boldsymbol{\Sigma}_{22}^{-1}\right| \mathbf{I}_{m}\right] \mathbf{w}=\mathbf{C}\left[0\left|-\boldsymbol{\Sigma}_{32} \boldsymbol{\Sigma}_{22}^{-1} \boldsymbol{\Sigma}\right| \mathbf{I}_{m}\right]$ $\Delta y$. Since $\operatorname{cov}(\mathbf{y}, \mathbf{e})=\operatorname{cov}(\mathbf{Y}+\mathbf{e}, \mathbf{e})=0+\operatorname{var}(\mathbf{e})$, we see that $\operatorname{cov}(\mathbf{f}, \mathbf{e})=\mathbf{C}\left[0\left|-\Sigma_{32} \Sigma_{22}^{-1}\right| \mathbf{I}_{m}\right] \Delta \operatorname{var}(\mathbf{e}) . \quad \operatorname{Cov}(\mathbf{b}, \mathbf{e})$ is computed in an analogous fashion by noting that, $\mathbf{b}=(-1)^{r} \mathbf{C}^{\prime}\left(\mathbf{w}^{-}-\hat{\mathbf{w}}_{b}\right)=(-1)^{r} \mathbf{C}^{\prime}\left(\mathbf{w}_{b}-\boldsymbol{\Sigma}_{12} \boldsymbol{\Sigma}_{22}^{-1} \mathbf{w}_{o}\right)=(-1)^{r} \mathbf{C}^{\prime}$ $\left[\mathbf{I}_{m}\left|-\boldsymbol{\Sigma}_{12} \boldsymbol{\Sigma}_{-2}^{-p}\right| 0\right] \Delta \mathbf{y}$, so that $\operatorname{cov}(\mathbf{b}, \mathbf{e})=(-1)^{r} \mathbf{C}^{\prime}$ $\left[\mathbf{I}_{m}\left|-\boldsymbol{\Sigma}_{12} \boldsymbol{\Sigma}_{22}^{-1}\right| 0\right] \Delta \operatorname{var}(\mathbf{e})$.
Example-U.S. 5+ Unit Housing Starts (continued). Figure 2 shows that the contribution of $\operatorname{cov}[(\mathbf{b}, \mathbf{0}, \mathbf{f}), \mathbf{e}]$ is zero or near zero in the middle of the series, but it becomes increasingly negative towards the ends of the series, in a pattern similar, though opposite in sign and of smaller magnitude, to that of $\operatorname{var}(\mathbf{b}, \mathbf{0}, \mathbf{f})$. At the very ends of the series, however, the pattern reverses and the covariance increases. The elements of $\operatorname{cov}[(\mathbf{b}, \mathbf{0}, \mathbf{f}) \mathbf{e}]$, are mainly positive, so its contribution to the seasonal adjustment variance is negative because $\operatorname{cov}[(\mathbf{b}, \mathbf{0}, \mathbf{f}), \mathbf{e}]$ and its transpose are subtracted from $\operatorname{var}(\mathbf{e})+\operatorname{var}(\mathbf{b}, \mathbf{0}, \mathbf{f})$. The net effect is that subtracting $\boldsymbol{\Omega}\left\{\operatorname{cov}[(\mathbf{b}, \mathbf{0}, \mathbf{f}), \mathbf{e}]+\operatorname{cov}[(\mathbf{b}, \mathbf{0}, \mathbf{f}), \mathbf{e}]^{\prime}\right\} \boldsymbol{\Omega}^{\prime}$ tends to offset the effect of adding $\boldsymbol{\Omega} \operatorname{var}(\mathbf{b}, \mathbf{0}, \mathbf{f}) \boldsymbol{\Omega}^{\prime}$, except near the very ends of the series. Thus, the graph of the variances of the seasonally adjusted series in Figure 2 is very similar to the graph of the contribution of $\operatorname{var}(\mathbf{e})$, except near the very ends of the series. We observed this type of "cancellation effect" in several other examples, including those of section 4.

### 2.4 Comparison with the First Approach of Wolter and Monsour

The first approach of Wolter and Monsour (1981) proposed use of $\boldsymbol{\Omega}_{w m} \operatorname{var}\left(\mathbf{e}_{o}\right) \boldsymbol{\Omega}_{w m}^{\prime}$ as the variance-covariance matrix of the $\mathrm{X}-11$ seasonal adjustment errors, where $\boldsymbol{\Omega}_{w m}$ is an $n+n$ matrix whose rows contain the $\mathrm{X}-11$ linear filter weights, both symmetric and asymmetric. That is, the
middle rows (rows $t$ such that $m<t<n-m+1$, assuming $n>2 m$ ) of $\Omega_{w m}$ contain the X-11 symmetric filter weights, but the first and last $m$ rows of $\boldsymbol{\Omega}_{w m}$ contain X-11's asymmetric filter weights. The middle rows of $\boldsymbol{\Omega}_{w m}$ and $\boldsymbol{\Omega}$ thus contain the same filter weights, but the first and last $m$ rows do not. This means that our approach will give the same results as that of Wolter and Monsour for $m<t<n$ $m+1$, that is, for time points at which the symmetric filter is being used. The results of the two approaches will differ for the first and last $m$ time points. Since the most recent seasonally adjusted data receive the most attention, this difference is potentially important.

Wolter and Monsour also considered use of a matrix $\boldsymbol{\Omega}^{*}$ instead of $\boldsymbol{\Omega}_{w m}$ where $\boldsymbol{\Omega}^{*}$ is $(n+12) \times(n+12)$ to include 12 additional rows of weights corresponding to year-ahead seasonal adjustment filters. Though year-ahead adjustment was the common practice through the early 1980 s, it has now mostly been replaced in the United States by concurrent adjustment (McKenzie 1984).

The differences between our approach and that of Wolter and Monsour can be viewed in two ways. One view is that since Wolter and Monsour did not consider forecast and backcast extension, their approach ignores the contribution of forecast and backcast errors to seasonal adjustment error. This contribution affects results for the first and last $m$ time points, although the examples of section 4 show that this contribution is often small. However, in some cases it is not small, including those time series not subject to sampling error. For such series Wolter and Monsour's approach would assign zero variance to the adjustments, even though initial adjustments would be revised as new data became available.

The other way to view the differences between the approaches centers on the difference in "targets". The seasonal adjustment error under Wolter and Monsour's approach can be thought of as $\boldsymbol{\Omega}_{w m}\left(\mathbf{Y}_{o}-\mathbf{y}_{o}\right)=-\boldsymbol{\Omega}_{w m} \mathbf{e}_{o}$. Since this results in zero error for series with no sampling error ( $\mathbf{Y}_{o}=\mathbf{y}_{o}$ ), Wolter and Monsour implicitly define the seasonal adjustment target to be $\boldsymbol{\Omega}_{w m} \mathbf{Y}_{o^{*}}$. This definition of target has the undesirable property that the target value for a given time point changes as additional data are acquired, since the rows of $\boldsymbol{\Omega}_{w m}$ contain different filter weights. Our target value for any given time point $t$ is always $\omega(B) Y_{t}$.
Example-U.S. 5+ Unit Housing Starts (continued). We compared results using our methodology with that of Wolter and Monsour's using the default X-11 seasonal adjustment filter although, as noted earlier, this example series is adjusted using the optional $3 \times 9$ seasonal moving average filter. This comparison used the default filter for convenience: asymmetric X -1 1 filter weights are needed to obtain results for the Wolter-Monsour approach and we were given a computer program by Nash Monsour that produced them only for the default filter. Figure 3 gives the results for both approaches. The non-constant variances over time from the Wolter-Monsour approach result from applying different filters at different time points. An
interesting consequence of this is that, despite the stationarity of the sampling error, the Wolter-Monsour seasonal adjustment variance is noticeably higher in the middle of the series than for many time points toward (but not close to) either end of the series. This carries the implausible implication that use of less data produces estimators with lower variance. Similar behavior can be observed in several examples presented by Pfeffermann (1994).


Figure 3. U.S. 5+ Units Housing Starts: Comparison with Approach of Wolter and Monsour (1981). The panel descriptions are as for Figure 2. The Wolter and Monsour approach (dotted lines) uses the asymmetric X-1 1 filters for the ends of the series and accounts only for sampling error. Our approaches agree in the middle of the series where there is no contribution from back/forecast error. The Wolter and Monsour variances inappropriately decrease near the ends of the series, suggesting that use of less data produces estimates with lower variances. The results here, in contrast to Figures 1, 2, 4, 5, and 6, use default X-11 filters. (See text.)

The results from using the default X-11 seasonal adjustment filter with our approach are also useful for comparing with the $3 \times 9$ seasonal moving average filter, for which results are given in Figure 2. Differences between results from using the two filters are not great. The contribution of the sampling error is somewhat lower and that of the fore/backcast error somewhat higher when using the default seasonal adjustment filter.

## 3. EXTENSIONS TO THE METHODOLOGY

This section discusses three extensions to the general methodology of section 2. The first two extensions are straightforward, the third more involved.

### 3.1 Variances for Seasonal, Trend, and Irregular Estimates; Variances with Time-Varying Filters

The only way the nonseasonal (seasonally adjusted) component is distinguished in the derivation of section 2 is through the filter weights placed in the matrix $\boldsymbol{\Omega}$. Therefore, corresponding variances for $\mathrm{X}-11$ estimates of the seasonal, trend, and irregular components follow from the same expressions simply by changing the matrix $\Omega$ to contain the desired filter weights. This also changes the dimension of $\Omega$, since the length of the seasonal adjustment, trend, and irregular filters (for given options) differs, and the filter length determines the size of $\boldsymbol{\Omega}$.

A similar extension handles the case of different seasonal moving averages (MAs) selected for different months (or quarters), an option allowed by X-11. This changes the seasonal adjustment (and seasonal, trend, and irregular) filters applied in the different months. The results of section 2 also accommodate this extension through a simple modification of $\boldsymbol{\Omega}$. Since the rows of $\boldsymbol{\Omega}$ correspond to the time points being adjusted, we simply define row $t$ of $\boldsymbol{\Omega}$ to contain the weights (along with sufficient zeroes) from whatever filter is being applied in month $t$. Some care must be taken to dimension $\boldsymbol{\Omega}$ appropriately if the longest selected MA is not used in the first and last months of the series.
Example-U.S. 5+ Unit Housing Starts (continued). Figure 4 shows the variance of the $\mathrm{X}-11$ trend estimate, using the $3 \times 9$ seasonal MA and 13 -term Henderson. The most obvious difference from the seasonal adjustment results is the substantial effect of fore/backcast error at the very ends of the series. This occurs because the largest weights of the trend filter ( $\omega^{(T)}(B)$ ) are the center weight ( $\omega_{o}^{(T)}$ ) and the adjacent weights $\left(\omega_{1}^{(T)}, \omega_{2}^{(T)}, \omega_{3}^{(T)}\right)$ that are applied to data immediately before and after the observation being adjusted (Bell and Monsell 1992). At the very ends of the series, the weights ( $\omega_{1}^{(T)}, \omega_{2}^{(T)}, \omega_{3}^{(T)}$ ) apply to fore/ backcasted observations, which results in large increases in the contribution of fore/backcast error there. The result is that uncertainty about the trend increases sharply at the ends of
the series. In the center of the series, however, the trend variances of Figure 4 are substantially lower than the seasonal adjustment variances of Figure 3, due to the smoothing of the sampling error by the trend filter.


Figure 4. U.S. 5+ Units Housing Starts: Variance Decomposition of the Trend Estimate. The panel descriptions are as for Figure 2. Note the large jump in trend estimate variances at the ends of the series due to the contribution of back/forecast error (third panel).

### 3.2 Variances for Seasonally Adjusted Month-to-Month and Year-to-Year Changes

The variances of the errors of the seasonally adjusted estimates of month-to-month change are the quantities $\operatorname{var}\left(v_{t}-v_{t-1}\right), t=2, \ldots, n$. Given $\operatorname{var}(\mathrm{v})$, the complete error covariance matrix for the seasonally adjusted month-tomonth changes can be calculated as $\Delta_{1} \operatorname{var}(\mathbf{v}) \Delta_{1}^{\prime}$, where

$$
\boldsymbol{\Delta}_{1}=\left(\begin{array}{cccccc}
-1 & 1 & 0 & \cdots & 0 & 0 \\
0 & -1 & 1 & \cdots & 0 & 0 \\
\vdots & & \ddots & \ddots & \vdots & \vdots \\
0 & 0 & & -1 & 1 & 0 \\
0 & 0 & 0 & \cdots & -1 & 1
\end{array}\right)
$$

is of dimension $(n-1) \times n$. The error covariance matrix for the seasonally adjusted year-to-year changes in a quarterly series is calculated similarly as $\Delta_{4} \operatorname{var}(\mathbf{v}) \Delta_{4}^{\prime}$, where

$$
\boldsymbol{\Delta}_{4}=\left(\begin{array}{ccccccccc}
-1 & 0 & 0 & 0 & 1 & 0 & \cdots & 0 & 0 \\
0 & -1 & 0 & 0 & 0 & 1 & \cdots & 0 & 0 \\
\vdots & & \ddots & & & & \ddots & \vdots & \vdots \\
0 & 0 & \cdots & -1 & 0 & 0 & 0 & 1 & 0 \\
0 & 0 & 0 & \cdots & -1 & 0 & 0 & 0 & 1
\end{array}\right)
$$

is of dimension $(n-4) \times n$. The corresponding $(n-12) \times n$ matrix $\Delta_{12}$ for monthly series follows a similar pattern with additional zeroes.

Variances of month-to-month or year-to-year changes in the trend are also easily obtained, as can be seen from this discussion and that of section 3.1.
Example-U.S. 5+ Unit Housing Starts (continued). We produced the standard errors for seasonally adjusted month-to-month and year-to-year changes for this series (Figure 5). Since this time series has been log transformed, standard errors can be approximately interpreted as percentages on the original (unlogged) scale. Compared to the standard errors for the seasonally adjusted series, there are slight increases in the standard errors of the month-to-month changes near the ends of the series, but the standard errors of the year-to-year changes show almost no such increase. Thus, for this series and filter, the uncertainty about month-to-month and year-to-year percent change in the seasonally adjusted data is almost constant across the series. The standard errors of the month-to-month and year-to-year changes are both about 50 percent higher than those for the seasonally adjusted series.

### 3.3 Variances of X-11 Seasonal Adjustments with Estimated Regression Effects

Seasonal adjustment often involves the estimation of certain regression effects to account for such things as calendar variation, known interventions, and outliers (Young 1965; Cleveland and Devlin 1982; Hillmer, Bell, and Tiao 1983; Findley, Monsell, Bell, Otto, and Chen submitted). (Outlier effects are often estimated in the same way as known interventions even though inference about outliers should ideally take account of the fact that the series was searched for the most "significant" outliers.) This section shows how the results already obtained can be extended to include the contribution to seasonal adjustment error of error in estimating regression parameters. We still assume the other model parameters, which determine the covariance structures of $\mathbf{Y}$ and e , are known. In practice these other model parameters will also be estimated, but
accounting for error in estimating them is much more difficult. A Bayesian approach for doing so in the context of model-based seasonal adjustment is investigated by Bell and Otto (submitted).


Figure 5. U.S. $5+$ Units Housing Starts: Standard Errors. These panels contrast the standard errors (not variances, as in previous figures) of the seasonally adjusted data (top panel) with the larger standard errors of seasonally adjusted month-to-month (middle panel) and year-to-year (bottom panel) change estimates.

We extend the model for $Y_{t}$ to include regression terms by writing $Y_{t}=\mathrm{x}_{t}^{\prime} b+Z_{t}$, where $\mathrm{x}_{t}$ is the vector of regression variables at time $t, \beta$ is the vector of regression parameters, and $Z_{i}$ is the series of true population quantities with regression effects removed. Extending our matrixvector notation, we write $\mathbf{Y}=\mathbf{X} \boldsymbol{\beta}+\mathbf{Z}, \mathbf{Y}_{o}=\mathbf{X}_{o} \boldsymbol{\beta}+\mathbf{Z}_{o}$, etc. The regression matrix $\mathbf{X}$ can be partitioned by its rows corresponding to the backcast, observation, and forecast periods: $\mathbf{X}=\left(\mathbf{X}_{b}^{\prime}\left|\mathbf{X}_{o}^{\prime}\right| \mathbf{X}_{f}^{\prime}\right)^{\prime}$. We assume $e_{t}$ has mean zero, so its model does not involve any regression effects. We then have $\mathbf{y}=\mathbf{Y}+\mathbf{e}=(\mathbf{X} \boldsymbol{\beta}+\mathbf{Z})+\mathbf{e}$, with the usual partitioning applying. Letting $z_{t}$ denote the series $y_{\text {, }}$ with the regression effects removed, we have $\mathbf{z}=\mathbf{y}-\mathbf{X} \boldsymbol{\beta}=\mathbf{Z}+\mathbf{e}$.

An additional partition is needed of the matrix $\mathbf{X}$ and vector $\beta$. This is because some of the regression effects in $\mathbf{x}_{t}^{\prime} \beta$ may be assigned to the nonseasonal component while
others, such as trading-day or holiday effects, may be removed as part of the seasonal adjustment. See Bell (1984b) for a discussion. Partition $\mathbf{x}_{t}^{\prime}$ as ( $\mathbf{x}_{S_{t}}^{\prime} \mid \mathbf{x}_{N_{t}}^{\prime}$ ) where $\mathbf{x}_{N_{t}}$ represents the regression variables assigned to the nonseasonal and $\mathbf{x}_{s t}$ the variables whose effects are to be removed in the seasonal adjustment. Correspondingly partition $\beta$ so $\mathbf{x}_{t}^{\prime} \beta=\mathbf{x}_{s t}^{\prime} \beta_{S}+\mathbf{x}_{N_{t}}^{\prime} \beta_{N}$ and $\mathbf{X} \beta=\mathbf{X}_{s} \beta_{s}+\mathbf{X}_{N} \beta_{N}$ $=\left(\mathbf{X}_{s} \mid \mathbf{X}_{N}\right)\left(\beta_{s}^{\prime} \mid \beta_{N}^{\prime}\right)^{\prime} . \quad\left(\mathbf{x}_{s t} \beta_{s}\right.$ is assigned to the "combined" seasonal component.) The matrix $\mathbf{X}$ can thus be partitioned two ways: by seasonal versus nonseasonal regression effects, and by the backcast, observation, and forecast periods. Thus we write

$$
\mathbf{X}=\left[\begin{array}{ll}
\mathbf{x}_{S b} & \mathbf{X}_{N b} \\
\mathbf{X}_{S o} & \mathbf{X}_{N o} \\
\mathbf{X}_{s f} & \mathbf{X}_{N f}
\end{array}\right] .
$$

If $\beta$ were known we could compute $\mathbf{z}_{o}=\mathbf{y}_{o}-\mathbf{X}_{o} \beta=\mathbf{Z}_{o}$ $+\mathbf{e}_{o}$,forecast and backcast extend this series (call the extended series $\hat{\mathbf{z}}$ ), adjust $\hat{\mathbf{z}}$ by X-11 ( $\boldsymbol{\Omega} \hat{\mathbf{z}})$, and add back the required regression effects $\mathbf{X}_{N o} \beta_{N o}$. The target of the seasonal adjustment would be $\mathbf{X}_{N o} \beta_{N o}+\boldsymbol{\Omega} \mathbf{Z}=\mathbf{X}_{N o} \beta_{N o}$ $+\boldsymbol{\Omega}(\mathbf{Y}-\mathbf{X} \boldsymbol{\beta})$, and the seasonal adjustment error would then be $\left(\mathbf{X}_{N o} \beta_{N o}+\boldsymbol{\Omega} \mathbf{Z}\right)-\left(\mathbf{X}_{N o} \beta_{N o}+\boldsymbol{\Omega} \hat{\mathbf{z}}\right)=\boldsymbol{\Omega}(\mathbf{Z}-\hat{\mathbf{z}})$. Thus, if the regression parameters were known they would not contribute to the seasonal adjustment error, and the results already given could be used to compute $\operatorname{var}(\boldsymbol{\Omega}(\mathbf{Z}-\hat{\mathbf{z}}))$.

In practice, $\beta$ will be estimated as part of the model fitting, say by maximum likelihood assuming normality. Given the estimates of the other model parameters, and taking these parameters as if they were known, the maximum likelihood estimate of $\beta$ and its variance are given by

$$
\begin{align*}
\hat{\boldsymbol{\beta}}= & {\left[\mathbf{X}_{o}^{\prime} \boldsymbol{\Delta}_{o}^{\prime} \boldsymbol{\Sigma}_{22}^{-1} \boldsymbol{\Delta}_{o} \mathbf{X}_{o}\right]^{-1} \mathbf{X}_{o}^{\prime} \boldsymbol{\Delta}_{o}^{\prime} \boldsymbol{\Sigma}_{22}^{-1} \boldsymbol{\Delta}_{o} \mathbf{y}_{o} }  \tag{3.1}\\
& \operatorname{var}(\hat{\boldsymbol{\beta}})=\left[\mathbf{X}_{o}^{\prime} \boldsymbol{\Delta}_{o}^{\prime} \boldsymbol{\Sigma}_{22}^{-1} \boldsymbol{\Delta}_{o} \mathbf{X}_{o}\right]^{-1}, \tag{3.2}
\end{align*}
$$

where $\Delta_{o}$ is of dimension $(n-d) \times n$, containing that part of the larger matrix $\Delta$ which differences the observed series $\mathbf{y}_{0}$. The expressions (3.1) and (3.2) are generalized least squares results using the regression equation for the differenced data, $\mathbf{w}_{o}=\Delta \mathbf{y}_{o}=\left(\Delta \mathbf{X}_{o}\right) \beta+\left(\mathbf{u}_{o}+\Delta \mathbf{e}_{o}\right)$, where the error term, $\mathbf{u}_{o}+\Delta \mathbf{e}_{o}$, has covariance matrix $\operatorname{var}\left(\mathbf{w}_{o}\right)=\boldsymbol{\Sigma}_{22}$, which is determined by the other model parameters.

Given the estimated regression parameters $\hat{\beta}$, the seasonally adjusted series would be obtained by subtracting the estimated regression effects from the data (call the resulting series $\hat{\mathbf{z}}_{o}=\mathbf{y}_{o}-\mathbf{X}_{o} \hat{\mathrm{\beta}}$ ), extending this series with forecasts and backcasts using the model (denote this extended series $\hat{\mathbf{z}}=\left[\hat{\mathbf{z}}_{b}, \hat{\mathbf{z}}_{o}, \hat{\mathbf{z}}_{f}\right]$ ), applying X-11 to the extended series ( $\mathbf{\Omega} \hat{\mathbf{z}}$ ), and adding back the estimated regression effects assigned to the nonseasonal component ( $\mathbf{\Omega} \hat{\mathbf{z}}+\mathbf{X}_{N_{o}} \hat{\boldsymbol{\beta}}_{N_{o}}$ ). The target of the seasonal adjustment is still $\mathbf{X}_{N o} \boldsymbol{\beta}_{\mathrm{No}}+\boldsymbol{\Omega} \mathbf{Z}$, discussed above. The seasonal adjustment
error is then $\mathbf{v}=\left(\mathbf{X}_{N_{o}} \boldsymbol{\beta}_{N_{o}}+\boldsymbol{\Omega} \mathbf{Z}\right)-\left(\boldsymbol{\Omega} \hat{\mathbf{z}}+\mathbf{X}_{N_{o}} \hat{\boldsymbol{\beta}}_{N_{o}}\right)=$ $\mathbf{X}_{N_{o}}\left(\beta_{N o}-\hat{\beta}_{N o}\right)+\boldsymbol{\Omega}(\mathbf{Z}-\hat{\mathbf{z}})$.

The expression for $\mathbf{v}$ can be simplified by rewriting $\hat{\mathbf{z}}$. First, let $\mathbf{G}=\left[\mathbf{B}^{\prime}|\mathbf{I}| \mathbf{F}^{\prime}\right]^{\prime}$, where $\mathbf{F}$ is the matrix that produces forecasts $\hat{\mathbf{y}}_{f}$ from $\mathbf{y}_{o}$ and $\mathbf{B}$ is the corresponding matrix that produces backcasts $\hat{\mathbf{y}}_{b}$ from $\mathbf{y}_{o}$. We will not need explicit expressions for $\mathbf{F}$ or $\mathbf{B}$. $\mathbf{G}$ applied to $\mathbf{z}_{0}$ produces $\hat{\mathbf{z}}$ while $\mathbf{G}$ applied to $\hat{\mathbf{z}}_{o}$ produces $\hat{\mathbf{z}}$. Therefore, $\hat{\mathbf{z}}=\hat{\mathbf{z}}-$ $(\hat{\mathbf{z}}-\hat{\mathbf{z}})=\hat{\mathbf{z}}-\left[\mathbf{G}\left(\mathbf{y}_{o}-\mathbf{X}_{o} \boldsymbol{\beta}\right)-\mathbf{G}\left(\mathbf{y}_{o}-\mathbf{X}_{o} \hat{\mathbf{\beta}}\right)\right]=\hat{\mathbf{z}}+\mathbf{G X}{ }_{o}$ ( $\beta-\hat{\beta}$ ). Note that $\mathbf{G} \mathbf{X}_{o}$ is obtained by applying the procedure for forecast and backcast extension (from the model for $z_{i}$ ) to each column of $\mathbf{X}_{o}$. The approach we used to do this is described in Appendix B. Continuing, we have

$$
\begin{aligned}
\mathbf{v} & =\mathbf{X}_{N o}\left(\boldsymbol{\beta}_{N o}-\hat{\boldsymbol{\beta}}_{N o}\right)+\boldsymbol{\Omega}\left[(\mathbf{Z}-\hat{\mathbf{z}})-\mathbf{G} \mathbf{X}_{o}(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}})\right] \\
& =\boldsymbol{\Omega}(\mathbf{Z}-\hat{\mathbf{z}})+\left\{\left[\mathbf{0} \mid \mathbf{X}_{N o}\right]-\boldsymbol{\Omega} \mathbf{G} \mathbf{X}_{o}\right\}(\beta-\hat{\boldsymbol{\beta}}) .
\end{aligned}
$$

Now, $\mathbf{Z}-\hat{\mathbf{z}}=\mathbf{z}-\mathbf{e}-\hat{\mathbf{z}}=[\mathbf{b}|\mathbf{0}| \mathbf{f}]-\mathbf{e}$. Note that $[\mathbf{b}|\mathbf{0}| \mathbf{f}]$, the error vector from projecting $\mathbf{z}$ on $\mathbf{z}_{o}$ or $\mathbf{y}_{o}$, is orthogonal to (uncorrelated with) $\beta-\hat{\beta}$, since $\hat{\beta}$ is a linear function of the data $\mathbf{y}_{o}$. Therefore, letting $\mathbf{K}=\left[0 \mid \mathbf{X}_{N_{o}}\right]-\boldsymbol{\Omega} \mathbf{G} \mathbf{X}_{o}$, we have the variance-covariance matrix of the seasonal adjustment error allowing for error in estimating $\beta$ :

$$
\begin{align*}
& \operatorname{var}(\mathbf{v})=\boldsymbol{\Omega} \operatorname{var}(\mathbf{Z}-\hat{\mathbf{z}}) \boldsymbol{\Omega}^{\prime}+\mathbf{K} \operatorname{var}(\hat{\boldsymbol{\beta}}) \mathbf{K}^{\prime} \\
& +\boldsymbol{\Omega} \operatorname{cov}(\mathbf{e}, \hat{\boldsymbol{\beta}}) \mathbf{K}^{\prime}+\mathbf{K} \operatorname{cov}(\hat{\boldsymbol{\beta}}, \mathbf{e}) \boldsymbol{\Omega}^{\prime} \tag{3.3}
\end{align*}
$$

where $\operatorname{var}(\hat{\boldsymbol{\beta}})$ is given by (3.2). In (3.3) $\boldsymbol{\Omega} \operatorname{var}(\mathbf{Z}-\hat{\mathbf{z}}) \boldsymbol{\Omega}^{\prime}$ is computed by the results of section 2 , and computation of $\mathbf{K} \operatorname{var}(\hat{\beta}) \mathbf{K}^{\prime}$ is straightforward once $\mathbf{G X} \mathbf{X}_{o}$ has been computed. To compute the other two terms requires

$$
\begin{align*}
& \operatorname{cov}(\hat{\boldsymbol{\beta}}, \mathbf{e}) \\
& =\operatorname{cov}\left(\left[\mathbf{X}_{o}^{\prime} \boldsymbol{\Delta}_{o}^{\prime} \boldsymbol{\Sigma}_{22}^{-1} \boldsymbol{\Delta}_{o} \mathbf{X}_{o}\right]^{-1} \mathbf{X}_{o}^{\prime} \boldsymbol{\Delta}_{o}^{\prime} \boldsymbol{\Sigma}_{22}^{-1} \boldsymbol{\Delta}_{o} \mathbf{y}_{o}, \mathbf{e}\right) \\
& =\operatorname{cov}\left(\left[\mathbf{X}_{o}^{\prime} \boldsymbol{\Delta}_{o}^{\prime} \boldsymbol{\Sigma}_{22}^{-1} \boldsymbol{\Delta}_{o} \mathbf{X}_{o}\right]^{-1} \mathbf{X}_{o}^{\prime} \boldsymbol{\Delta}_{o}^{\prime} \boldsymbol{\Sigma}_{22}^{-1}\left[\mathbf{u}_{o}+\boldsymbol{\Delta}_{o} \mathbf{e}_{o}\right], \mathbf{e}\right) \\
& =\left[\mathbf{X}_{o}^{\prime} \boldsymbol{\Delta}_{o}^{\prime} \boldsymbol{\Sigma}_{22}^{-1} \boldsymbol{\Delta}_{o} \mathbf{X}_{o}\right]^{-1} \mathbf{X}_{o}^{\prime} \boldsymbol{\Delta}_{o}^{\prime} \boldsymbol{\Sigma}_{22}^{-1} \boldsymbol{\Delta}_{o}\left[\mathbf{0}_{n \times m} \mid \mathbf{I}_{n \times n}\right. \\
& \left.\mid \mathbf{0}_{n \times m}\right] \operatorname{var}(\mathbf{e}) . \tag{3.4}
\end{align*}
$$

Note that $\left[\mathbf{0}_{n \times m}\left|\mathbf{I}_{n \times n}\right| \mathbf{0}_{n \times m}\right] \operatorname{var}(\mathbf{e})=\left[\operatorname{cov}\left(\mathbf{e}_{o}, \mathbf{e}_{b}\right)\left|\operatorname{var}\left(\mathbf{e}_{o}\right)\right|\right.$ $\left.\operatorname{cov}\left(\mathbf{e}_{o}, \mathbf{e}_{f}\right)\right]$ is the middle $n$ rows of $\operatorname{var}(\mathbf{e})$. Using (3.4) and the aforementioned results, (3.3) can be computed. We can compare the resulting diagonal elements of $\operatorname{var}(\mathbf{v})$ with those of the sum of the last three terms in (3.3), to see if allowing for the error due to estimating the regression parameters is important.

There is an important qualification to make about the results of this section. Since the first term on the right hand side of (3.3), $\boldsymbol{\Omega} \operatorname{var}(\mathbf{Z}-\hat{\mathbf{z}}) \boldsymbol{\Omega}^{\prime}$, is the seasonal adjustment variance we would get by ignoring error in estimating the regression parameters, it is tempting to interpret the sum of
the last three terms in (3.3) as the contribution to seasonal adjustment variance of error due to estimating regression parameters. Unfortunately, this sum is not itself a variance (it can in fact be written as $\operatorname{var}(\mathbf{K} \hat{\beta}+\boldsymbol{\Omega e})-\operatorname{var}(\boldsymbol{\Omega e})$ ), and so it can actually be negative. When this happens the seasonal adjustment variances that allow for error due to estimating regression parameters are actually lower than those that ignore this error. We were in fact able to achieve such a result by artificially modifying model parameters in the following example with trading-day variables (though, as in the results shown, the effects were quite small). This situation contrasts with comparable results for model-based approaches which express the seasonal adjustment error as the sum of two orthogonal terms: the error when all parameters are known, plus the contribution to error from estimating regression parameters. The seasonal adjustment variance in this case is thus the sum of the variances of these two terms, and so the "regression contribution" is always nonnegative. This result is analogous to $\boldsymbol{\Omega} \operatorname{var}(\mathbf{Z}-\hat{\mathbf{z}}) \boldsymbol{\Omega}^{\prime}+\mathbf{K} \operatorname{var}(\hat{\beta}) \mathbf{K}^{\prime}$ in (3.3). The problem in (3.3) is that the $\mathbf{X}-11$ estimate $\boldsymbol{\Omega z}$ is not an optimal (MMSE) estimator of the target $\boldsymbol{\Omega} \mathbf{Z}$, hence the error $\boldsymbol{\Omega}(\mathbf{Z}-\hat{\mathbf{z}})$ is correlated with $\hat{\beta}$ through the sampling error e, leading to the two covariance terms in (3.3). This situation results partly from our choice of target ( $\mathbf{X} \beta+\boldsymbol{\Omega} \mathbf{Z}$ ) and partly from the fact that $\mathrm{X}-11$ cannot be assumed to produce an optimal estimator of anything (note comments related to this in the Introduction).
Example-U.S. 5+ Unit Housing Starts (continued). We use the same example to illustrate the contribution to seasonal adjustment error of adding trading-day variables (Bell and Hillmer 1983), although the corresponding regression coefficients were not statistically significant when estimated with this series. Figure 6a shows the results. In this illustration, the lowest line is the "contribution" to the seasonal adjustment variance from estimating the trading-day effects (but see remarks above). When added to the original estimate of variance (dotted line), we obtain the variance of the seasonally and trading-day adjusted series, allowing for error in estimating the trading-day coefficients (top solid line). We see that, for this example, the increase in variance due to including estimated trading-day effects in the model is slight. Figure 6 b gives results for the trend filter. Here the contribution to trend uncertainty due to estimating the trading-day coefficients is certainly negligible.

The contribution to seasonal adjustment variance of adding three additive outlier variables and one level shift variable is illustrated in Figure 6c. These regression variables were identified as potential outlier effects using the Regarima program (produced by the Time Series Staff at the U.S. Census Bureau) with a critical $t$-statistic of 2.5 . Regarima uses an outlier detection methodology similar to those discussed in Bell (1983) and Chang, Tiao, and Chen (1988). The contributions of the additive outliers appear as three spikes while that of the level shift is a single smaller hump in the middle of the series. In comparison to the
trading-day regression variables, the effect of these outlier variables is mainly local but much stronger. In particular, there is additional uncertainty about seasonal adjustments for observations considered additive outliers.


Figure 6. U.S. 5+ Units Housing Starts: Including the "Contribution" from Regression Effects in the Variance Estimates. The top panel shows both the original variances from the first panel of Figure 2 (dotted curve) and the variances allowing for additional uncertainty due to estimating trading-day regression effects (top solid curve). The regression contribution is also shown (bottom solid curve). The second panel shows the corresponding results for the variances of the trend estimates. Note that the regression contribution to the seasonal adjustment variances is small, and to the trend estimate variances it is essentially zero. The third and fourth panels show analogous results when the trading-day regression effects are replaced by three additive outliers and a level shift. Notice that these have important local effects on the seasonal adjustment and trend estimate variances.

Results for the trend filter (Figure 6d) differ in that uncertainty is much greater around the observation where a level shift was detected, which approaches the level of uncertainty at the ends of the series. A level shift is considered part of the trend, so an estimated level shift effect would first be subtracted from the series (in $\mathbf{X} \hat{\boldsymbol{\beta}}$ ), and then added back following application of the $\mathrm{X}-11$ trend filter. (This is analogous to the treatment of regression effects assigned to the nonseasonal or seasonal components in seasonal adjustment as discussed above.) In contrast, since both additive outliers and level shifts are considered part of the nonseasonal component, all four effects were added back as part of the seasonal adjustment when producing results for Figures $6 a$ and $6 b$.

Actually, these sorts of results for outliers should only be regarded as crude approximations, since they treat the time of occurrence and types of outliers as known, leaving only the magnitude of the effects to be estimated. Ideally, one would like to recognize that the series was searched for significant outliers, but this is much more difficult.

## 4. EXAMPLES

We illustrate our approach using several additional economic time series whose sampling errors follow different models. The models used for these example series are taken from previous work as noted.

### 4.1 Retail Sales of Department Stores

Department store sales are estimated in the Census Bureau's monthly retail trade survey. Essentially all sales come from department store chains, all of which are included in the survey, hence, there is virtually no sampling error in the estimates. Thus, the variance of the $\mathrm{X}-11$ seasonal adjustment comes only from fore/backcast error and from error in estimating regression effects. (Note that the Wolter-Monsour seasonal adjustment variance would be zero for this series.) The model used for this series (Bell and Wilcox 1993), for the period August 1972 through March 1989 for the logs of the observations, is $(1-B)\left(1-B^{12}\right)\left[Y_{-}-x_{t}^{\prime} \beta\right]=(1-0.53 B)\left(1-0.52 B^{12}\right) a_{t}$ with $\sigma_{a}^{2}=4.32 \times 10^{-4}$, where $\mathrm{x}_{t}$ includes variables to account for trading-day and Easter holiday effects, and $Y_{t}=y_{t}$ is the $\log$ of the original series divided by length-of-month factors. In adjusting the series at the Census Bureau, the default X-11 adjustment filter and 13 -term Henderson trend filter are used.

Figure 7a shows the standard errors for the seasonally adjusted data over time, with and without the contribution of regression effects. Unlike the 5+ units housing starts series, there are marked increases in the standard errors of seasonally adjusted data at the ends of series, due entirely to fore/backcast error. The contribution to the standard error due to estimating regression effects is also more pronounced for this series. An interesting feature in Figure 7
is the sets of small downward projecting spikes that occur one year apart in triplets. These occur at non-leap year Februaries, for which there is no trading-day effect (the trading-day regression variables are all zero). There is still a small regression contribution to seasonal adjustment error at these time points since the adjustment averages in these contributions from adjacent time points. (Dips at non-leap year Februaries are also visible on close inspection of Figure 6a.) In addition, for some years, the error in estimating the Easter effect produces a noticeable upward projecting spike involving the two months March and April.


Figure 7. U.S. Department Stores, with Trading-Day and Easter Effects. This series has no sampling error. The four panels give standard errors with and without the contribution from estimating regression effects. For the seasonally adjusted data and corresponding month-to-month and year-to-year changes (first three panels), the "contribution" from estimating regression effects is substantial and erratic in the middle of the series (where it is the sole contributor) but, at either end, diminishes for reasons explained in the text. The regression contribution to the trend estimate standard errors is small.

The regression relative contribution to the seasonal adjustment standard errors diminishes towards the ends of the series. This results from two factors: (1) the magnitude of the regression contribution to $\operatorname{var}\left(v_{t}\right)$ decreases somewhat towards the ends of the series, and, more importantly, (2) $\operatorname{var}\left(Z_{t}-\hat{Z}_{t}\right)$ increases dramatically towards the ends of the series, diminishing the relative contribution to $\operatorname{var}\left(v_{t}\right)$ due to regression (and this is further accentuated when square roots are taken).

The pattern of the standard errors of seasonally adjusted month-to-month changes (Figure 7b) is similar to that for the standard error of the seasonally adjusted data (Figure 7a). The regression contribution is slightly larger than it is for the seasonally adjusted data. Standard errors of year-toyear changes (Figure 7c) follow similar patterns but the regression contribution is considerably larger than it is for the month-to-month changes, and it remains important at the ends of the series.

A similar set of calculations was performed using the default X-11 trend filter, and results for the standard errors of the trend estimates, with and without the regression contribution, are depicted in Figure 7d. The patterns over time of these standard errors are similar to the corresponding figures for the $5+$ units housing starts series, but the standard errors are much smaller due to the absence of sampling error. The regression contribution is small.

The standard errors for all plots in Figure 7 are small -none exceed 0.8 percent. For this series, the regression contribution is small and probably ignorable near the very ends of the series, for all but the year-to-year changes. However, in the middle of the series, the sole contributor to standard errors is that due to the regression effects.

### 4.2 Teenage Unemployment

The Bureau of Labor Statistics (BLS) publishes the monthly time series of number of U.S. unemployed teenagers estimated from the Current Population Survey (CPS). Data from January 1972 to December 1983 ( $(n=144)$ were used by Bell and Hillmer (submitted) to develop a model for this series. The sampling error variance $h_{t}{ }^{2}$ changes over time, so is nonstationary. The sampling error model they developed is

$$
\begin{equation*}
e_{t}=h_{t} \tilde{e}_{t} \text { where }(1-0.6 B) \tilde{e}_{t}=(1-0.3 B) b_{t} \tag{4.1}
\end{equation*}
$$

with $\sigma_{b}^{2}=0.87671$ so that $\operatorname{var}\left(\tilde{e}_{t}\right)=1$. CPS sampling error variances can be approximated by generalized variance functions (Wolter 1985, Chapter 5; Hanson 1968). The generalized variance function Bell and Hillmer used for the teenage unemployment series is

$$
\begin{equation*}
h_{t}^{2}=1.971 y_{t}-\left(1.53 \times 10^{-5}\right) y_{t}^{2} \tag{4.2}
\end{equation*}
$$

where $y_{t}$ is the estimate of the number in thousands of unemployed teenagers at time $t$. The estimated model for the signal component $Y_{t}$ is

$$
\begin{equation*}
(1-B)\left(1-B^{12}\right) Y_{t}=(1-0.27 B)\left(1-0.68 B^{12}\right) a_{t} \tag{4.3}
\end{equation*}
$$

with $\sigma_{a}^{2}=4294$. There are no regression effects in the model and the series is not transformed. BLS uses the default $X-11$ seasonal adjustment filter (so $m=84$ ).

In applying the methods of this paper to this example, problems arise from the fact that the (estimated) sampling error variance $h_{t}^{2}$ depends on the estimate $y_{t}$ through the generalized variance function (4.2). In the backcast and forecast periods $y_{t}$ is unknown. To obtain $h_{t}^{2}$ in these periods we forecast and backcast $y_{t}$ using a simple ARIMA ( $\left.\begin{array}{llll}0 & 1 & 1\end{array}\right)\left(\begin{array}{lll}0 & 1 & 1\end{array}\right)_{12}$ model for $y_{t}$ (not for $Y_{t}$, as in (4.3)). The resulting 84 forecasts and backcasts were then used in (4.2) to produce $h_{t}^{2}$ in the forecast and backcast periods. More refined treatments are possible, such as using the component model given by (4.1) and (4.3) to forecast $y_{t}$.
(a) variance of seasonally adjusted series

(b) contribution of samping error (e)

Cc) contribution of fore/backcast error ( b 0 f)

Cd) contribution of covariances of e and (b 0 f)


Figure 8. Teenage Unemployment, with Default X-11 Options. The panel descriptions are as for Figure 2. The seasonal pattern of the sampling error variance contribution (second panel) results from its dependence on the level of the series through a generalized variance function (see text).

The seasonal adjustment variance for this series (Figure 8a) is dominated at most times $t$ by the sampling error contribution (Figure 8b). This is because, while the contribution of $\operatorname{var}(\mathbf{b}, \mathbf{0}, \mathbf{f})$ is substantial for this series (Figure $8 \mathrm{c})$, it tends to be offset by the contribution of $\operatorname{cov}[(\mathbf{b}, \mathbf{0}, \mathbf{f})$, e] $+\operatorname{cov}[(\mathbf{b}, \mathbf{0}, \mathbf{f}), \mathbf{e}]^{\prime}$ (Figure 8d), except at the first and last
few time points. The patterns of variances of seasonally adjusted month-to-month changes and year-to-year changes (not shown) are similar to that of Figure 8a. The variances of the month-to-month changes are slightly larger than those of the adjusted series, those of the year-to-year changes are larger still.

### 4.3 Retail Sales of Drinking Places

Retail sales of drinking places are estimated in the Census Bureau's monthly retail trade survey. In this survey, (noncertainty) sample cases are independently redrawn approximately every 5 years, so the covariance matrix of the sampling errors is block diagonal. Bell and Hillmer (1990) developed the following model for the sampling error of the logged series within a given sample:

$$
\begin{gather*}
\left(1-0.75 B-0.66 B^{3}+0.50 B^{4}\right)\left(1-0.71 B^{12}\right) e_{t} \\
=(1+0.13 B) b_{t}, \tag{4.4}
\end{gather*}
$$

with $\sigma_{b}^{2}=9.301 \times 10^{-5}$. For time points $t$ and $j$ in different samples, $\operatorname{cov}\left(e_{t}, e_{j}\right)=0$. Bell and Hillmer developed a model for the signal component of the logged series using unbenchmarked estimates from September 1977 to December 1986. We shall instead use the following model fit by Bell and Wilcox (1993) using additional data through October 1989:

$$
\begin{gathered}
(1-B)\left(1-B^{12}\right)\left[Y_{\mathrm{t}}-X_{t}^{\prime} b\right]=(1-0.23 B) \\
\left(1-0.88 B^{12}\right) a_{\mathrm{f}},
\end{gathered}
$$

where $X_{t}$ contains trading-day regression variables, and $\sigma_{a}^{2}=4.16 \times 10^{-4}$.

In seasonally adjusting this series, the default X - 11 filters are used. The contribution of error due to estimating regression parameters is small for this series, and so is not included in the results to follow. Since the contribution of sampling error overwhelms the contributions from fore/ backcast error and $\operatorname{cov}[(\mathbf{b}, \mathbf{0}, \mathbf{f}), \mathbf{e}]+\operatorname{cov}[(\mathbf{b}, \mathbf{0}, \mathbf{f}), \mathbf{e}]^{\prime}$, we also do not illustrate these separate variance contributions. Figure 9a gives the standard error of the seasonally adjusted data (shown over 232 observations to better illustrate the pattern, with vertical lines indicating sample redraws) and Figure $9 b$ the standard error of seasonally adjusted month -to-month changes.

Note the strong pattern in Figure 9a, b due to the redrawing of the sample every five years. In particular, this produces a large spike in the standard error of seasonally adjusted month-to-month changes (Figure 9b) when the sample is redrawn. Similar jumps in standard deviations of year-to-year changes occur for the first year of a new sample. We also found similar patterns for other series from the retail trade survey using models from Bell and Wilcox (1993).

The preceding discussion and results ignored certain aspects of how estimation for the retail trade survey is actually carried out. In fact, to avoid large increases in
variances of change estimates around the sample redraw, such as those reflected in Figure 9b, simple modifications are made to estimates in a newly introduced sample to make their level consistent with that from the old sample. The simplest version of the modification is as follows. Let $z_{\text {(old }) \mathrm{t}}\left(=\exp \left(y_{(\text {old })}\right)\right)$ denote estimates from the old sample, and $z_{\text {(new)t }}$ unmodified estimates from the new sample. Assume that the old sample provides estimates for $t \leq \tau$, and that the new sample is to provide estimates for $t>\tau$. To provide overlap data for the modification, the new sample is begun one month early, so that both $z_{\text {(old)t }}$ and $z_{(\text {new }) \text { t }}$ are available. The modified new sample estimates are defined as $z_{\text {(new)t }}^{\prime}=z_{\text {(new) }}\left(z_{\text {(old) })} / z_{\text {(new)r }}\right)$ for $t \geq \tau$. This modification is carried out each time a new sample is introduced. In terms of the corresponding logged estimates $y_{t}$, the modification is $y_{(\text {new }) t}^{\prime}=y_{(\text {new }) t}+\left(y_{(\text {old }) t}-y_{(\text {new }) t}\right)$. Since the modification to $y_{r}$ is linear, it is easy to account for its effects on the seasonal adjustment variance calculations here. The month-to-month change at time $\tau+1$ before the modification (and without seasonal adjustment) is $y_{(\text {new })+1}-y_{\text {(old }) \tau}$. Note that this change has a large variance since $y_{(\text {new }) \tau+1}$ and $y_{(\text {old }) \tau}$ come from different, independent samples. After modification, this change is $y_{(\text {new }) \tau+1}-y_{(\text {new }) \tau}$, which has a much lower variance due to strong positive correlation between $y_{(\text {new }) \tau+1}$ and $y_{(\text {new }) \text { r }}$ (arising from the the sampling error model (4.4)). Unadjusted month-to-month change estimates for time points other than $\tau+1$ are unaltered by the modification.

Figure 9 d shows that modifying new sample estimates eliminates the large increases in the standard deviation of seasonally adjusted month-to-month changes at the transitions to new samples. Similar effects were seen for year-toyear changes over a one year period. The price paid for this improvement is a steadily increasing error in the level estimates (Figure 9c) following introduction of new samples. This occurs because the modification introduces a transient error into the level estimates that persists throughout the new sample. Thus, the modification trades off worse accuracy of level estimates for improvements in change estimates. (Figure 9c shows no increase for the first five years because we assume the estimates there are not modified to agree with those from a previous sample.) Moreover, the strong patterns in Figure 9a occur because the sampling errors from unmodified estimates in adjacent samples are uncorrelated. On the other hand, sampling errors in the modified estimates are fairly strongly correlated between adjacent samples. The effect of this, after applying the seasonal adjustment filter, is a much different pattern (almost no pattern) in the first five years of Figure 9 c , and slight oscillations around the linear increase thereafter.

The standard errors for the X-11 trend estimates and changes (not shown) look like smoothed versions of those shown in Figure 9.

In practice, final estimates from the retail trade survey are even more complicated than what was just described and illustrated. First, more than one month of overlapping


Figure 9. Retail Sales of Drinking Places: Samples Redrawn Every Five Years. The top panel shows the standard error of the seasonally adjusted data and the second panel the standard error of the corresponding month-to-month changes. The strong pattern results from independently drawing a new sample every five years (at the dotted vertical lines). For month-to-month changes, this produces large increases in standard errors at the time of the sample redraw. To eliminate this problem, a new sample is drawn to overlap with the previous sample for one or more months and the new sample's estimates are modified using data from the overlap to make them consistent in level with estimates from the previous sample (see text). This eliminates the increases in standard errors of change estimates when the sample is redrawn (fourth panel), but introduces a transient error into the modified level estimates, whose effects accumulate over time (third panel).
data are collected and may be used to modify level estimates when a new sample is introduced. More importantly, monthly estimates are benchmarked to agree with annual totals obtained from the more accurate annual retail trade survey or five year economic census. Benchmarking should
thus alleviate the problem of level variances increasing over time seen in Figure 9c. However, since benchmarking imposes linear sum constraints on the original (unlogged) estimates, its effects on seasonal adjustment variances are difficult to investigate under the approach developed here, and we have not done so. (We have used a model for unbenchmarked data to avoid this problem.) Durbin and Quenneville (1995) develop a model-based approach to benchmarking that accounts for the nonlinearities that such benchmark constraints impose on logged data.

## 5. CONCLUSIONS

This paper presented an approach to the long-standing problem of obtaining variances for X-11 seasonal adjustments. Our goal was the development and application of an approach to obtain variances accounting for two sources of error. The first error source is sampling error ( $e_{t}$ ), which arises because we do not observe the true series, $Y_{t}$, but instead observe estimates $y_{t}=Y_{t}+e_{t}$ from a repeated survey. The second error source results from the need to extend the observed series with forecasts and backcasts to apply the symmetric X-11 filters. This second error source leads to seasonal adjustment revisions. To account for these two sources of error, we defined the seasonal adjustment variance as the variance of the error in using the $\mathrm{X}-11$ adjustment to estimate a specific target. This target, $\omega(B) Y_{t}$, is what would result from applying the symmetric, linear X-11 filter, $\omega(B)$, to the true series if its values were available far enough into the future and past for the symmetric filter to be used. (The application to additive $\mathrm{X}-11$ with fore/backcast extension is immediate, and log-additive $\mathrm{X}-11$ is taken as an approximation to multiplicative X -11.)

Our approach was also applied to produce variances of $\mathrm{X}-11$ trend estimates, and to produce variances of month-tomonth and year-to-year changes in both the seasonally adjusted data and trend estimates. A further extension was made to allow for error in estimating regression parameters (e.g., to model calendar effects), though this was more involved and had some limitations.

The variances we obtain ignore uncertainty due to time series variation in the seasonal and nonseasonal components. We argued in section 2 that this may be appropriate for typical users of X-11 seasonally adjusted data. If one desires to account for this time series variation, however, we suggest that consideration be given to model-based approaches to seasonal adjustment, since time series models provide a means to explicitly account for variation in all the components. Alternatively, Pfeffermann (1994) developed an approach to X-11 seasonal adjustments that attempts to account for irregular variation and sampling error.

Our approach builds on the first approach suggested by Wolter and Monsour (1981), by accounting for the contribution of forecast and backcast error that was ignored by
them. An alternative view of the difference between our approach and theirs is that we define a consistent seasonal adjustment target, whereas, in using X-11's asymmetric filters, Wolter and Monsour implicitly used targets that change over time. Because of this, our approach avoids the unrealistic feature of seasonal adjustment variances that decrease towards the ends of the series, which can be seen in results of Wolter and Monsour, and also of Pfeffermann.

In the empirical results presented, the contribution of sampling error often dominated the seasonal adjustment variances. This is partly because sampling error was often large relative to fore/backcast error, and partly because the contribution of fore/backcast error tended to be offset by the contribution of the covariance of fore/backcast error with the sampling error. On the other hand, empirical results for trend estimate variances showed large increases at the ends of series due to the effects of fore/backcast error. Since the largest contribution of fore/backcast error occurs at the ends of the series, and variances for the most recent seasonal adjustments and trend estimates are of the most interest, one should not ignore the contribution of fore/backcast error.

The relative contribution to our variances of error in estimating trading-day or holiday regression coefficients tended to be small, unless the series had no sampling error. Error due to estimating additive outlier and level shift effects was substantial around the time point of the outlier. The effects of AOs were large on seasonal adjustment variances; the effects of LSs were large on trend estimate variances.

Nonstationarities in the sampling errors produced interesting patterns in the seasonal adjustment and trend estimate variances. Two types of sampling error nonstationarities were examined. Seasonal patterns in sampling error variances produced corresponding seasonal patterns in seasonal adjustment variances. Independent redrawings of the sample, which yield sampling errors correlated within but not across samples, produced erratic patterns in seasonal adjustment and trend estimate variances over time within a sample. These patterns approximately repeat across different samples if the samples remain in force for approximately equal time spans.

Computations for the examples shown (given the fitted models, which were obtained from the references cited) were done by programming the expressions of Sections 2 and 3 in the S+ programming statistical language. The resulting computer code is available on request.

## ACKNOWLEDGEMENTS

We thank Nash Monsour for providing the FORTRAN program to calculate default X-11 asymmetric filter weights. In addition, he patiently explained how estimates from the monthly retail trade surveys are modified when new samples are drawn. This paper reports the general results of research undertaken by Census Bureau staff. The views expressed are attributed to the authors and do not necessarily reflect those of the Census Bureau.

## APPENDIX A

Several expressions to be calculated in this paper are of the general form

$$
\begin{equation*}
\mathbf{A} \Sigma^{-1} \mathbf{B} \tag{A.1}
\end{equation*}
$$

where $\boldsymbol{\Sigma}$ is a positive definite matrix, and $\mathbf{A}$ and $\mathbf{B}$ are conformable to $\boldsymbol{\Sigma}$. Let $\boldsymbol{\Sigma}=\mathbf{L} \mathbf{L}^{\prime}$ be the Cholesky decomposition of $\boldsymbol{\Sigma}$. Then $\mathbf{A} \boldsymbol{\Sigma}^{-1} \mathbf{B}=\mathbf{A}\left(\mathbf{L}^{-1}\right)^{\prime} \mathbf{L}^{-1} \mathbf{B}$ and (A.1) can be computed as follows:
(1) Solve $\mathbf{L} \mathbf{Q}_{1}=\mathbf{B}$ for $\mathbf{Q}_{1}$
(2) Solve $\mathbf{L} \mathbf{Q}_{2}=\mathbf{A}^{\prime}$ for $\mathbf{Q}_{2}$

Compute $\mathbf{A} \boldsymbol{\Sigma}^{-1} \mathbf{B}=\mathbf{Q}_{2} \mathbf{Q}_{1}$.
(1) and (2) can be solved efficiently since $L$ is lower triangular.

## APPENDIX B

Two steps are required to obtain $\mathbf{G} \mathbf{X}_{0}$, used in section 3.3. The first step produces "forecast" and "backcast" extension of the differenced regression variables. The second step uses these results and the difference equation to produce forecast and backcast extension of the original (undifferenced) regression variables.

Let $\mathbf{R}_{o}=\Delta_{o} \mathbf{X}_{o}$, where $\Delta_{o}$ is that part of the matrix $\Delta$ which differences the observed series $\mathbf{y}_{o}$. Analogous to the computation of $\hat{\mathbf{w}}_{f}$ and $\hat{\mathbf{w}}_{b}$ in section 2.2, forecast extensions of the differenced regression variables are calculated as $\mathbf{R}_{f}=\boldsymbol{\Sigma}_{32} \boldsymbol{\Sigma}_{22}^{-1} \mathbf{R}_{o}$ and backcast extensions as $\mathbf{R}_{b}=$ $\Sigma_{12} \boldsymbol{\Sigma}_{22}^{-1} \mathbf{R}_{o} \quad \mathbf{R}_{f}$ and $\mathbf{R}_{b}$ are of the form (A.1) and can be computed by the technique given above.

For the second step, let $x_{t}$ denote any one of the regression variables in $\mathbf{X}$. Let the required forecast extensions be denoted $\hat{x}_{n+l}$ for $l=1,2, \ldots, m$. Let the differencing operator in the model be $\delta(B)=1-\delta_{1} B-\ldots-$ $\delta_{d} B^{d}$, and let $\hat{r}_{n+l}$ be the forecast extension of $\delta(B) x_{t}=r_{t}$ at time $n+l\left(\hat{r}_{n+l}\right.$ is an element of $\left.\mathbf{R}_{f}\right)$. The $\hat{x}_{n+l}$ are calculated iteratively as

$$
\hat{x}_{n+l}=\delta_{1} \hat{x}_{n+l-1}+\ldots+\delta_{d} \hat{x}_{n+l-d}+\hat{r}_{n+l}, \text { for } l=1, \ldots, m
$$

where $\hat{x}_{n+j}=x_{n+j}$ if $j \leq 0$.
The required backcast extensions of $x_{t}$ are denoted $\hat{x}_{1-1}$ for $l=1, \ldots, m$. These are also obtained recursively from the difference equation $\delta(B) \hat{x}_{t}=\hat{r}_{t}$ by solving for $\hat{x}_{1-1}$ in the expression

$$
\hat{x}_{d+1-l}=\delta_{1} \hat{x}_{d-l}+\ldots+\delta_{d} \hat{x}_{1-l}+\hat{r}_{d+1+l}
$$

and substituting previously computed backcasts as needed. Thus,

$$
\begin{aligned}
& \hat{x}_{1-l}=\delta_{d}^{-1}\left(\hat{x}_{d+1-l}-\delta_{1} \hat{x}_{d-l}-\ldots-\delta_{d-1} \hat{x}_{2-l}-\hat{r}_{d+1-l}\right), \\
& \quad \text { for } l=1, \ldots, m,
\end{aligned}
$$

where $\hat{x}_{1-j}=x_{1-j}$ for $j \leq 0$.

## REFERENCES

BELL, W. R. (1983). A computer program for detecting outliers in time series. Proceedings of the Business and Economic Statistics Section, American Statistical Association, 634-639.
BELL, W. R. (1984a). Signal extraction for nonstationary time series. Annals of Statistics, 12, 646-664.
BELL, W. R. (1984b). Seasonal Decomposition of Deterministic Effects, Research Report Number 84/01. Statistical Research Division, Bureau of the Census.
BELL, W. R., and HILLMER, S. C. (1983). Modeling time series with calendar variation. Journal of the American Statistical Association, 78, 526-534.
BELL, W. R., and HILLMER, S. C. (1984). Issues involved with the seasonal adjustment of economic time series (with discussion). Journal of Business and Economic Statistics, 2, 291-320.

BELL, W. R., and HILLMER, S. C. (1988). A Matrix Approach to Signal Extraction and Likelihood Evaluation for ARIMA Component Time Series Models. Research Report Number 88/22, Statistical Research Division, Bureau of the Census.
BELL, W. R., and HILLMER, S. C. (1990). The time series approach to estimation for repeated surveys. Survey Methodology, 16, 195-215.

BELL, W. R., and HILLMER, S. C. (submitted). Applying time series models in survey estimation.
BELL, W. R., and MONSELL, B. C. (1992). X-11 Symmetric Linear Filters and Their Transfer Functions. Research Report 92/15, Statistical Research Division, Bureau of the Census.
BELL, W. R., and OTTO, M.C. (submitted). Bayesian Assessment of Uncertainty in Seasonal Adjustment with Sampling Error Present.
BELL, W. R., and WILCOX, D. W. (1993). The effect of sampling error on the time series behavior of consumption data. Journal of Econometrics, 55, 235-265.
BINDER, D. A., and DICK, J. P. (1989). Modelling and estimation for repeated surveys. Survey Methodology, 15, 29-45.
BINDER, D. A., and DICK, J. P. (1990). A method for the analysis of seasonal ARIMA models. Survey Methodology, 16, 239-253.
BOBBITT, L., and OTTO, M. C. (1990). Effects of forecasts on the revisions of seasonally adjusted values using the $\mathrm{X}-11$ seasonal adjustment procedure. Proceedings of the Business and Economics Statistics Section, American Statistical Association, 449-454.

BOX, G.E.P., and JENKINS, G.M. (1976). Time Series Analysis: Forecasting and Control. San Francisco: Holden Day.
BREIDT, F. J. (1992). Variance estimation in the frequency domain for seasonally adjusted time series. Proceedings of the Business and Economic Statistics Section, American Statistical Association, 337-342.
BURRIDGE, P., and WALLIS, K.F. (1984). Unobserved components models for seasonal adjustment filters. Journal of Business and Economic Statistics, 2, 350-359.
BURRIDGE, P., and WALLIS, K. F. (1985). Calculating the variance of seasonally adjusted series. Journal of the American Statistical Association, 80, 541-552.

CHANG, I., TIAO, G. C., and CHEN, C. (1988). Estimation of time series parameters in the presence of outliers. Technometrics, 30 , 193-204.
CLEVELAND, W. S., and DEVLIN, S. J. (1982). Calendar effects in monthly time series: modeling and adjustment. Journal of the American Statistical Association, 77, 520-528.
DAGUM, E. B. (1975). Seasonal factor forecasts from ARIMA models. Proceedings of the 40 th Session of the International Statistical Institute, Warsaw, Poland, 206-219.
DURBIN, J., and QUENNEVILLE, B. (1995). Benchmarking monthly time series with structural time series models. Proceedings of the Survey Methods Section, 23rd Annual Meeting of the Statistical Society of Canada, Montreal, Canada, 13-18.
FINDLEY, D. F., MONSELL, B. C., BELL, W. R., OTTO, M. C., and CHEN, B. (1998). New capabilities and methods of the X-12-ARIMA seasonal adjustment program. Journal of Business and Economic Statistics, 16, 127-152.
GEWEKE, J. (1978). Revision of Seasonally Adjusted Time Series. SSRI Report No. 7822, Department of Economics, University of Wisconsin.
HANSON, R. H. (1968). The Current Population Survey: Design and Methodology. Technical Paper 40, U.S. Census Bureau, Washington, D.C., Government Printing Office.
HAUSMAN, J. A., and WATSON, M. W. (1985). Errors in variables and seasonal adjustment procedures. Journal of the American Statistical Association, 80, 531-540.

HILLMER, S.C., BELL, W.R., and TIAO, G. C. (1983). Modeling considerations in the seasonal adjustment of economic time series. Applied Time Series Analysis of Economic Data, (Ed. A. Zellner), U.S. Department of Commerce, Bureau of the Census, 74-100.

McKENZIE, S.K. (1984). Concurrent seasonal adjustment with Census X-11. Journal of Business and Economic Statistics, 2, 235-249.

McLEOD, I. (1975). Derivation of the theoretical autocovariance function of autoregressive-moving average time series. Applied Statistics, 24, 255-256.

McLEOD, I. (1977). Correction to derivation of the theoretical autocovariance function of autoregressive-moving average time series. Applied Statistics, 26, 194.

PFEFFERMANN, D. (1994). A general method for estimating the variances of X-11 seasonally adjusted estimators. Journal of Time Series Analysis, 15, 85-116.

PFEFFERMANN, D., MORRY, M., and WONG, P. (1993). Estimation of the Variances of X-11-ARIMA Seasonally Adjusted Estimators for a Multiplicative Decomposition and Heteroscedastic Variances. Working Paper METH-93-005, Time Series Research and Analysis Division, Statistics Canada, Ottawa, Canada.

PIERCE, D. A. (1980). Data revisions with moving average seasonal adjustment procedures. Journal of Econometrics, 14, 95-114.

PRESIDENT'S COMMITTEE TO APPRAISE EMPLOYMENT AND UNEMPLOYMENT STATISTICS (1962). Measuring Employment and Unemployment, Washington, D.C.: U.S. Government Printing Office.

SHISKIN, J., YOUNG, A. H., and MUSGRAVE, J. C. (1967). The X-11 Variant of the Census Method II Seasonal Adjustment Program. Technical Paper No. 15, U.S. Department of Commerce, Bureau of Economic Analysis.

TILLER, R. B. (1992). Time series modeling of sample survey data from the U.S. current population survey. Journal of Official Statistics, 8, 149-166.

WALLIS, K. F. (1982). Seasonal adjustment and revision of current data: linear filters for the X-11 method. Journal of the Royal Statistical Society, Series A, 145, 74-85.
WILSON, G. T. (1979). Some efficient computational procedures for high order ARIMA models. Journal of Statistical Computation and Simulation, 8, 301-309.

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

WOLTER, K. M., and MONSOUR, N. J. (1981). On the problem of variance estimation for a deseasonalized Series. In Current Topics in Survey Sampling, (Eds. D. Krewski, R. Platek, and J. N. K. Rao). New York: Academic Press, 367-403.

YOUNG, A. (1965). Estimating Trading-day Variation in Monthly Economic Time Series. Technical Paper 12, Bureau of the Census.

YOUNG, A. H. (1968). Linear approximations to the Census and BLS seasonal adjustment methods. Journal of the American Statistical Association, 63, 445-471.

# Item Selection in the Consumer Price Index: Cut-off Versus Probability Sampling 

JAN DE HAAN, EDDY OPPERDOES and CECILE M. SCHUT ${ }^{1}$


#### Abstract

Most statistical offices select the sample of commodities of which prices are collected for their Consumer Price Indexes with non-probability techniques. In the Netherlands, and in many other countries as well, those judgemental sampling methods come close to some kind of cut-off selection, in which a large part of the population (usually the items with the lowest expenditures) is deliberately left unobserved. This method obviously yields biased price index numbers. The question arises whether probability sampling would lead to better results in terms of the mean square error. We have considered simple random sampling, stratified sampling and systematic sampling proportional to expenditure. Monte Carlo simulations using scanner data on coffee, baby's napkins and toilet paper were carried out to assess the performance of the four sampling designs. Surprisingly perhaps, cut-off selection is shown to be a succesful strategy for item sampling in the consumer price index.


KEY WORDS: Laspeyres price index; Monte Carlo simulation; Sampling; Scanner data; Substitution bias.

## 1. INTRODUCTION

Outsiders may think that measuring inflation is an easy job: just visit shops, collect a lot of prices and average them. However, statisticians engaged in the compilation of the Consumer Price Index (CPI), which is the most widely used measure of inflation, face many theoretical and practical problems. In most countries the CPI is essentially a Laspeyres price index. This index weights the partial price indexes of the various commodities by expenditure shares that are fixed at base period levels. Sampling procedures are needed to estimate the population value. Ideally, the mean square error of the estimator would be minimized. Even though the Laspeyres index formula is extremely simple, the estimation procedures applied to the CPI make it a rather complex statistic. Described in a stylized way, the estimation involves three different kinds of samples. A sample of households taking part in an expenditure survey is used to estimate the commodity group expenditure weights. From each commodity group a sample of commodities (items for short) is selected. The prices of these items are collected in a sample of outlets.

In this paper we focus on the sampling of items. Only a few statistical agencies, e.g., the U.S. Bureau of Labor Statistics, use probability sampling to select items to be priced. Most others, for instance Statistics Netherlands, rely on the judgements of experts working at the central office for determining which items should represent the commodity group. In the past this method could be defended by referring to the lack of appropriate sampling frames. Due to the rapidly increasing automation of the retail industry, registers of consumer goods become more and more available, and probability sampling of items comes in sight. Before changing over to a new sampling strategy, however,
it seems worthwile to experiment with altemative strategies in order to assess their impact on the accuracy of the estimated price index numbers. The question to be answered is whether current non-probabilistic selection practices perform worse, in terms of the mean square error, than probability techniques. This is the main topic of the present paper. Simulation studies were carried out for three commodity groups, i.e., coffee, disposable baby's napkins and toilet paper.

Not so long ago, empirical price index number research was hampered by the fact that highly disaggregated expenditure and quantity information at the individual outlet level was lacking or at best available for small samples. Nowadays, some market research firms have managed to set up vast micro data bases on sales of consumer goods, especially in the field of fast moving consumer goods. These are derived from electronic scanning by bar-code reader or the associated bar-code typed in at the cashier's desk. Bradley, Cook, Leaver and Moulton (1997) give an overview of potential uses of scanner data in CPI construction. Processing large scanner data bases is a rather time-consuming task. For CPI compilation as such, this could prevent an extensive use in the near future. But scanner data certainly provide a rich source of information for empirical analysis. In addition to studies into sampling, scanner data also enable us to calculate price index numbers according to different index formulas. The (fixed weight) Laspeyres price index does not take the households' reactions to relative price changes into account. We therefore examined to what extent the Laspeyres population price indexes of the three commodity groups are biased with respect to the Fisher price index, an index that does account for commodity substitution.

[^3]Section 2 gives an overview of the scanner data that we used. Section 3 addresses four different commodity sampling designs. Three of these (i.e., simple random sampling, stratified sampling, and sampling proportional to size) are probability techniques, whereas the fourth (cut-off sampling) is a judgemental one that mimics official practices in the Netherlands. Section 4 describes Monte Carlo experiments we performed to determine the accuracy of the estimated commodity group price indexes under the various sampling designs mentioned. Section 5 deals with the use of Fisher indexes at the item level and the item group level, respectively. The within-group substitution bias of the Laspeyres commodity group price indexes is shown. Section 6 summarizes and discusses the findings.

## 2. BAR-CODE SCANNING DATA

### 2.1 An Overview

Scannable products are defined in Europe by the European Article Number (EAN). Manufacturers should assign one and only one EAN to every variety, size, type of packaging, etc. of a product. This has two implications. In the first place, EANs sometimes change very rapidly, for instance because of a new packaging. Clearly, this makes it difficult to follow a specific item over time. Secondly, some EANs have negligible expenditures. It seems that the system of classification is too detailed; what is really one item has been classified as a multitude of items. In a test study using scanner data on coffee, Reinsdorf (1995) also found that "items that are, for all practical purposes, the same may occasionally have different UPC's" (the US Universal Product Code). Some aggregation over EANs is required. Fortunately, several product characteristics such as brandname and subname are included in the scanner data sets. We will treat EANs having the same product characteristics as identical items. If the number of characteristics is insufficient, there will of course be a danger of overaggregation, that is of putting heterogeneous items together.

From A.C. Nielsen (Nederland) B.V. we received scanner data sets containing weekly supermarket sales on coffee, disposable baby's napkins and toilet paper. The initial data sets contained 320, 569 and 294 different EANs, respectively. For each EAN, the number of packages sold and the corresponding value is included, together with several product and outlet characteristics. Prices are not included; average prices (unit values) must be calculated from the values and quantities. The coffee data relate to sales over a period of two and a half years, beginning with week 1 of 1994 and ending in week 24 of 1996, in a sample of 20 supermarkets located in a Dutch urban area unknown to us. The data on the other two item groups refer to a sample of 149 shops spread over the whole country, and cover a period of two years, beginning with week 1 of 1995 and ending with week 52 of 1996.

For reasons of convenience we deleted the minor brands. In the case of coffee, only the 15 brands with the highest turnover during the entire observation period were selected from the 55 brands actually sold. After aggregating over EANs with identical product characteristics, we further limited the population to those items that were sold in the base year 1994 and every month thereafter in order to have a complete data set for each month. We ended up with a total of 68 items (excluding coffee beans), among which 40 items of ground coffee (including decaffeinated coffee) and 28 items of instant coffee. These account for $94.5 \%$ of total base year coffee expenditure in the initial data set. For napkins and toilet paper (leaving out moist toilet paper), the brands with a turnover share of less than $1 \%$ were removed. Next, only those items were selected that were sold in the base year 1995 and at least eight months thereafter. This resulted in 58 napkins items and 70 toilet paper items, accounting for $90 \%$ and $86 \%$ of total 1995 expenditure in the initial data sets.

### 2.2 Descriptive Statistics

The most striking feature of the item expenditures is the skewness of the distribution. Figure 1 shows the inequality of the base period expenditures in our adjusted data sets by means of so-called Lorenz curves. The vertical axis depicts the cumulative expenditure total, the horizontal axis the cumulative number of items, both expressed as percentages. The items are sorted in increasing order of expenditure. In case of equal expenditures, the Lorenz curve would lie on the diagonal. The more unequal the distribution becomes, the lower its position will be. Coffee item expenditures are distributed extremely unequal, with the three largest items accounting for over half of total base year (1994) coffee expenditure. For baby's napkins and toilet paper the largest six and eight items, respectively, account for nearly half of total base year (1995) expenditure.

Figure 2 shows unit value index numbers, that is the change in the value per package, irrespective of quantity, brandname, type etc., taken over all outlets. This gives a first impression of the change in "prices" during the period under study. For coffee, there was a remarkable decrease in the second half of 1995 following large price rises in 1994 due to bad harvests in Brazil. Coffee prices are largely determined by world market prices for coffee beans. We did not find evidence of significant differences in price changes between outlets. Baby's napkins differ in this respect. A heavy competition was going on between the various producers (which may have caused the decline of the unit values during 1996), while discounts and other kinds of special actions were offered frequently. Hence, the unit value taken over all items and outlets gives an inaccurate picture of the aggregate price change of baby's napkins.


Figure 1. Distribution of base period item expenditures


Figure 2. Unit value index numbers (1994=100 for coffee, 1995=100 for baby's napkins and toilet paper)

## 3. ESTIMATING LASPEYRES-TYPE PRICE INDEXES

We start this section by introducing some notation. Let commodity group $A$ consist of a finite number, say $N$, of commodities (items); $g \in A$ means that item $g$ belongs to group $A$. We assume that $A$ is fixed during time. In real life this is not true: some products disappear from the market, while new products enter. In the short run, however, the constant item group assumption seems reasonable. Note that we adjusted our initial data set accordingly. The reason behind this is that we want to concentrate solely on the sampling aspect. The Laspeyres (fixed weight) price index of commodity group $A$ in period $t$ is

$$
\begin{equation*}
P^{t}=\frac{\sum_{g \in A} e_{g}^{0} P_{g}^{t}}{\sum_{g \in A} e_{g}^{0}}=\sum_{g \in A} w_{g}^{0} P_{g}^{t}, \tag{1}
\end{equation*}
$$

where $P_{g}^{\prime}$ denotes the price index of item $g, e_{g}^{0}$ the expenditure on $g$ during base period 0 and $w_{g}^{0}$ the corresponding expenditure share of $g$ within item group $A$. In the base period a sample $\hat{A}$ with fixed size $n$ is taken from $A$. Because $A$ is supposed to be fixed during time, it seems natural to keep $\hat{A}$ fixed as well.

### 3.1 Simple Random Sampling

Probability sampling refers to situations in which all possible samples have a known probability of selection. Under simple random sampling (without replacement), all possible samples have equal selection probabilities. The Horvitz-Thompson estimator $\hat{P}_{A}^{t}=(N / n) \sum_{g \in \hat{A}} w_{g}^{0} P_{g}^{t}$ is unbiased for $P^{t}$, that is $E\left(\hat{P}_{A}^{t}\right)=P^{t}$ where the expectation $E($.$) denotes the mean over all possible samples under a$ given sampling design, in this particular case simple random sampling. Despite its unbiasedness, $\hat{P}_{A}^{t}$ will not be used in practice because of two undesirable properties.

Firstly, if the price indexes of all sampled items are equal, the estimated item group index differs from that value, unless the population and sample means of expenditures coincide. Price index makers probably dislike this feature. Secondly, and more importantly, $\hat{P}_{A}^{t}$ is bound to exhibit extraordinary large sampling variance. To overcome both difficulties, $P^{t}$ is estimated by taking unbiased estimators of the numerator and denominator:

$$
\begin{equation*}
\hat{P}_{B}^{t}=\frac{(N / n) \sum_{g \in A} e_{g}^{0} P_{g}^{t}}{(N / n) \sum_{g \in \hat{A}} e_{g}^{0}}=\sum_{g \in \hat{A}} \hat{W}_{g}^{0} P_{g}^{t} \tag{2}
\end{equation*}
$$

where $\hat{w}_{g}^{0}$ is the expenditure share of item $g$ in the sample. Using a first-order Taylor linearization (Särndal, Swensson and Wretman 1992, pp. 172-176), the variance of $\hat{P}_{B}^{t}$ can be written as

$$
V\left(\hat{P}_{B}^{t}\right) \approx V\left(\hat{P}_{A}^{t}\right)-\left(P^{t}\right)^{2}\left(2 \mu^{\prime} \rho^{t}-1\right)\left(\sigma^{0}\right)^{2}
$$

where $\sigma^{0}$ denotes the coefficient of variation or relative standard error of the sample mean of base period expenditures, $\mu^{2}$ is the ratio of the relative standard errors of the average base period expenditures expressed in prices of period $t$ and 0 , and $\rho^{t}$ is the correlation coefficient between the average base period expenditures in prices of $t$ and 0 (which is expected to have a positive sign). The choice for $\hat{P}_{B}^{\prime}$ instead of $\hat{P}_{A}^{t}$ can thus be elucidated by the fact that the former exploits the panel character of the sample; with $\rho^{t}>1 /\left(2 \mu^{t}\right)$, a substantial reduction in variance is expected. An alternative expression for the variance of $\hat{P}_{B}^{\prime}$ is:

$$
\begin{equation*}
V\left(\hat{P}_{B}^{t}\right) \approx \frac{1-f}{n} \frac{N^{2}}{N-1} \sum_{g \in A}\left(w_{g}^{0}\right)^{2}\left(P_{g}^{t}-P^{\prime}\right)^{2} \tag{3}
\end{equation*}
$$

which can be estimated using sample data provided that the sampling fraction $f=n / N$ is known. This formula, earlier mentioned by Balk (1989), shows that the variance depends on the within-group dispersion of the item price indexes. Hence, the variance could be lowered either by constructing item groups made up of items having similar price changes or by enlarging the sample. Särndal et al., (1992, p. 176) caution that "the Taylor linearization method has a tendency to lead to underestimated variances in not so large samples". The CPI item samples are generally quite small. For some item groups there may even be only one or two representative items. Thus, besides being unstable (having a large variance itself), the variance will probably also be underestimated when based on (3).

We note that estimator $\hat{P}_{B}^{\prime}$, being a ratio, suffers from small sample bias of approximately $\mathrm{o}(1 / n)$. It can easily be verified that its absolute value $\left|B\left(\hat{P}_{B}^{\prime}\right)\right| \leq \sigma^{0} \sqrt{ } V\left(\hat{P}_{B}^{\prime}\right)$. If $\sigma^{0}$ is small, say less than 0.1 , the bias of $\hat{P}_{b}^{t}$ may safely be regarded as negligible in relation to its standard error. However, with a small item sample and a large variability of base period expenditures, $\sigma^{0}$ could easily exceed 0.1 by far. We add that the all items CPI is unlikely to be biased to
a large extent on this account, since the bias is a (weighted) average of positive and negative biases of the various item group indexes.

### 3.2 Sampling Proportional to Size

Sampling proportional to size has the advantage that the most important items have a big chance of being sampled. We will restrict ourselves to fixed size sampling without replacement, since this seems most likely to be chosen in case of item sampling proportional to size (see for example the Swedish case described by Dalén and Ohlsson 1995). Base period expenditure acts as our measure of size, and the required first-order inclusion probability for item $g$ is $\pi_{g}=n e_{g}^{0} / e^{0}=n w_{g}^{0}$, where $e^{0}=\sum_{g \in A} A_{g}^{0}$. It follows that $\sum_{g \in \hat{A}} p_{g}^{i} / r$ is an unbiased estimator of $P^{i}$.

Sampling proportional to size without replacement, combined with the Horvitz-Thompson or $\pi$ estimator, is sometimes called $\pi \mathrm{ps}$ sampling. Most existing schemes for fixed-size $\pi$ ps sampling are draw-sequential and rather complicated. We will therefore use systematic $\pi$ ps selection instead. This scheme can be described by imagining the expenditures $e_{g}^{0}(g \in A)$ as cumulatively laid out on a horizontal axis, starting at the origin and ending at $e^{0}$. A real number is randomly chosen in the interval $\left(0, e^{0} / n\right]$, and we proceed systematically by taking the items $g$ identified by points at the constant distance $e^{0 / n}$ apart. This method yields exactly the desired sample size. For commodity groups with large variation in base period expenditures, it may not always be possible to select an item sample strictly proportional to expenditure. Obviously, $\pi_{g} \leq 1$ must be satisfied for all $g$. If $n>1$ and some $e_{g}^{0}$ values are extremely large, it may be true for some items that $n e_{g}^{0} / e^{0}>1$, contradicting the requirement $\pi_{g} \leq 1$. The conflict will be dealt with as follows. The $N$ items are ordered according to descending expenditures. First, if $e_{1}^{0}>e^{0} / n$, we set $\pi_{1}=1$. Next, if $e_{2}^{0}>\left(e^{0}-e_{1}^{0}\right) /(n-1)$, we also set $\pi_{2}=1$. The procedure is repeated until the requirement for sampling proportional to base period expenditure is met for all remaining items. Our recursive approach differs somewhat from the method proposed by Sämdal et al., (1992, p. 90). They suggest to set $\pi_{\rho}=1$ for all $g$ with $e_{g}^{0}>e^{0} / n$. In our data sets this would lead to unnecessary large numbers of items with $\pi_{g}=1$. The subgroup $A_{H}$ of items with the highest base period expenditures which is selected with certainty will be called the self-selecting part of the sample. From the remaining low-expenditure subgroup $A_{L}$ a sample $\hat{A}_{L}$ with size $n_{L}$ is drawn strictly proportional to expenditure. The resulting unbiased estimator is an expenditure weighted average of $P^{t}(H)$, the population Laspeyres price index of $A_{H}$, and $\sum_{g \in \hat{A}_{L}} P_{g}^{t} / n_{L}$, the estimated price index of $A_{L}$.

### 3.3 Stratified Sampling

The obvious advantage of simple random sampling as opposed to sampling proportional to expenditure is that,
apart from a register of items serving as a sampling frame, no other data are required. See also Balk (1994). With very unequally distributed item expenditures chances are big that the market leaders fall outside the sample, a situation that seems intuitively unappealing. We will argue that it would indeed be preferable if they were selected. Recall that the variance of the item group price index under simple random sampling depends on the within-group dispersion of the item price indexes. A variance reduction could be achieved if it were possible to stratify the item group into homogeneous subgroups according to their price changes. However, a priori knowledge of item price changes is not available. Another way to lower the variance might be to stratify the item group into two subgroups, one $\left(A_{H}\right)$ with high base period expenditures which is observed entirely and the other one ( $A_{L}$ ) with low expenditures from which a random sample $A_{L}$ is taken. The new item group price index estimator is an expenditure weighted average of $\hat{P}_{B}^{t}(L)$, the Laspeyres index of the low-expenditure subgroup, estimated in accordance with (2), and $P^{t}(H)$. Its sampling variance is $\left(1-\tau_{H}\right)^{2} V\left[\hat{P}_{B}^{t}(L)\right]$, where $\tau_{H}$ is the expenditure share of $A_{H}$ within $A$. This method does not necessarily reduce the variance of the estimated price index, but it is likely to do so under certain conditions. The variance of the new estimator will be smaller than the variance of $\hat{P}_{B}^{f}$ when

$$
\begin{equation*}
1-\tau_{H}<\frac{\operatorname{se}\left(\hat{P}_{B}^{t}\right)}{\operatorname{se}\left[\hat{P}_{B}^{t}(L)\right]} \tag{4}
\end{equation*}
$$

where $\operatorname{se}($.$) denotes standard error. Inequality (4) is$ expected to hold if the item expenditures are distributed extremely unequal, since $1-\tau_{H}$ will then become much smaller than 1 . Stratification may be especially productive as the overall sample size $n$ increases.

The choice of $\tau_{H}$ and thus of the size $N_{H}$ of the "takeall" stratum $A_{H}$ is a bit of a problem. Preferably we would have some optimality criterion in order to minimize the variance. But since a priori knowledge of item price changes is lacking and past trends do not forecast future price changes very accurately, the optimal size of $A_{H}$ can hardly be computed in practice. In the empirical analysis we will try two different relative sample sizes $\lambda_{H}=N_{H} / n$ of $A_{H}$, namely $\lambda_{H}=1 / 3$ and $\lambda_{H}=2 / 3$. These values suffice to give a clear indication of the performance.

### 3.4 Cut-off Sampling

When the sample size is very small it seems rather likely that stratification with $\lambda_{H}=2 / 3$ leads to a larger standard error of the estimated price index than with $\lambda_{H}=1 / 3$. But what happens if $A_{L}$ is not observed at all, so that $\lambda_{H}=1$ and thus $n=N_{H}$ ? We would then be using (a special type of) cut-off sampling. The item group price index is estimated simply by $\hat{P}_{C}^{t}=P^{t}(H)$. All $g \in A_{H}$ now have an inclusion probability of 1 , whereas all $g \in A_{L}$ have zero inclusion probability (Sämdal et al., 1992, pp. 531-533). Since we know exactly which items will be selected there
is no randomness involved and the sampling variance of the $\hat{P}_{C}^{t}$ is zero by definition. The bias equals the actual error, i.e., the difference between the estimated value and the true population index

$$
\begin{equation*}
\hat{P}_{C}^{t}-P^{t}=\left(1-\tau_{H}\right)\left[P^{t}(H)-P^{t}(L)\right] \tag{5}
\end{equation*}
$$

With an extremely unequal distribution of item expenditures, even a small sample size would cause a large value for $\tau_{H}$. In that case cut-off estimation may outperform stratification, in terms of the mean square error. We may either fix the cut-off rate $\tau_{H}$, so that the sample size $n$ is determined by $\tau_{H}$, or fix the sample size, in which case $\tau_{H}$ depends on the choice of $n$. The latter option was chosen by us since fixed size sampling designs are common practice in selecting CPI items, and because this allows a suitable comparison with other fixed size designs.

The use of cut-off procedures can be justified on the grounds that i) the costs prohibit the construction of a reliable sampling frame for the whole population, and ii) the bias is deemed negligible. Assumption ii) cannot be verified in general, of course. The deliberate exclusion of part of the target population from sample selection may nevertheless give satisfactory results when appropriate corrections are made. Statistics Netherlands makes use of cut-off sampling in various other business surveys, for instance in production and foreign trade statistics where very small enterprises are left unobserved. In the Dutch National Accounts, that use production and foreign trade data as important inputs, explicit estimates are being made for small firms. The cut-off method for CPI item selection, on the other hand, does not correct for the excluded items. In addition to cost-considerations, this method is sometimes defended by the belief that, at least in the longer run, the price changes of less important items will not differ much from those of the market leaders within the same product category because of similar production cost structures.

## 4. EMPIRICAL ESTIMATION

### 4.1 Monte Carlo Simulation

With the exception of cut-off selection it is difficult to find reliable measures of the sampling distributions based on a single sample. Under simple random sampling the estimator $\hat{P}_{C}^{\prime}$ has an unknown bias whereas variance estimation based on Taylor linearization techniques gives inaccurate results because CPI item samples are generally very small. Systematic $\pi \mathrm{ps}$ sampling raises the question of how to estimate the variance of the estimator since the second-order inclusion probabilities are unknown. To obtain the exact sampling distribution we would have to consider all samples $\hat{A}$ that are possible under a certain sampling design. For every $\hat{A}$ the probability of drawing $\hat{A}$ and the estimated value of the commodity group price index must be known in order to calculate the exact values of the expected value, the bias and the variance of the estimator.

This is virtually impossible because of the extremely large number of possible samples. To describe the sampling distribution, we will therefore carry out Monte Carlo simulations. A large number of samples, say $K$, is drawn from the (same) population $A$ according to the given design and for each sample the estimate is calculated. If $K$ is large enough, the distribution of the $K$ estimates will closely approximate the exact sampling distribution. Let $\hat{P}_{k}^{2}$ denote the result for the $k$-th sample under a certain sampling design. Then

$$
\overline{\hat{P}}^{t}=\frac{1}{K} \sum_{k=1}^{K} \hat{P}_{k}^{2}
$$

is an unbiased estimate of the expected value $E\left(\hat{P}^{t}\right)$. We will calculate

$$
\overline{\hat{P}}^{t}-P^{t}
$$

which is an unbiased estimate of the bias $B\left(\hat{\boldsymbol{P}}^{t}\right)$;

$$
S_{P}^{2}=\frac{1}{K-1} \sum_{k=1}^{K}\left(\hat{P}_{k}^{t}-\overline{\hat{P}}^{t}\right)^{2}
$$

which is an unbiased estimate of the variance $V\left(\hat{P}^{t}\right)$; and

$$
\sqrt{\left(\overline{\hat{P}}^{t}-P^{t}\right)^{2}+S_{P}^{2}}
$$

which is an approximately unbiased estimate of the root mean square error (rmse) of $\hat{P}^{t}$. Särndal et al. (1992, p. 280), remark that "the imperfection caused by the finite number of repetitions is more keenly felt in the case of a variance measure ... than in the case of measures calculated as means".

### 4.2 Results

Monte Carlo simulations were carried out with three different sample sizes: $n=3, n=6$ and $n=12$. The number of repetitions ( $K$ ) per experiment was set to 500,000 . Table 1 shows the results for coffee in January 1995 (1994=100), tables 2 and 3 those for baby's napkins and toilet paper, respectively, in January 1996 (1995=100). The choice of the formula with which individual price observations are aggregated into a single item price index is discussed in the Appendix. Throughout this section all item price indexes are calculated as unit value indexes over all outlets. Simple random sampling performs particularly bad. For example, with $n=3$ the true (Laspeyres) coffee price increase of $17.2 \%$ is understated by $1.4 \%$-points. Together with a standard error of $5.1 \%$-points, the rmse amounts to $5.3 \%$-points, that is almost one third of the true price increase. Even with $n=12$, so that the sampling fraction is 0.18 (which would be unusually large in practical situations), the rmse still remains considerably high. Notice that, as expected, the small sample bias is halved when the sample size is doubled. Stratification works reasonably well with larger sample sizes but gives disappointing results with $n=3$. In the latter case, stratification increases the rmse compared to simple random sampling for baby's napkins and toilet paper when $N_{H}=2$ (that is, when $\lambda_{H}=2 / 3$ ). Our favourite probabilistic design would be systematic sampling proportional to expenditure because the estimates are unbiased and their standard errors relatively low. The most surprising finding perhaps is the good performance of cut-off selection. Except for $n=3$ and $n=6$ in case of baby's napkins, this method produces the best results.

Table 1
Estimated Laspeyres Price Index Numbers for Coffee (1994=100), January 1995 ( $N=68$ )

| Sampling scheme | $n=3$ |  |  |  | $n=6$ |  |  |  | $n=12$ |  |  |  |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
|  | exp. value | se | bias | rmse | exp. value | se | bias | rmse | exp. value | se | bias | rmse |
| S.R. ${ }^{*}$ ) | 115.7 | 5.1 | -1.4 | 5.3 | 116.4 | 3.4 | -0.7 | 3.5 | 116.7 | 2.3 | -0.4 | 2.3 |
| $\pi \mathrm{ps}$ | 117.2 | 2.2 | 0 | 2.2 | 117.2 | 1.3 | 0 | 1.3 | 117.2 | 0.7 | 0 | 0.7 |
| Stratified |  |  |  |  |  |  |  |  |  |  |  |  |
| $\lambda_{H}=1 / 3$ | 116.4 | 3.9 | -0.7 | 4.0 | 116.6 | 2.3 | -0.5 | 2.3 | 117.0 | 1.2 | -0.1 | 1.2 |
| $\lambda_{H}=2 / 3$ | 115.6 | 4.5 | -1.5 | 4.7 | 116.4 | 2.5 | -0.7 | 2.6 | 117.0 | 1.1 | -0.2 | 1.1 |
| Cut-off | 117.0 | 0 | -0.2 | 0.2 | 117.2 | 0 | 0.0 | 0.0 | 117.5 | 0 | 0.3 | 0.3 |

*) Simple random
Table 2
Estimated Laspeyres Price Index Numbers for Baby's Napkins (1995=100), January 1996 ( $N=58$ )

| Sampling scheme | $n=3$ |  |  |  | $n=6$ |  |  |  | $n=12$ |  |  |  |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
|  | exp. value | se | bias | rmse | exp. value | se | bias | rmse | exp. value | se | bias | rmse |
| S.R. | 99.4 | 5.0 | 2.3 | 5.5 | 98.7 | 3.9 | 1.5 | 4.2 | 97.9 | 2.9 | 0.8 | 3.0 |
| $\pi \mathrm{ps}$ | 97.2 | 2.8 | 0 | 2.8 | 97.2 | 1.6 | 0 | 1.6 | 97.2 | 1.5 | 0 | 1.5 |
| Stratified |  |  |  |  |  |  |  |  |  |  |  |  |
| $\lambda_{H}=1 / 3$ | 98.9 | 5.0 | 1.8 | 5.3 | 98.1 | 3.3 | 1.0 | 3.4 | 97.4 | 1.7 | 0.2 | 1.7 |
| $\lambda_{H}=2 / 3$ | 98.3 | 5.8 | 1.1 | 5.9 | 97.4 | 3.3 | 0.3 | 3.3 | 97.0 | 1.6 | -0.2 | 1.6 |
| Cut-off | 92.0 | 0 | -5.1 | 5.1 | 93.4 | 0 | -3.8 | 3.8 | 95.5 | 0 | -1.6 | 1.6 |

Table 3
Estimated Laspeyres Price Index Numbers for Toilet Paper (1995=100), January 1996 ( $N=70$ )

| Sampling scheme | $n=3$ |  |  |  | $n=6$ |  |  |  | $n=12$ |  |  |  |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
|  | exp. value | se | bias | rmse | exp.value | se | bias | rmse | exp. value | se | bias | rmse |
| S.R. | 103.9 | 4.5 | 0.1 | 4.5 | 103.9 | 3.5 | -0.1 | 3.5 | 103.9 | 2.6 | 0.1 | 2.6 |
| $\pi \mathrm{ps}$ | 103.9 | 3.4 | 0 | 3.4 | 103.9 | 1.8 | 0 | 1.8 | 103.9 | 1.2 | 0 | 1.2 |
| Stratified |  |  |  |  |  |  |  |  |  |  |  |  |
| $\lambda_{H}=1 / 3$ | 103.5 | 4.3 | -0.3 | 4.3 | 103.7 | 3.2 | -0.1 | 3.2 | 104.0 | 2.1 | 0.1 | 2.1 |
| $\lambda_{H}=2 / 3$ | 103.7 | 4.6 | -0.2 | 4.6 | 104.2 | 3.4 | 0.4 | 3.4 | 103.9 | 1.6 | 0.0 | 1.6 |
| Cut-off | 105.0 | 0 | 1.1 | 1.1 | 104.0 | 0 | 0.1 | 0.1 | 104.0 | 0 | 0.1 | 0.1 |

For coffee we also tried another form of stratified sampling. The entire population of items was subdivided into ground coffee and instant coffee, and we took random samples from each stratum. Although the price changes of instant coffee are smoothed and lag behind as compared to ground coffee, Monte Carlo results using stratified sampling were similar to those using unstratified sampling for all four sampling methods. This contradicts earlier findings (see De Haan and Opperdoes 1997a). The reason is that we deleted some instant coffee items for this study to have a complete data set for each month, and ended up with a minor fraction ( $8 \%$ ) of instant coffee in total base year coffee expenditures.

It would be hazardous to draw conclusions about the performance of the various sampling designs based on simulations for one particular month since it is likely that the outcomes depend on the frequency distribution of the item price indexes. Figure 3 shows these distributions for coffee and baby's napkins in two months. Both distributions move to the left, indicating that the unweighted mean has declined. Apart from that, the frequency distribution for coffee remains quite stable. The shape of the curve for
napkins, on the other hand, changes dramatically; the variance of the item indexes has grown.

Monte Carlo experiments were run for each month of the period under study. Figure 4 shows the rmse with $n=3$. The pattern that emerges for coffee and toilet paper is surprisingly robust: cut-off selection always comes out as best. Apparently, if sample sizes are small, the exclusion of the smaller items does not seem to matter much. This is what many statistical offices have been appreciating for a long time, without being able to test it empirically before. The reason why cut-off selection performs better than sampling proportional to expenditure is in case of toilet paper partly caused by the fact that there is no self-selecting part under the latter sampling scheme. With larger samples the results under cut-off selection and sampling proportional to size are very much alike. For baby's napkins the outcomes differ somewhat. Because of the high volatility of the item indexes, the rmse under cut-off selection varies considerably; it seems to meander around the rmse resulting from systematic sampling proportional to expenditure. The high variability of the error can be considered a drawback of cut-off selection.


Figure 3. Frequency distribution of item price index numbers

## Coffee



## Baby's Napkins



## Toilet paper



Figure 4. Rmse of estimated Laspeyres price indexes ( $n=3$ )

## 5. THE USE OF FISHER INDEXES

### 5.1 Unit Value Versus Fisher Item Indexes

In section 4 the item price indexes were calculated as unit value indexes over all outlets. To assess the impact of the choice of the item index formula on the outcomes of the simulation study, Table 4 compares Monte Carlo results with $n=3$ based on unit value item index numbers (as in tables 1-3) and Fisher item price index numbers; see the Appendix for details. For coffee, we notice hardly any differences. For napkins and toilet paper, on the other hand,
the rmse decreases when Fisher index numbers are used instead, especially in case of simple random sampling. This is caused by the fact that unit value indexes tend to show a more erratic pattern. If physically identical types of napkins or toilet paper are deemed heterogeneous across outlets, so that the Fisher formula would be more appropriate, the use of unit value indexes overstates the price variability of particularly small items and exaggerates the poor performance of simple random sampling. Nevertheless, we would still have to conclude that simple random sampling does not work very well.

### 5.2 Within-group Substitution Bias

Many statistical agencies and users are of the opinion that the CPI should be an approximation to the true cost of living index. This theoretical concept is derived from microeconomics and measures the change in the minimum costs for a representative consumer, or household, necessary to retain the same standard of living or utility. Since utility cannot be measured, a feasible index formula should be chosen that closely approximates the concept. Diewert (1976) showed that (what he calls) superlative indexes provide second order approximations to the cost of living index. The most important feature of superlative price indexes is that they take account of consumers' substitution towards goods and services exhibiting relatively small price increases. These index formulas make use of expenditure data relating to both the base period 0 and the current period $t$. In practice it takes some time before expenditure data are known, so that superlative indexes cannot be compiled in real time. For the sake of timeliness most national statistical offices adopt the Laspeyres (fixed weight) formula for constructing their CPIs.

The Fisher index is one of the best-known superlative indexes. When applied to the item group level, the difference between the population price indexes calculated according to the Laspeyres and the Fisher formula can be interpreted as within-group item substitution bias (Figure 5). For coffee it is less than $1 \%$-point per year. For toilet paper and particularly for napkins the biases are very large, about $1.5-3 \%$-points per year. Within-group substitution bias is generally positive and increases over time. Notice, however, that for baby's napkins in a few months of the first half of 1996 the Laspeyres index numbers are lower than the Fisher index numbers. This unexpected effect, and possibly also the large magnitude of the positive bias in other months, may be due to a deficiency of the data set which only contains supermarkets. It is well-known that baby's napkins are bought in the Netherlands also in other kinds of shops such as drugstores that do not make use of bar-code scanning. Substitution between the included and excluded outlets in the data base may damage our population index numbers as accurate approximations of the true values. We are convinced though that it does not seriously affect the assessment of the sampling methods presented in section 4.

Table 4
Estimated Laspeyres Price Index Numbers Using Alternative Item Indexes ( $n=3$ )

| Sampling scheme | Coffee, January 1995 (1994=100) |  |  |  | Napkins, January 1996 (1995=100) |  |  |  | Toilet paper, January 1996 (1995=100) |  |  |  |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
|  | (1) |  | (2) |  | (1) |  | (2) |  | (1) |  | (2) |  |
|  | exp. <br> value | rmse | exp. <br> value | rmse | exp. <br> value | rmse | exp. <br> value | rmse | exp. <br> value | rmse | exp. value | rmse |
| S.R. *) | 115.7 | 5.3 | 115.8 | 5.3 | 99.4 | 5.5 | 100.4 | 4.2 | 103.9 | 4.5 | 104.0 | 3.5 |
| $\pi \mathrm{ps}$ | 117.2 | 2.2 | 117.2 | 2.2 | 97.2 | 2.8 | 98.6 | 2.1 | 103.9 | 3.4 | 104.3 | 3.6 |
| Stratified |  |  |  |  |  |  |  |  |  |  |  |  |
| $\lambda_{H}=1 / 3$ | 116.4 | 4.0 | 116.5 | 4.0 | 98.9 | 5.3 | 100.1 | 4.1 | 103.5 | 4.3 | 103.3 | 3.6 |
| $\lambda_{H}=2 / 3$ | 115.6 | 4.7 | 115.6 | 4.8 | 98.3 | 5.9 | 99.5 | 4.6 | 103.7 | 4.6 | 103.2 | 3.9 |
| Cut-off | 117.0 | 0.2 | 117.0 | 0.2 | 92.0 | 5.1 | 94.8 | 3.8 | 105.5 | 1.1 | 104.7 | 1.0 |

(1) Based on unit value item index numbers
(2) Based on Fisher item index numbers


Figure 5. Difference between Laspeyres and Fisher population price index numbers

## 6. DISCUSSION

Although bar-code scanning data have some deficiencies, they provide an excellent opportunity to undertake empirical research into various sampling issues concerning CPIs. Our simulations show that, for coffee, disposable baby's napkins and toilet paper at least, simple random sampling of items should be advised against. We believe that this recommendation can be extended to all item groups where the distribution of expenditures is very skewed. If statistical offices want to apply probability sampling, they would do a better job using sampling proportional to expenditure. However, cut-off selection might be a good or even better alternative for those item groups where the various item price changes are not too volatile. As a matter of fact, as far as we are aware this is the first study to supply empirical evidence in support of cut-off CPI item selection methods. Aggregated scanner data - that is, scanner data aggregated over outlets - should give a clear indication of the required cut-off rate. Statistics Netherlands already made use of aggregated Nielsen data on a range of commodity groups in the past in order to select items for the CPI sample.

Cut-off methods are applied extensively in the Netherlands and many other European countries (Boon 1997). In the Netherlands the actual item selection is a little more complex than the situation described above. First, a number of item subgroups instead of specific items are chosen using the cut-off method. Next, a number of specific items are selected from each subgroup through so-called judgemental sampling. The selection of these representative items is based on the judgement of experts working at the central office who should have a firm knowledge of the consumer market in question. Usually the most frequently bought items or those with the highest turnover will be selected, so that the entire sampling scheme is a two-stage cut-off procedure. It is unlikely that such a two-stage method would yield results much different from the single-stage procedure we have used in this paper.

In some other European countries, e.g., the United Kingdom, cut-off selection does not take place at the central office but by field staff at the outlets where prices are measured. To illustrate this method, we choose one item per outlet, namely the item with the highest base period sales in the outlet. For coffee, baby's napkins and toilet paper this yields 2, 12 and 24 different items, respectively. The Laspeyres item group index is estimated in accordance with expression (2), where the item price indexes are calculated as outlet-specific unit value indexes and weighted by outlet-specific weights. Figure 6 shows the rmse resulting from this method. If we compare this with Figure 4 (cut-off selection done at the central office for $n=3$ ), the accuracy of both cut-off selection methods seems "on average" to be of the same order of magnitude, although the pattern is slightly more erratic under selection at the outlets. But such a comparison is quite arbitrary. Why not compare cut-off
selection at the outlets with cut-off selection at the central office for $n=6$, or $n=12$, or indeed for any other sample size? Another problem is that we treated the item price indexes as if they were known with certainty. In reality they will be based on a sample of outlets, so that our results are conditional on this sample. For a proper assessment of both cut-off selection procedures we need to take both the sampling of items and the sampling of outlets into account. However, that is beyond the scope of this paper.

Scanner data not only offer challenging perspectives for statistical research in the field of CPI sampling issues, they also enable us to compile all sorts of index numbers, including superlative indexes, using real and highly disaggregated data at the individual outlet level. We demonstrated that the Laspeyres item group price indexes used by statistical agencies can be biased by more than $+1 \%$-point on a yearly basis with respect to the (superlative) Fisher price index that accounts for item substitution. A related type of bias, caused by neglecting products that are introduced after the base period (see e.g., Boskin, Dulberger, Gordon, Griliches and Jorgenson 1996), was not addressed by us. Scanner data do provide a good opportunity to investigate this new goods bias.


Figure 6. Rmse resulting from cut-off selection at the outlets

## ACKNOWLEDGEMENTS

This research was partially supported by Eurostat (the Statistical Office of the European Communities) under SUP-COM 1996, Lot 1 : Development of methodologies in consumer price indices and purchasing power parities. The authors are grateful to A.C. Nielsen (Nederland) B.V. for providing scanner data at marginal costs. They also wish to thank Bert M. Balk, Leendert Hoven and two anonymous referees for helpful comments on an earlier draft.

## APPENDIX: THE CHOICE OF THE ITEM INDEX FORMULA

To perform sampling simulations we need item index numbers. What index formula should be chosen? Statistical offices are generally forced to calculate indexes at the
lowest level of aggregation based on price data alone because quantity or expenditure data is lacking. See Szulc (1987), Dalén (1991), Balk (1994), and Diewert (1995) for a comprehensive treatment of this subject. With scanner microdata at hand, we are in the unique position to construct genuine price indexes (Silver 1995, Hawkes 1997). Consider a set of outlets $B_{g}$, assumed fixed during time, where item $g$ can be bought; $b \in B_{g}$ means that $g$ can be bought in outlet $b$. The price of $g$ at outlet $b$ in period $s$ ( $s=0, t$ ) and the corresponding quantity sold are denoted $P_{g b}^{s}$ and $x_{g b}^{s}$, respectively. The item will be taken as the lowest aggregation level where price indexes are constructed. As a start we restrict ourselves to item indexes that can be written as ratios of weighted arithmetic mean prices in period $t$ and period 0 :

$$
\begin{equation*}
P_{g}^{t}=\frac{\sum_{b \in B_{g}} w_{g b}^{s} P_{g b}^{t}}{\sum_{b \in B_{g}} w_{g b}^{u} P_{g b}^{0}}, \tag{6}
\end{equation*}
$$

where $w_{g b}^{z}=x_{g b}^{2} / \sum_{b \in B_{g}} x_{g b}^{z}$ denotes the share of outlet $b$ in the total quantity sold of item $g$ in period $z(z=s, u)$. If $u=0$ and $s=t$, the prices in period 0 and period $t$ are weighted by the corresponding relative quantities. The average prices are called unit values, and $P_{g}^{t}$ is a unit value index. De Haan and Opperdoes (1997b) and Balk (1998) discuss its merits. Adding up quantities makes sense only if item $g$ can be conceived of as being homogeneous, that is identical across all $b \in B_{g}$. Unit values then yield the appropriate average transaction prices and the unit value index is the appropriate item price index.

The problem, of course, is to define homogeneity. It can be argued that physically identical products sold in different outlets are not identical items because of different services that accompany the transactions, so that homogeneity across outlets never occurs. Another index formula should then be chosen. If $u=s$ in expression (6), $P_{g}^{t}$ can be called a fixed quantity price index with $u$ acting as the quantity reference period. For $u=s=0, P_{g}^{t}$ turns into the Laspeyres price index, and for $u=s=t, P_{g}^{\prime}$ is the Paasche price index. On theoretical grounds we cannot favour either one. For reasons of symmetry it seems natural to take the (unweighted) geometric average of the Paasche and the Laspeyres index, which is the Fisher (ideal) price index.

## REFERENCES

BALK, B.M. (1989). On Calculating the Precision of Consumer Price Indices. Report, Department of Price Statistics, Statistics Netherlands, Voorburg.
BALK, B.M. (1994). On the first step in the calculation of a consumer price index. Proceedings of the First International Conference on Price Indices, Statistics Canada, Ottawa.

BALK, B.M. (1998). On the use of unit value indices as consumer price subindices. Proceedings of the Fourth International Conference on Price Indices, U.S. Bureau of Labor Statistics, Washington, D.C.
BOON, M. (1997). Sampling Designs in Constructing Consumer Price Indices: Current Practices at Statistical Offices. Research Paper no. 9717, Research and Development Division, Statistics Netherlands, Voorburg.
BOSKIN, M.J., DULBERGER, E.R., GORDON, R.J., GRILICHES, Z., and JORGENSON, D. (1996). Toward a More Accurate Measure of the Cost of Living. Final report to the U.S. Senate Finance Committee, Washington, D.C.
BRADLEY, R., COOK, B., LEAVER, S.G., and MOULTON, B.R. (1997). An overview of research on potential uses of scanner data in the U.S. CPI. Proceedings of the Third International Conference on Price Indices, Statistics Netherlands, Voorburg.
DALÉN, J. (1991). Computing elementary aggregates in the Swedish consumer price index. Journal of Official Statistics, 8, 129-147.
DALÉN, J., and OHLSSON, E. (1995). Variance estimation in the Swedish consumer price index. Journal of Business \& Economic Statistics, 13, 347-356.
DIEWERT, W.E. (1976). Exact and superlative index numbers. Journal of Econometrics, 4, 115-145.

DIEWERT, W.E. (1995). Axiomatic and Economic Approaches to Elementary Price Indexes. Working Paper no. 5104, National Bureau of Economic Research, Cambridge.

HAAN, J. De, and OPPERDOES, E. (1997a). Estimation of the Coffee Price Index Using Scanner Data: The Sampling of Commodities. Research Paper, Socio-economic Statistics Division, Statistics Netherlands, Voorburg.

HAAN, J. De, and OPPERDOES, E. (1997b). Estimation of the coffee price index using scanner data: The choice of the micro index. Proceedings of the Third International Conference on Price Indices, Statistics Netherlands, Voorburg.

HAWKES, W.J. (1997). Reconciliation of consumer price index trends with corresponding trends in average prices for quasi-homogeneous goods using scanning data. Proceedings of the Third International Conference on Price Indices, Statistics Netherlands, Voorburg.

REINSDORF, M. (1995). Constructing Basic Component Indexes For The U.S. CPI From Scanner Data: A Test Using Data On Coffee. Presented at NBER Conference on Productivity, Cambridge, Mass., July 17, 1995.

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

SILVER, M. (1995). Elementary aggregates, micro-indices and scanner data: some issues in the compilation of consumer price indices. Review of Income and Wealth, 41, 427-438.

SZULC, B.J. (1987). Price indices below the basic aggregation level. ILO Bulletin of Labour Statistics, Reprinted in Turvey (1989).

TURVEY, R. (1989). Consumer Price Indices: An ILO Manual. Geneva: International Labour Office.

# Robust Calibration Estimators 

PIERRE DUCHESNE ${ }^{1}$


#### Abstract

We consider the use of calibration estimators when outliers occur. An extension is obtained for the class of Deville and Sämdal (1992) calibration estimators based on Wright (1983) QR estimators. It is also obtained by minimizing a general metric subject to constraints on the calibration variables and weights. As an application, this class of estimators helps us consider robust calibration estimators by choosing parameters carefully. This makes it possible, e.g., for cosmetic reasons, to limit robust weights to a predetermined interval. The use of robust estimators with a high breakdown point is also considered. In the specific case of the mean square metric, the estimator proposed by the author is a generalization of a Lee (1991) proposition. The new methodology is illustrated by means of a short simulation study.


KEY WORDS: Calibration estimator; Regression estimator; Range restrictions; Robustness.

## 1. INTRODUCTION

The problem of outliers is an important one in all branches of statistics. In sampling theory, the background is different from that of parametric statistics since the objective is often to estimate the total of a variable of interest $y$. An outlier may have its full weight within the population total. Moreover, methodologists may assume, at the estimation stage, that the values of units are recorded without error, since the gathered units are often processed within an editing system (Särndal, Swensson and Wretman 1992, section 1.7). This step is part of the sampling procedure in large statistical agencies such as Statistics Canada. Lee (1995) has provided an overview of robustness developments within sampling theory.

Nevertheless, since populations for economic surveys are often asymmetric, some units might be extreme as compared to others, as was discussed by Kish (1965). The complete elimination of such units would lead to biased estimates, while maintaining them with their full weight might make an estimator such as the generalized regression (GREG) estimator highly variable. This would suggest a compromise between bias and variance. When outliers occur, the challenge is to propose robust estimators of the total that are little affected by certain units that deviate sharply from others. Such estimators should have little bias and a small mean square error. Traditionally, sampling theory has been deeply involved in the development of unbiased or asymptotically design unbiased (ADU) estimators. See for example Särndal et al., (1992, Section 7.12). However, this ADU property is perhaps undesirable within the context of outliers. This was discussed by Chambers and Kokic (1993), who showed the conflict between the ADU property and the robustness of an estimator.

We consider the Horvitz-Thompson (HT) estimator defined by $\hat{T}_{y \mathrm{HT}}=\sum_{s} d_{k} y_{k}$, where $d_{k}=\pi_{k}^{-1}, \pi_{k}$ being the inclusion probability (If $A$ is a set of units, $A \subseteq U$, then $\sum_{A}$
is a notation signifying $\sum_{k \in A}$ ). Let us assume a positive variable of interest $y$ and an asymmetric population. As the HT estimator is a mean weighted by the $d_{k}$, it is vulnerable to large values of $y$. A unit with a high weight $d_{k}$ may also have a considerable impact on the estimation step by including variable estimates. Lee (1995) defined these units as influential. An extreme unit is not necessarily influential if its weight $d_{k}$ is sufficiently small. Traditionally, methodologists have sought to limit the impact of influential units when they are known prior to sampling, by assigning for example sampling weights close to 1 to extreme units. Gambino (1987) and Lee (1995) have nevertheless discussed situations in which this cannot be done. In a major article, Hidiroglou and Srinath (1981) considered changing the sampling weights when outliers occur. Their approach gave much legitimacy to weight modification within sampling procedures.

Many of the first robust alternatives to the total were based on M estimators and GM-estimators. Nevertheless, much interest has been shown recently for estimators that also provide good overall robustness, as measured by the breakdown point of an estimator. These concepts are discussed for example in Donoho and Huber (1983), Hampel, Ronchetti, Rousseeuw and Stahel (1986) and Rousseeuw and Leroy (1987). The breakdown point measures the percentage of outliers within the sample that the estimator can tolerate while providing nonetheless a good estimate of a given characteristic of the population. Lee, Ghangurde, Mach and Yung (1992) required estimators of the total that were based on robust estimators with a high breakdown point.

We will be considering calibration estimators of the total $T_{y}$ written as $\sum_{s} w_{k} y_{k}$. These estimators were developed for example in Deville and Särndal (1992). We are looking for weights $w_{k}$ that are as close as possible to sampling weights $d_{k}=\pi_{k}^{-1}$, while meeting benchmark constraints, denoted CE (also known as calibration constraints),

[^4]\[

$$
\begin{equation*}
\sum_{s} w_{k} x_{k}=T_{x} \tag{1.1}
\end{equation*}
$$

\]

where $x_{k}$ is a vector of dimension $m$ that corresponds to the available auxiliary information of known total $T_{x}=\sum_{U} x_{k}$. These estimators are popular as they are easily interpreted, since methodologists are used to assigning weights $w_{k}$ to units $y_{k}$. Several metrics are studied to measure the proximity between $d_{k}$ and $w_{k}$. The GREG estimator is an important example with $w_{k}=d_{k}\left(1+\left(T_{x}-\hat{T}_{x H T}\right)^{\prime} M_{s}^{-1} x_{k} / c_{k}\right)$, where $M_{s}=\sum_{s} d_{k} x_{k} x_{k}^{\prime} / c_{k}$. It is obtained by minimizing the mean square metric $\sum_{s} c_{k}\left(w_{k}-d_{k}\right)^{2} / d_{k}$. Constants $c_{k}$ are weighting factors which can take into account problems of heteroscedasticity (for example). Särndal (1996) discussed the selection of these constants. However, since the $g$ weights $g_{k}=w_{k} / d_{k}$ of the GREG estimator are not generally restricted, other metrics are proposed as a means of limiting them so that they might meet certain constraints applicable to the range of values (CARV). Specifically, this makes it possible to avoid undesirable negative weights $w_{k}$. See Deville and Särndal (1992), Singh and Mohl (1996) and Stukel, Hidiroglou and Särndal (1996).

As was noted by Fuller, Loughin and Baker (1994, p. 81), there is a link between calibration estimators and robust methods. However, it is wrong to assume that calibration estimators necessarily have good properties of robustness, given that all the calibration estimators considered by Deville and Särndal (1992) were asymptotically equivalent to the GREG estimator, which, being ADU, is not robust. Moreover, a traditional calibration estimator is not robust as it depends linearly on $w_{k}$, and $w_{k}$ and does not take into account $y_{k}$.

The purpose of this paper is to build estimators in the form of $\sum_{s} w_{k} y_{k}$ where the weights $w_{k}$ provide robustness while meeting constraints on the calibration variables and the weights $w_{k}$. The starting point of our approach is the class of Wright (1983) estimators QR. Let us assume we have available constants $\left\{\left(q_{k}, r_{k}\right), q_{k}>0, r_{k} \geq 0, \forall k \in U\right\}$, such that $\sum_{U} \pi_{k} q_{k} x_{k} x_{k}^{\prime}>0$ and $\sum_{s} q_{k} x_{k} x_{k}^{\prime}>0, \forall s$. (If $A$ is a symmetric matrix, $A>0$ means that $A$ is definite as positive.) The QR estimators are defined on the basis of $q_{k}$ and $r_{k}$ by the relation

$$
\begin{equation*}
\hat{T}_{\mathrm{yQR}}=T_{x}^{\prime} \hat{B}_{q}+\sum_{s} r_{k} e_{k}, \tag{1.2}
\end{equation*}
$$

where $\hat{B}_{q}$ assumes a form weighted by the $q_{k}$

$$
\begin{equation*}
\hat{B}_{q}=\left(\sum_{s} q_{k} x_{k} x_{k}^{\prime}\right)^{-1} \sum_{s} q_{k} x_{k} y_{k}, \tag{1.3}
\end{equation*}
$$

and

$$
\begin{equation*}
e_{k}=y_{k}-x_{k}^{\prime} \hat{B}_{q} . \tag{1.4}
\end{equation*}
$$

It will be shown in section 2 that the QR estimators are calibration estimators, and a new class of estimators, denoted $R Q R$, will be introduced, also based on the choice of constants $q_{k}$ and $r_{k}$. It generalizes to a certain extent the QR estimators as well as the class of Deville and Särndal (1992) estimators. The RQR class is interesting in
that it makes it possible to obtain weights $w_{k}$ that are limited to a given interval, say $[L, U]$. Some of the properties of classes QR and RQR are provided in section 2.

Section 3 describes applications of the RQR class in the building of robust calibration estimators. The main goal is to modify robust default weights so that they meet calibration constraints. Section 3.1 discusses the choice of constants $q_{k}$ and $r_{k}$ using arguments suited to calibration estimators. This is a new and unifying approach, and in section 3.2 it guides our choice of $q_{k}$ and $r_{k}$ when there is auxiliary information. One important element is the use of a robust estimator allowing for the weighted form (1.3), providing the $q_{k}$. Note that this is the case for GMestimators. Usually, estimators with a high breakdown point do not have a weighted form. Consideration is given to reweighting these estimators, allowing the breakdown point to be kept under control and making it possible to have estimators written in the form (1.3). See Rousseeuw and Leroy (1987) and Simpson and Chang (1997). We then discuss the choice of $q_{k}$ and $r_{k}$, so as to calculate an RQR estimator and obtain a robust calibration estimator with restricted weights. Various robust estimators, including the Lee (1991) estimator and the Chambers (1986) estimator, are compared in section 4 with RQR estimators as well as with the GREG estimator and one calibration estimator considered in Deville and Särndal (1992) whose weights are limited. The Lee (1991) estimator can be considered a specific case of our approach. It allows us to also consider a new estimator with restricted weights. Four populations that have already been studied in the literature are considered. It will be noted that estimators free of weight constraints are subject to negative weighting problems. With the RQR class of estimators, robust estimators having positive weights can be obtained, and they compare well with estimators free of weight constraints. Finally, conclusions are drawn in section 5. Appendix B contains a list of abbreviations, and Appendix $C$ contains a list of the various constants found in this paper, with definitions.

## 2. RQR CLASS ESTIMATORS

Consider a finite population $U=\{1,2, \ldots, N\}$ of size $N$ whose total $T_{y}=\sum_{U} y_{k}$ we wish to estimate for a variable of interest $y$ that is positive. A sample $s$ of size $n_{s}$ is drawn following a sampling design $p(s)$. The inclusion probability of a unit $k$ is denoted $\pi_{k}$, and the second-order inclusion probabilities are denoted $\pi_{k l}$. We assume that the auxiliary information $x_{k}$ is of unit value, i.e., $x_{k}$ is known from a reliable source $\forall k \in U$.

Wright (1983) introduced a class of QR estimators written in the form (1.2) with the primary objective of unifying a large number of common estimators. We find the best linear unbiased prediction (BLUP) estimator of Royall (1970) derived from the model-based theory,
obtained by assuming $\left(q_{k}, r_{k}\right)=\left(1 / c_{k}, 1\right)$, and the GREG estimator of Cassel, Sämdal and Wretman (1976) by considering the choice $\left(q_{k}, r_{k}\right)=\left(d_{k} / c_{k}, d_{k}\right)$. Alternately, (1.2) can be written as

$$
\hat{T}_{y \mathrm{QR}}=\sum_{s} d_{k} g_{k} y_{k}
$$

where $d_{k} g_{k}$ satisfies

$$
\begin{equation*}
d_{k} g_{k}=r_{k}+\left(T_{x}-\hat{T}_{x r}\right)^{\prime}\left(\sum_{s} q_{k} x_{k} x_{k}^{\prime}\right)^{-1} q_{k} x_{k}, \tag{2.1}
\end{equation*}
$$

with $\hat{T}_{x r}=\sum_{s} r_{k} x_{k}$. Assuming $r_{k}=d_{k}, g_{k}$ corresponds to the $g$-weight of the GREG estimator.

The QR estimators are calibration estimators, obtained by minimizing the mean square metric subject to the CEs

$$
\begin{equation*}
\min \frac{1}{2} \sum_{s}\left(w_{k}-r_{k}\right)^{2} / q_{k}, \text { as of } \sum_{s} w_{k} x_{k}=T_{x} . \tag{2.2}
\end{equation*}
$$

The weights $w_{k}$ are chosen as close as possible to the $r_{k}$ and the $q_{k}$ are weighting factors. In other words, the starting weights $r_{k}$ are transformed into calibration weights $w_{k}$. The solution to problem (2.2) is $w_{k}=d_{k} g_{k}$, where $d_{k} g_{k}$ is given by the formula (2.1).

Nothing, however, guarantees that the weights $w_{k}$ of the QR estimator are positive, which might be undesirable in practice. See Brewer (1994), who formalized the interpretation of weights. To limit the weights $w_{k}$ in $[L, U]$, we wish to resolve

$$
\begin{align*}
& \min \sum_{s} G\left(w_{k} ; q_{k}, r_{k}\right), \\
& \quad \text { as of } \sum_{s} w_{k} x_{k}=T_{x} \text { and } w_{k} \in[L, U] . \tag{2.3}
\end{align*}
$$

The calibration estimator of the total is

$$
\begin{equation*}
\hat{T}_{y \mathrm{RQR}}=\sum_{s} w_{k} y_{k} \tag{2.4}
\end{equation*}
$$

where the $w_{k}$ are obtained by resolving problem (2.3). It is assumed that function $G(w ; q, r)$ is strictly convex and can be derived in $w$ for fixed $r$ and $q$. We denote $g(u ; q, r)=$ $G^{\prime}(u ; q, r)$ and $h(u ; q, r)=g^{-1}(u ; q, r)$. Moreover, it is assumed that $h(0 ; q, r)=r$ and $h^{\prime}(0 ; q, r)=q$. The resulting estimators are called $\mathrm{QR}(\mathrm{RQR})$ restricted calibration estimators.

Fuller et al., (1994) favoured regression estimators having reasonable invariance properties. It can be shown that $R Q R$ estimators are regression equivariant and to scale when constants $q_{k}$ and $r_{k}$ are transformation invariant. Useful definitions may be found in Bolfarine and Zacks (1992).

There is no guarantee that there is a solution to problem (2.3). We refer to the simulation study in Stukel et al., (1996). There may, for example, be realizations of the sample for which even the CEs cannot be satisfied (1.1). Thus, the sample is so imbalanced that it is impossible for the weighted sum of the components for each dimension to provide the corresponding population total. The only
recourse for the practitioner, then, is to relax the constraints by reducing the dimension of the number of auxiliary variables. See also the discussion in Fuller et al., (1994). As for the calibration estimators considered in Deville and Särndal (1992), it was shown, in result 1, that there is a solution with a probability approaching one. Under certain conditions, this result can be adapted to class RQR estimators.

The metric on which we will focus our attention so that the weights may satisfy the CARVs is a slight modification of case No. 7 in Deville and Särndal (1992). We call it the restricted mean square metric. The $G$-function that corresponds to the choice of this metric is

$$
G\left(w_{k} ; q_{k}, r_{k}\right)= \begin{cases}\frac{1}{2}\left(w_{k}-r_{k}\right)^{2} / q_{k} & \text { if } w_{k} \in[L, U] \\ \infty & \text { otherwise }\end{cases}
$$

whereas the $h$-function is

$$
h\left(x_{k}^{\prime} \lambda ; q_{k}, r_{k}\right)= \begin{cases}L & r_{k}+q_{k} x_{k}^{\prime} \lambda<L, \\ r_{k}+q_{k} x_{k}^{\prime} \lambda & r_{k}+q_{k} x_{k}^{\prime} \lambda \in[L, U] \\ U & r_{k}+q_{k} x_{k}^{\prime} \lambda>U\end{cases}
$$

Given this modification, it is the weight $w_{k}$ that is constrained and not only $w_{k} / d_{k}$ as for case No. 7 in Deville and Särndal (1992). In our situation, $w_{k}$ can "correct" an initial weight that is an outlier. It will be noted that, as it is formulated, the Deville and Särndal metric (1992) subtly inserts the constraints on the $w_{k}$ in the $G$-function. In order to calculate the estimator (2.4) according to this metric, it is sufficient to follow the same approach as Deville and Särndal (1992), which leads us to a solution, using Newton's method, for the following equation in $\lambda$

$$
\begin{equation*}
\sum_{s} h\left(x_{k}^{\prime} \lambda ; q_{k}, r_{k}\right) x_{k}=T_{x} \tag{2.5}
\end{equation*}
$$

The final estimator is $\hat{T}_{y \mathrm{RQR}}=\sum_{s} h\left(x_{k}^{\prime} \lambda_{\infty} ; q_{k}, r_{k}\right) y_{k}$, where $\lambda_{\text {© }}$ is the solution to equation (2.5).

It is interesting to know whether the weight constraint changes the properties of the estimator as compared to a QR estimator that is free of weight constraints. The following result (as proven in the Appendix) shows that, under certain conditions, the two estimators are asymptotically equivalent. In practice, using the restricted mean square metric, we have not observed any significant deviations.

Proposition 1. According to hypotheses $C_{1}$ and $C_{2}$ given in the Appendix,

$$
\begin{equation*}
N^{-1}\left|\hat{T}_{y \mathrm{QR}}-\hat{T}_{y \mathrm{RQR}}\right|=o_{p}\left(n^{-1 / 2}\right) \tag{2.6}
\end{equation*}
$$

This result can possibly be obtained using the approach leading to result No. 5 in Deville and Särndal (1992) dealing with the asymptotic equivalence between the GREG and calibration estimators considered by the authors.

However, proposition 1 is of some use to understand the type of conditions needed to reach the result described in our situation.

Since (1.2) shows the same asymptotic behaviour as quantity $T_{x}^{\prime} B_{q}+\sum_{-1} w_{k} E_{k}$, where $E_{k}=y_{k}-x_{k}^{\prime} B_{q}$ and $B_{q}=$ $\left(\sum_{U} \pi_{k} q_{k} x_{k} x_{k}^{q}\right)^{-1} \sum_{U} \pi_{k} q_{k} x_{k} y_{k}$, this would suggest ${ }^{q}$ as variance estimator

$$
\begin{equation*}
v_{L}=\sum \sum_{s} \check{\Delta}_{k l}\left(w_{k} e_{k}\right)\left(w_{l} e_{l}\right), \tag{2.7}
\end{equation*}
$$

where $\Delta_{k l}=\pi_{k l}-\pi_{k} \pi_{l}, \check{\Delta}_{k l}=\Delta_{k l} / \pi_{k l}$ and $e_{k}$ are given by (1.4). See Särndal et al., (1989) and Särndal et al., (1992, p. 234). It can be shown that the asymptotic bias of a QR estimator is given under general conditions by

$$
E_{p}\left(\hat{T}_{y \mathrm{QR}}-T_{y}\right)=\sum_{U}\left(\pi_{k} r_{k}-1\right) E_{k} .
$$

Then, a possible bias estimator is $b=\sum_{s} d_{k}\left(\pi_{k} r_{k}-1\right) e_{k}$, which can be used in conjunction with formula (2.7) to build an estimator of the mean square error for a QR and RQR estimator, using proposition 1.

RQR estimators make it possible to obtain calibration estimators with constrained weights. Given set $q_{k}$ and $r_{k}$, it is sufficient to resolve problem (2.3). In the sections which follow, the RQR class is applied within a context of robustness. We will show how to direct the selection of constants $q_{k}$ and $r_{k}$, chosen in practice using sample $s$.

## 3. BUILDING ROBUST AND CALIBRATED ESTIMATORS

### 3.1 Methods Based on Weight Reduction and Value Modification

Lee (1995) discussed various propositions based on the weight reduction method for simple random sampling. Once outlier observations have been detected, these methods consist in reducing the weight of extreme observations. These methods are to be preferred to those which eliminate doubtful observations entirely, since all the observations in the sample are legitimate, as was discussed by Lee et al., (1992).

With respect to calibration estimators, we begin by considering the situation in which there is no auxiliary information available and the only constraint is $\sum_{s} w_{k}=N$. This case will guide our path. Consider the QR estimator with $q_{k}=r_{k}$. For the sake of our discussion, we consider constants $r_{k}$, known and fixed. The weights minimizing (2.2) subject to $\sum_{s} w_{k}=N$ are $w_{k}=C_{s}(r) r_{k}$, where $C_{s}(r)=N / \sum_{s} r_{k}$, so that $\hat{T}_{y \mathrm{QRR}}$ becomes

$$
\begin{equation*}
\hat{T}_{y \mathrm{QR}}=C_{s}(r) \sum_{s} r_{k} y_{k} . \tag{3.1}
\end{equation*}
$$

Whenever an observation is extreme, it might represent few units like itself within the population, and its weight should
perhaps be reduced. In order to satisfy the CEs, this means finding weights $w_{k}$ that come closest to the sampling weights $d_{k}$ for units which are not outliers but come as close as possible to a reduction factor $r$ for outlier units, where $r$ is chosen by the statistician. Specifically, we denote $s=s_{1} \cup s_{2}$, where $s_{1}$ of cardinality $n_{1}$ represents those units that are not reported as outliers, whereas $s_{2}=s-s_{1}$ of cardinality $n_{2}=n-n_{1}$ represents the outlier units of $s$. The reduction factor $r$ will typically satisfy $r \leq d_{k}, \forall k \in s_{2}$. For example, consider the estimator (3.1) with $q_{k}=r_{k}=$ $B_{k}=d_{k} I_{k 1}+r\left(1-I_{k 1}\right)$, where $I_{k 1}$ is the variable indicating affiliation to $s_{1}$. In this way, constants $q_{k}$ and $r_{k}$ are reduced for units of $s_{2}$ so as to reflect the fact that units of $s_{2}$ are extreme. The estimator (3.1) becomes

$$
\hat{T}_{y \mathrm{QR}}=C_{s}(B)\left(\sum_{s 1} d_{k} y_{k}+r \sum_{s 2} y_{k}\right) .
$$

In the case of simple random sampling, $d_{k}=N / n$ and we obtain

$$
\hat{T}_{y \varrho R}=C_{s}(B) \hat{T}_{y B},
$$

where $\hat{T}_{y B}=N / n \sum_{s 1} y_{k}+r \sum_{s 2} y_{k}$ is the Bershad (1960) estimator discussed in Lee (1995). Other methods based on weight reduction have been discussed in Lee (1995), who also discussed the choice of $r$.

One disadvantage of methods based on weight reduction is that the analyst must identify the outlier units. Methods based on value modification avoid this difficulty by providing gradual weight reduction for units that are more extreme. We consider a case of simple random sampling. We assume

$$
\begin{equation*}
m\left(y_{k} ; t, a, b\right)=b+(a-b) \min \left(1, t / y_{k}\right) . \tag{3.2}
\end{equation*}
$$

Thus, this function assigns a starting weight of value $a$ for the $y_{k}<t$, and gradually reduces this to a final weight $b$, as $y_{k}$ becomes extreme. Value $t$ is called the threshold. The constants $a, b$ and $t$ are chosen by the statistician. Several values for $a$ and $b$ have been considered in the literature. Thus, instead of assigning a fixed reduction factor to the units of $s_{2}$, we select $q_{k}=r_{k}=W_{k}=m\left(y_{k} ; t, N / n, f N / n\right)$, where $f$ is a constant between 0 and 1. The estimator (3.1) becomes

$$
\begin{aligned}
\hat{T}_{y \mathrm{QR}} & =C_{s}(W) \sum_{s} W_{k} y_{k} \\
& =C_{s}(W) \hat{T}_{y W} .
\end{aligned}
$$

The estimator $\hat{T}_{y w}$ has been discussed in Gross, Taylor and Lloyd-Smith (1986) as well as in Chambers and Kokic (1993), who called it the winsorized estimator. This is a special case of the approach used by Chambers (1982, 1986). When $f=0$, the estimator (3.1) becomes $\hat{T}_{y \mathrm{QR}}=C_{s}\left(W_{t}\right) \hat{T}_{y W_{1}}$ with $q_{k}=r_{k}=W_{l k}=m\left(y_{k} ; t, N / n, 0\right)$,
where $\hat{T}_{y w_{1}}=N / n\left(\sum_{s 1} y_{k}+n_{2} t\right)$, denoting the part of $s$ containing units that satisfy $y_{k}<t$. The estimator $\hat{X}_{y W_{1}}$ has been discussed in Lee (1995), as well as in Gross et al., (1986), who called it the type I winsorized estimator. When $f=n / N$, Gross et al., (1986) called it the type II winsorized estimator. It has also been discussed in Bruce (1991).

In a design $\pi \mathrm{ps}$, Dalén (1987) inserted the design by assuming $D_{k}=m\left(y_{k} ; \pi_{k} t, d_{k}, 1\right)$. Thus, if $k$ and $l$ are two extreme observations such that $y_{k}=y_{l}$, then the observation whose sampling weight is largest will have a higher weight $D_{k}$. Selecting $r_{k}=q_{k}=D_{k}$ makes it possible to obtain essentially the Dalén estimator, $\hat{T}_{y Q R}=C_{s}(D) \sum_{s} D_{k} y_{k}$. The estimator $\hat{T}_{y D}=\sum_{s} D_{k} y_{k}$ has been studied for example in Tambay (1988).

Table 3.1
Estimator (3.1) Based on Weight Reduction and Value Modification

| Estimator | Value of $q_{k}=r_{k}$ |
| :--- | :--- |
| Bershad | $B_{k}=d_{k} I_{k l}+r\left(1-I_{k}\right)$ |
| Winsorised | $W_{k}=m\left(y_{k} t, N / n, f N / n\right)$ |
| Winsorised, type I | $W_{l k}=m\left(y_{k} ; t, N / n, 0\right)$ |
| Winsorised, type II | $W_{l k}=m\left(k_{k} ; t, N / n, 1\right)$ |
| Dalén | $D_{k}=m\left(y_{k} ; \pi_{k} t, d_{k}, 1\right)$ |

Note: $m\left(y_{k} ; t, a, b\right)=b+(a-b) \min \left(\mathrm{l}, t / y_{k}\right)$.
The approach used in this section suggests that we may occasionally seek estimators whose weights are close to $r_{k}$ rather than the sampling weights $d_{k}$. The constants $r_{k}$ will themselves be chosen close to $d_{k}$ for the proper units, but will be reduced once a unit is declared extreme. The QR estimators allow the weight reduction and value modification methods to be unified. Methods based on value modification help us choose weights that are adapted to the specific sample $s$ chosen. As was noted in Chambers and Kokic (1993), this is not surprising since the problem of outliers occurs after the selection of sample $s$. We must use the sample at our disposal to overcome the problem. These methods are generalized in the following section using auxiliary information.

### 3.2 Estimators of the Total Based on Robust Statistics

One of the first attempts to obtain robust alternatives to population totals using auxiliary information can be found in Chambers (1982, 1986), who proposed a robust ratio estimator based on BLUP estimator decomposition. One recent extension of the work carried out by Chambers can be found in Welsh and Ronchetti (1998). Gwet and Rivest (1992) also proposed a robust version of the ratio estimator using an approach based on the design in simple random sampling. Rivest and Rouillard (1991) carried out a comparative study of several robust estimators, and examined several estimators of the mean square error. For designs
with unequal probabilities, Hulliger (1995) considered robustifying the HT estimator when inclusion probabilities are obtained using auxiliary information. Gwet and Rivest (1992) and Hulliger (1995) considered a version of the influence function for finite populations, emphasizing the need for procedures having good properties of local robustness and the use of estimators having limited influence functions. Influence functions were discussed generally in Hampel et al., (1986).

The following sections will deal with building robust estimators having constrained weights. The building of such estimators is based on the following steps:

- Identifying the constants $q_{k}$ and $r_{k}$; this provides a QR estimator.
- Resolving the problem (2.3) so as to provide an RQR estimator.
In terms of robustness, the coefficients $q_{k}$ are selected such that $\hat{B}_{q}$ is a robust estimator. Thus, the first part of the QR estimator, $T_{x}^{\prime} \hat{B}_{q}$, provides a good predicted value for the entire population. The second part of the QR estimator, $\sum_{s} r_{k} e_{k}$, corrects the first part for the $y_{k}$ observed in the sample. The constants $r_{k}$ ensure that with this correction, the outliers in the sample will not return with full weight.


### 3.2.1. Choice of $\boldsymbol{q}_{\boldsymbol{k}}$ Based on a GM-Estimator

Consider the estimator (1.2) in which $\hat{B}_{g}$ is replaced by a robust estimator of a regression coefficient. Such estimators have been discussed for example in Huber (1981) and Hampel et al., (1986). We thus obtain

$$
\begin{equation*}
\hat{T}_{g}=T_{x}^{\prime} \hat{B}_{g}+\sum_{s} r_{k}\left(y_{k}-x_{k}^{\prime} \hat{B}_{g}\right) . \tag{3.3}
\end{equation*}
$$

The estimator (3.3) does not have the form of QR estimators unless $\hat{B}_{8}$ assumes a weighted form. This is the case if $\hat{B}_{g}$ is a GM-estimator defined by the equation

$$
\begin{equation*}
\sum_{s} d_{k} h_{k} x_{k} \psi\left(\left(y_{k}-x_{k}^{\prime} B\right) /\left(\sigma h_{k}^{a} \sqrt{c_{k}}\right)\right) / \sqrt{c_{k}}=0 \tag{3.4}
\end{equation*}
$$

since the solution to (3.4) can be expressed as

$$
\hat{B}_{g}=\left(\sum_{s} d_{k} h_{k}^{1-a} u_{k} x_{k} x_{k}^{\prime} / c_{k}\right)^{-1} \sum_{s} d_{k} h_{k}^{1-\alpha} u_{k} x_{k} y_{k} / c_{k},
$$

where

$$
u_{k}=\frac{\psi\left(\left(y_{k}-x_{k}^{\prime} \hat{B}_{g}\right) /\left(\sigma h_{k}^{a} \sqrt{c_{k}}\right)\right)}{\left(y_{k}-x_{k}^{\prime} \hat{B}_{g}\right) /\left(\sigma h_{k}^{a} \sqrt{c_{k}}\right)}
$$

The properties of GM-estimators have been discussed in Simpson and Chang (1997). To simplify our discussion, $\sigma$ is assumed to be known, and the role of $c_{k}$ is the same as in the case of the GREG estimator. The function $\psi$ is determined by the analyst. A current example would be the Huber function

$$
\Psi_{H u b}(x ; c)=\left\{\begin{align*}
c & \text { if } x>c,  \tag{3.5}\\
x & \text { if }|x| \leq c \\
-c & \text { if } x<-c .
\end{align*}\right.
$$

A value of $c$ around 2 is often used in calculating GMestimators. See for example Hampel et al., (1986), Gwet and Rivest (1992) and Hulliger (1995).

The choice of $h_{k}$ makes it possible to limit the influence of auxiliary information that is too extreme. The constant $\alpha=0$ leads to the choice of Mallows whereas $\alpha=1$ makes it possible to obtain the Schweppe version. The Schweppe version is sometimes preferred. See Coakley and Hettmansperger (1993) and Hampel et al., (1986, p. 322). When there is minimal auxiliary information, i.e., when we only have available a real variable $x_{k}, \forall k \in U$, a possible choice for function $h_{k}$ is

$$
\begin{equation*}
h_{k}=\min \left(1, \frac{t}{x_{k} / \operatorname{med}\left(x_{k}\right)}\right) \tag{3.6}
\end{equation*}
$$

For a design $\pi \mathrm{ps}$, a modification of $h_{\mathrm{k}}$ following Dalén (1987) so as to take various sampling weights into consideration would perhaps be desirable. The constant $t$ must be specified by the statistician. A value of $t$ around 1.5 is found in the applications. See for example Rivest and Rouillard (1991), who also provide other choices for functions $h_{k}$.

Writing $B_{g}$ as a weighted estimator makes it possible to write estimator (3.3) as a QR estimator with

$$
\left(q_{k}, r_{k}\right)=\left(d_{k} h_{k}^{1-a} u_{k} / c_{k}, r_{k}\right)
$$

The choice of constants $r_{k}$ is discussed in section 3.2.3.

### 3.2.2 Choice of $\boldsymbol{q}_{k}$ Based on a High-Breaking-Point Estimator

The choice of a GM-estimator is only a first step towards obtaining a very robust estimator of the total. In fact, although the influence function of GM-estimators is restricted, the fact remains that such estimators do not have a high breakdown point, which usually diminishes according to the dimension of the auxiliary information (Rousseeuw and Leroy 1987, p. 13). This section will explain how to build robust calibration estimators based on high breakdown point estimators. As such estimators do not usually assume a weighted form, we will consider reweighting them. This will allow us to obtain, as in the previous section, the constants $q_{k}$ needed to compute the RQR estimator metric. Specifically, the following weights $\hat{u}_{k}$ are considered:

$$
\begin{equation*}
\hat{u}_{k}=\frac{\psi\left(\left(y_{k}-x_{k}^{\prime} \hat{B}_{0}\right) /\left(\sigma h_{k}^{a} \sqrt{c_{k}}\right)\right)}{\left(y_{k}-x_{k}^{\prime} \hat{B}_{0}\right) /\left(\sigma h_{k}^{a} \sqrt{c_{k}}\right)} \tag{3.7}
\end{equation*}
$$

where $\hat{B}_{0}$ is an equivariant estimator with a high breakdown point meeting certain regularity conditions. The reweighted estimator is

$$
\begin{align*}
\hat{B}_{r}= & \left(\sum_{s} d_{k} h_{k}^{1-\alpha} \hat{u}_{k} x_{k} x_{k}^{\prime} / c_{k}\right)^{-1} \times  \tag{3.8}\\
& \sum_{s} d_{k} h_{k}^{1-\alpha} \hat{u}_{k} x_{k} y_{k} / c_{k} .
\end{align*}
$$

The asymptotic properties of this type of estimator have been studied in Simpson and Chang (1997).

The estimator $\hat{B}_{0}$ that is considered is the one-step GMestimator of Coakley and Hettmansperger (1993). This estimator has a high breakdown point. It is obtained as the first iteration of the Newton formula in equation (3.4), where the Schweppe version is used, assuming $\alpha=1$. Other robust estimators could have been chosen. However, the efficiency and robust properties of the Coakley and Hettmansperger (1993) estimator make it a good choice. Thus, the proposed constant $q_{k}$ is

$$
q_{k}=d_{k} h_{k}^{1-\alpha} \hat{u}_{k} / c_{k}
$$

with $\hat{B}_{0}=\hat{B}_{\mathrm{CH}}, \hat{B}_{\mathrm{CH}}$ denoting the Coakley and Hettmansperger (1993) estimator.

### 3.2.3 Choice of $\boldsymbol{r}_{\boldsymbol{k}}$.

Once the constants $q_{k}$, have been determined, the constants $r_{k}$ must be selected. If $d_{k}=r_{k}$ then under general conditions, the QR estimator is an ADU estimator. However, such a choice of $r_{k}$ yields an estimator that is sensitive to outliers. Alternately, choosing $r_{k}=0$ provides a robust estimator that might be very biased as was emphasized in Gwet and Rivest (1992, p. 1180). Lee (1991) suggested choosing $r_{k}=\theta d_{k}$, where $\theta \in[0,1]$. The asymptotic bias becomes under general conditions $(\theta-1) \sum_{U} E_{k}$, where $E_{k}$ represents the residuals obtained by adjusting a robust estimator for the entire population. Choosing $\theta$ makes it possible to control estimator bias. The discussion in section 3 leads us to suggest constants $r_{k}$ that are close to the $d_{k}$ for good units, and reduced gradually for doubtful observations. We suggest choosing

$$
\begin{equation*}
r_{k}=d_{k} u_{k}^{*} \tag{3.9}
\end{equation*}
$$

where

$$
u_{k}^{*}=\frac{\psi^{*}\left(\left(y_{k}-x_{k}^{\prime} \hat{B}_{r}\right) /\left(\sigma h_{k}^{\alpha} \sqrt{c_{k}}\right)\right)}{\left(y_{k}-x_{k}^{\prime} \hat{B}_{r}\right) /\left(\sigma h_{k}^{\alpha} \sqrt{c_{k}}\right)}
$$

The function $\psi^{*}$ which we will be considering is a modification of the Huber function

$$
\Psi^{*}(x)= \begin{cases}x & \text { if }|x| \leq a,  \tag{3.10}\\ a & \operatorname{sign}(x) \\ \text { if }|x|>a \text { and }|x|<a / b \\ b x & \text { if }|x|>a / b\end{cases}
$$

We choose $a=9, b=1 / 4$. The reason for this modification is that we do not want the outliers comprising large residuals or extreme auxiliary information to have weights that are too reduced. In this way, the sampling weight is fully maintained when the argument for $u_{k}^{*}$ is between -9 and 9 , and reduced gradually to one quarter. If the weight of large residuals is reduced too much, then the bias becomes too great, leading to the choice of $\psi^{*}$. The choice of constants $r_{k}$ has been done empirically, and seems to work well in practice.

Thus, we will consider the choice of constants $q_{k}$ and $r_{k}$ following

$$
\begin{equation*}
\left(q_{k}, r_{k}\right)=\left(d_{k} h_{k}^{1-\alpha} \hat{u}_{k} / c_{k}, d_{k} u_{k}^{*}\right) . \tag{3.11}
\end{equation*}
$$

We suggest a generalization of the Lee proposition (1991), since instead of considering $r_{k}=d_{k} \theta$ where $\theta$ is fixed, $r_{k}=d_{k} u_{k}^{*}$ will adapt automatically (or adaptively) to the sample. Having this choice of constants $q_{k}$ and $r_{k}$ at our disposal, and with the usual mean square metric, we obtain a QR estimator, but it is subject to negative weighting problems. However, with the constants (3.11), we can consider the restricted mean square metric, solve the problem (2.3) and obtain a robust estimator meeting the CEs and the CARVs.

There are in proposition 1 possible solutions for the asymptotic behaviour of the resulting RQR estimator as compared to the QR estimator free of weight constraints. However, since the constants (3.11) depend on $s$ in a complex way, there can be no automatic conclusion about asymptotic equivalence. Nevertheless, the simulation study in section 4 seems to suggest a very comparable behaviour for the estimator with and without constraint on the weights, with respect to the Monte Carlo mean square error. Thus, empirical evidence shows that if the $q_{k}$ and $r_{k}$ are chosen in such a way that the estimator without constraint on the weights is robust, then the version with constraints on the weights will also be robust.

Finally, the following is a summary of the steps in the proposed method used to obtain a robust RQR estimator.

1. Choice of constants $q_{k}$ and $r_{k}$. We suggest the constants found in equation (3.11). For this step, it is necessary to compute $\hat{B}_{\mathrm{CH}}$.
2. Choice of metric. If need be, choice of constants $L$ and $U$. These constants are chosen such that $L \leq r_{k} \leq U, \forall k \in s$.
3. Solution using Newton's method for equation (2.5).
4. Assume $w_{k}=h\left(x_{k}{ }^{\prime} \lambda_{\infty} ; q_{k}, r_{k}\right)$ for $\lambda_{\infty}$ solution to step 3 .
5. Assume $\hat{T}_{y r}=\sum_{s} w_{k} y_{k}$, which is the proposed RQR estimator.

The procedure requires a certain number of constants. The constants $\alpha, t$ and $c$ are found in the calculation of $q_{k}$ and $r_{k}$. The choice of these values is nevertheless justified using robustness theory, which helps guide the practitioner.

Thus, the value for $c$ in the Huber function can be obtained by taking into account efficiency concerns under normal errors. See Hampel et al., (1986, p. 333) and Gwet and Rivest (1992). Constants $a$ and $b$ are also found; they are more directly linked to the proposed estimators. Constant $b$ represents the maximum weight reduction that can be allowed when specifying the default weights $r_{k}$, and for this reason there is a link with the suggestion made by Lee (1991). The constant which it is most important to specify is possibly the value of $a$. We suggest here $a=9$. However, in our simulations, a value of $a$ between 6 and 12 yielded relatively comparable results. The choice of limits $L$ and $U$ rests on cosmetic considerations, so that the weights may be limited to one interval. This last consideration is perhaps secondary for the practitioner. As a result, it would seem that the most important aspect is to choose a value of $r_{k}$ that is close to $d_{k}$ for the proper values, then reduced as an observation is deemed extreme, and that is the goal which has guided our choice of $r_{k}$ in this section. Nevertheless, it would be useful to make a choice of $r_{k}$ that satisfies a certain optimality criterion.

### 3.3 Chambers Model-Assisted Estimator

Another approach is based on a decomposition proposed by Chambers $(1982,1986)$ which we now apply to QR estimators. Note that a QR estimator can always be written in the form

$$
\hat{T}_{y \mathrm{yR}}=\sum_{s} r_{k} y_{k}+\left(T_{x}-\hat{T}_{x r}\right)^{\prime} B+\sum_{s} z_{k} \sqrt{q_{k}}\left(y_{k}-x_{k}^{\prime} B\right)
$$

where $z_{k}=\left(T_{x}-\hat{T}_{x r}\right)^{\prime}\left(\sum_{s} q_{k} x_{k} x_{k}^{\prime}\right)^{-1} x_{k} \sqrt{q_{k}}, \hat{T}_{x r}=\sum_{s} r_{k} x_{k}$ and $B$ are arbitrary. Chambers (1986) had considered the specific case $\left(q_{k}, r_{k}\right)=\left(1 / \sigma_{k}^{2}, 1\right)$ for the ratio estimator. In order to limit the influence of outlier units, Chambers proposed

$$
\begin{gather*}
\hat{T}_{\text {yCHAM }}=\sum_{s} r_{k} y_{k}+\left(T_{x}-\hat{T}_{x r}\right)^{\prime} B+ \\
\sum_{s} z_{k} \psi_{p}\left(\sqrt{q_{k}}\left(y_{k}-x_{k}^{\prime} B\right)\right) . \tag{3.12}
\end{gather*}
$$

The function $\psi_{p}$ helps limit the influence of large residuals. The choice for $B$ is a robust estimator, e.g., $\hat{B}_{g}$. One function $\psi_{p}$ considered in Chambers (1986) was

$$
\begin{equation*}
\psi_{p}(t)=t \exp \left(-0.25(|t|-6)^{2}\right) \tag{3.13}
\end{equation*}
$$

It is interesting to note that (3.12) can be written as

$$
\hat{T}_{y \mathrm{CHAM}}=T_{x}^{\prime} B+\sum_{s}\left(r_{k}+\left(d_{k} g_{k}-r_{k}\right) \lambda_{k}\right) e_{k}(B),
$$

where $e_{k}(B)=y_{k}-x_{k}^{\prime} B, g_{k}$ is defined in formula (2.1) calculated using $q_{k}$ and $r_{k}$, and

$$
\begin{equation*}
\lambda_{k}=\frac{\psi_{p}\left(\sqrt{q_{k}}\left(y_{k}-x_{k}^{\prime} B\right)\right)}{\sqrt{q_{k}}\left(y_{k}-x_{k}^{\prime} B\right)} . \tag{3.14}
\end{equation*}
$$

Thus, the residuals $e_{k}(B)$ are weighted using a relation referring to formula (3.2). If $\lambda_{k} \equiv 1$, then $d_{k} g_{k}$ is applied to residuals $e_{k}(B)$ and it is easy to verify that we have the estimator $\hat{T}_{y \text { QR }}^{k}$. Altemately, if $\lambda_{k} \equiv 0$, we obtain (3.3) if we assume $B=\hat{B}_{g}$. If in (3.12) we assume $\left(q_{k}, r_{k}\right)=\left(1 / c_{k}, 1\right)$ and $B=\hat{B}_{g}$, then the Chambers estimator represents a compromise between the BLUP and a robust estimator based on a GM-estimator. Note that formally (3.12) is a QR estimator with

$$
\left(q_{k}^{*}, r_{k}^{*}\right)=\left(d_{k} h_{k}^{1-a} u_{k} / c_{k}, r_{k}+\left(d_{k} g_{k}-r_{k}\right) \lambda_{k}\right) .
$$

However, since $r_{k}^{*}$ is not necessarily positive, it is not always possible to undertake a change of metric in this case.

## 4. EMPIRICAL STUDY

To study the performance of robust calibration estimators, we carried out a Monte Carlo simulation study. We considered four populations comprising data from readily available works on sampling theory. For each population, $K=2000$ samples were drawn using simple
random sampling for various sample sizes. Our main objective was to determine whether it is possible to obtain estimators having good empirical properties (bias, mean square error) while satisfying the CEs and the CARVs. Note that all the programs were written in S-PLUS (Statistical Sciences 1991) and are available from the author.

### 4.1 Populations Under Study

The population graphs can be found in Figure (4.1). The first population, comprising 51 units, can be found in Mosteller and Tukey (1977, p. 560). It consists of the U.S. population in 1960 and in 1970 for each of the 50 states and the federal district of Columbia. It is called POPUSA. Looking at the scattergram of the 1970 population in terms of the 1960 population, we notice that all units seem to be on the same straight line, with some good leverage points. An example of a good leverage point is the point surrounded in this population. The second population, with 34 units, can be found in Singh and Chaudhary (1986, p. 177). It deals with the area of fields sown in 1971 and in 1974. This population is called AREA. There is a bad leverage point (see the surrounded point) in this population since the point (4170.99) does not respect the linear trend


Figure 4.1. The four populations under study
of the majority of units. Samples of size 10 and 15 are drawn from POPUSA and AREA. The third population, i.e. the MU284 in Särndal et al., (1992), comprises the 284 municipalities of Sweden. We considered variables $x=S 82$ concerning the total number of seats in the municipal council, and $y=P 85$ representing the population of Sweden in 1985. There are vertical outliers (e.g., the surrounded point) and one bad leverage point. Finally, we considered population made up of MU281 made up of MU284 from which the three largest municipalities were excluded. The variables considered were $x=$ REV84 representing the values of landed property based on the 1984 assessment, and $y=$ RMT85 representing municipal tax revenues in 1985. The unit of measurement was one million kronor for both variables. It seems this population has several bad leverage points. Samples of size $n=30$ and $n=60$ were drawn from MU284 and MU281. Table (4.1) contains totals for various populations.

Table 4.1
Totals for Various Populations and Totals Known From Auxiliary Information

| Population | $T_{x}$ | $T_{y}$ | $N$ |
| :--- | ---: | ---: | ---: |
| POPUSA | 179,972 | 203,923 | 51 |
| AREA | 29,118 | 6,781 | 34 |
| MU284 | 13,500 | 8,339 | 284 |
| MU281 | 757,246 | 53,124 | 281 |

### 4.2 Description of the Estimators

The two basic estimators were the GREG estimator and the estimator obtained by considering case No. 7 in Deville and Särndal (1992), i.e., a GREG estimator with restricted weights. These estimators were denoted GREG/U and

GREG/R respectively. We selected $c_{k} \equiv 1$ for populations POPUSA and AREA, and chose $c_{k}=x_{k}$ for populations MU284 and MU281. Our choice for the $c_{k}$ was motivated by the relationship between these constants and the heteroscedasticity of the superpopulation model. Of the robust estimators, we studied the Chambers (1986) estimator by considering

$$
\left(q_{k}^{*}, r_{k}^{*}\right)=\left(d_{k} \hat{u}_{k}\left(\hat{B}_{\mathrm{CH}}\right) / c_{k}, 1+\left(d_{k} g_{k}-1\right) \lambda_{k}\left(\hat{B}_{r}\right)\right),
$$

where in the formula $(2.1)\left(\left(q_{k}, r_{k}\right)=\left(1 / c_{k}, 1\right)\right.$, denoted CHAM, based on $\hat{B}_{r}$. The constants $\hat{u}_{k}\left(\hat{B}_{C H}\right)$ were obtained from formula (3.7). Selection $\alpha=1$ was used throughout the simulation. Huber's function $\psi$ was used with the constant $c=1.345$ for $\hat{B}_{r}$. The functions $\lambda_{k}$ are those given by formula (3.6), where we selected $t=1.46$. The function $\lambda_{k}$ is defined by equation (3.14). The function $\psi_{p}$ considered was that given by equation (3.13). The scale was estimated as in Coakley and Hettmansperger (1993). We also considered the model-assisted BLUP estimator in which the generalized least squares estimator was replaced by estimator $\hat{B}_{r}$, which we called MODEL. Moreover, we considered the Lee (1991) estimator on the basis of $\hat{\boldsymbol{B}}_{r}$ where $r_{k}=0,25 d_{k}$, using the mean square metric. We also studied an extension of the Lee (1991) estimator by considering the limited mean square metric. These estimators were denoted by LEE25/U and LEE25/R respectively. Finally, we considered the new method in section 3.2.3, selecting ( $q_{k}, r_{k}$ ) as given by equation ( 3.11 ) in accordance with the mean square metric and the limited mean square metric. They were denoted by QRROB/U and QRROB/R respectively. The choice of function $\psi^{*}$ was given by formula (3.10).

Table 4.2
Monte Carlo Results for Sampling From the POPUSA Population

| Estimators | $\mathrm{VAR}_{\mathrm{M}}$ | $\mathrm{MSE}_{\mathrm{M}}$ | $\mathrm{CV}_{\mathrm{m}}$ | $\mathrm{BR}_{\mathrm{M}}$ | MIN | MAX | CARV ${ }^{1}$ | CONV |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| $\underline{n=10}$ |  |  |  |  |  |  |  |  |
| GREG/U | 34.90 | 34.92 | 2.90 | -0.07 | -6.24 | 26.75 | 86.7 |  |
| GREG/R | 35.29 | 35.30 | 2.91 | -0.04 | 0.20 | 32.00 | 100.0 | 98.4 |
| CHAM | 32.43 | 33.75 | 2.85 | -0.56 | -19.61 | 40.96 | 84.0 |  |
| MODEL | 27.66 | 30.69 | 2.72 | -0.85 | -19.71 | 40.86 | 82.8 |  |
| LEE25/U | 27.48 | 30.07 | 2.69 | -0.79 | -19.38 | 39.64 | 83.2 |  |
| LEE25/R | 28.67 | 30.90 | 2.73 | -0.73 | 0.20 | 32.00 | 100.0 | 98.4 |
| QRROB/U | 27.40 | 28.40 | 2.61 | -0.49 | -15.68 | 40.10 | 83.2 |  |
| QRROB/R | 28.33 | 29.18 | 2.65 | -0.45 | 0.20 | 32.00 | 100.0 | 98.4 |
| $\underline{n}=15$ |  |  |  |  |  |  |  |  |
| GREG/U | 21.90 | 21.95 | 2.30 | -0.10 | -3.13 | 15.32 | 94.7 |  |
| GREG/R | 22.12 | 22.15 | 2.31 | -0.09 | 0.20 | 16.00 | 100.0 | 99.5 |
| CHAM | 18.11 | 20.14 | 2.20 | -0.70 | -5.79 | 16.44 | 92.4 |  |
| MODEL | 15.43 | 19.03 | 2.14 | -0.93 | -6.09 | 16.92 | 91.0 |  |
| LEE25/U | 15.44 | 19.54 | 2.17 | -0.99 | -6.19 | 17.06 | 90.8 |  |
| LEE $25 / \mathrm{R}$ | 15.72 | 19.68 | 2.18 | -0.98 | 0.20 | 16.00 | 100.0 | 99.5 |
| QRROB/U | 14.68 | 16.44 | 1.99 | -0.65 | -4.48 | 16.41 | 90.9 |  |
| QRROB/R | 14.85 | 16.56 | 2.00 | -0.64 | 0.20 | 16.00 | 100.0 | 99.5 |

[^5]
### 4.3 Frequency Measurements

The eight estimators in Section (4.2) were calculated for each sample. The results can be found in Tables 4.2, 4.3, 4.4 and 4.5. Since one asset of the new methods is the CARVs, statistics were calculated on these weights. Columns MIN and MAX in the tables of results contain the minimum and maximum values of weights calculated during the simulation for each estimator. Also shown is the percentage of samples for which the weights are within the CARVs in the CARV column in the tables of results. We also considered the percentage of samples for which there were convergent limited estimators in the CONV column. The intervals used $[L, U]$ for the limited intervals are specified in the different tables. In all cases, the various statistics were calculated using samples for which all estimators were convergent.

Another significant feature is related to the bias and efficiency of the proposed methods. Let $\hat{T}$ denote an estimator of the total $T_{y}$. Assume $\hat{T}_{i}$ is the estimator of the total calculated using sample $i, i=1, \ldots, K$. The relative Monte Carlo bias $\mathrm{BR}_{M}$, the mean value $E_{M}$ and the variance $V_{M}$ are given by the usual formulas, i.e.,

$$
\begin{aligned}
\mathrm{BR}_{M}= & \left(E_{M}(\hat{T})-T_{y}\right) / T_{y} \times 100, E_{M}(\hat{T})= \\
& \frac{1}{K} \sum_{i=1}^{K} \hat{T}_{i} \text { and } V_{M}=\frac{1}{K} \sum_{i=1}^{K}\left(\hat{T}_{i}-E_{M}(\hat{T})\right)^{2} .
\end{aligned}
$$

Our main criterion for efficiency will be the Monte Carlo mean square error defined by $\mathrm{MSE}_{M}=1 / K \sum_{i=1}^{K}\left(\hat{T}_{i}-T_{y}\right)^{2}$. The coefficients of variation $\mathrm{CV}_{M}$ are calculated in accordance with $\sqrt{\mathrm{MSE}_{\mathrm{M}}} / T_{y}$. The variance and mean square error are expressed in millions. The coefficient of variation, the relative bias, the CARVs and the convergence of limited versions are expressed as percentages.

### 4.4 Discussion

The POPUSA population had no outliers that did not satisfy the linear model. During sampling, the coefficients of variation of the estimators were small, which could be expected given the trend of the population. Columns MSE and VAR are very similar, indicating that bias is not a problem for this population. All relative bias was less than $1 \%$. The QRROB/U estimator provided a reduction in variance as compared to GREG/U that exceeded $21 \%$ for $n=10$ and $30 \%$ for $n=15$.

The size of the AREA population was small. This population had a bad leverage point leading to very high empirical relative bias for all the estimators. The GREG/U estimator had a relative bias of more than $7 \%$ in spite of a 44\% sampling for this population. The robust estimators had the most significant bias, though it was relatively comparable to the bias of the GREG/U estimator. The most
significant reduction in variance was achieved for the QRROB/U estimator, but at the cost of a relative bias of about $10 \%$.

Population MU284 had a vertical outlier and bad leverage points. Robust estimators reduced the variance radically, since they were not affected by the three extreme units in $y$ which were clearly moving away from the linear trend. The CHAM, QRROB/R and QRROB/U estimators were more than four times less variable than the GREG/U estimator. However, this led to a much higher negative bias. All the robust estimators were severely biased. The MODEL estimator showed a negative bias of more than $13 \%$, whereas QRROB/U had a negative bias of the order of $11 \%$. As for QRROB, a better choice of constants in function $\psi^{*}$ might help reduce a larger part of the bias at the cost of a lower variance reduction. Increasing the sample size to $n=60$ made it possible to reduce the bias below the $10 \%$ of the CHAM and QRROB estimators, but the other robust estimators remained more biased.

Population MU281 contained a fairly large number of bad leverage points. The variance dominated the MSE share of this population. The LEE25 estimator was the least variable, with a reduction of more than $35 \%$ as compared to GREG/U. However, although $\theta=0.25$ functions well for this population, our study shows that it is not always the best choice.

Note that all the robust estimators were more efficient than the GREG or its limited version. As was confirmed by the results of Deville and Särndal (1992), the limited version of the GREG estimator showed essentially the same behaviour as the GREG in terms of both bias and Monte Carlo variance for each population. Of all the estimators that were considered, GREG/U and GREG/R were the least biased. The robust versions all exhibited greater bias. However, this is more than offset by the reduction in variance so that the efficiency of robust estimators is always greater than that of GREG/U or of GREG/R estimators.

Concerning the constraints on the weights, it will be noted that the GREG/U, CHAM, MODEL, LEE 25 and QRROB/U estimators are all subject to problems of negative weighting, as can be seen in column MIN. This problem is avoided with limited estimators. The CARV column shows that the constraints were not met relatively frequently, depending on population and sample size, varying between $5 \%$ and $60 \%$. The general behaviour of the two limited robust estimators was comparable to that of their non-limited versions. Moreover, QRROB/R, in addition to meeting the CARVs, provided interesting properties of efficiency, as compared to other robust estimators. Limited versions were not as prone to convergence problems when sample sizes were greater. Note that we had to use wider bands in the case of POPUSA in order to obtain satisfactory convergence rates.

Table 4.3
Monte Carlo Results for Sampling From the AREA Population

| Estimators | VAR $_{\mathrm{M}}$ | MSE $_{\mathrm{M}}$ | $\mathrm{CV}_{\mathrm{M}}$ | BR $_{\mathrm{M}}$ | MIN | MAX | CARV $^{\prime}$ | CONV |
| ---: | ---: | ---: | ---: | ---: | ---: | ---: | ---: | ---: |
| $=10$ |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |
| GREG/U | 1.334 | 1.700 | 19.23 | 8.92 | -3.35 | 14.94 | 86.6 |  |
| GREG/R | 1.295 | 1.629 | 18.82 | 8.53 | 0.20 | 14.00 | 100.0 | 99.0 |
| CHAM | 1.187 | 1.541 | 18.30 | 8.77 | -4.09 | 14.90 | 87.2 |  |
| MODEL | 1.291 | 1.580 | 18.54 | 7.93 | -5.23 | 16.75 | 86.8 |  |
| LEE25/U | 1.279 | 1.593 | 18.61 | 8.26 | -5.28 | 16.89 | 86.6 |  |
| LEE25/R | 1.284 | 1.596 | 18.63 | 8.24 | 0.20 | 14.00 | 100.0 | 99.0 |
| QRROB/U | 1.026 | 1.440 | 17.70 | 9.50 | -4.74 | 15.38 | 87.6 |  |
| QRROB/R | 1.028 | 1.437 | 17.68 | 9.43 | 0.20 | 14.00 | 100.0 | 99.0 |
| n=15 |  |  |  |  |  |  |  |  |
| GREG/U | 0.940 | 1.178 | 16.00 | 7.18 | -1.40 | 7.03 | 93.0 |  |
| GREG/R | 0.928 | 1.154 | 15.85 | 7.01 | 0.20 | 6.00 | 100.0 | 99.8 |
| CHAM | 0.708 | 0.989 | 14.67 | 7.82 | -1.52 | 7.92 | 93.7 |  |
| MODEL | 0.757 | 0.997 | 14.73 | 7.22 | -1.66 | 8.39 | 93.1 |  |
| LEE25/U | 0.672 | 1.059 | 15.18 | 9.18 | -1.68 | 9.40 | 92.0 |  |
| LEE25/R | 0.671 | 1.056 | 15.15 | 9.15 | 0.20 | 6.00 | 100.0 | 99.8 |
| QRROB/U | 0.485 | 0.990 | 14.68 | 10.48 | -1.59 | 8.90 | 93.9 |  |
| QRROB/R | 0,485 | 0.986 | 14.64 | 10.44 | 0.20 | 6.00 | 100.0 | 99.8 |

${ }^{1}$ The limits for the CARVs are $[0.20,14]$ for $n=10$ and $[0.20,6]$ for $n=15$.

Table 4.4
Monte Carlo Results for Sampling From the MU284 Population

| Estimators | VAR $_{M}$ | MSE $_{M}$ | CV $_{M}$ | BR $_{M}$ | MIN | MAX | CARV $^{1}$ | CONV |
| ---: | ---: | ---: | ---: | ---: | ---: | ---: | ---: | ---: |
| $n=30$ |  |  |  |  |  |  |  |  |
| GREG/U | 2.833 | 2.925 | 20.51 | -3.64 | -6.83 | 23.90 | 89.8 |  |
| GREG/R | 2.813 | 2.910 | 20.46 | -3.73 | 0.20 | 16.00 | 100.0 | 99.2 |
| CHAM | 0.645 | 1.639 | 15.35 | -11.95 | -11.80 | 31.26 | 77.0 |  |
| MODEL | 0.709 | 2.037 | 17.11 | -13.82 | -12.06 | 31.91 | 68.40 |  |
| LEE25/U | 0.887 | 1.877 | 16.43 | -11.93 | -11.06 | 30.93 | 73.5 |  |
| LEE25/R | 0.871 | 1.847 | 16.30 | -11.85 | 0.20 | 26.00 | 100.0 | 99.2 |
| QRROB/U | 0.719 | 1.532 | 14.84 | -10.81 | -9.46 | 25.84 | 86.5 |  |
| QRROB/R | 0.720 | 1.525 | 14.81 | -10.76 | 0.20 | 16.00 | 100.0 | 99.2 |
| n=60 |  |  |  |  |  |  |  |  |
| GREG/U | 1.473 | 1.489 | 14.63 | -1.49 | -1.19 | 10.03 | 90.1 |  |
| GREG/R | 1.467 | 1.484 | 14.61 | -1.57 | 0.20 | 7.00 | 100.0 | 99.7 |
| CHAM | 0.357 | 0.990 | 11.93 | -9.54 | -2.53 | 15.59 | 69.8 |  |
| MODEL | 0.380 | 1.255 | 13.43 | -11.22 | -4.93 | 14.52 | 58.1 |  |
| LEE25/U | 0.403 | 1.201 | 13.14 | -10.72 | -4.80 | 14.20 | 60.3 |  |
| LEE25/R | 0.396 | 1.203 | 13.16 | -10.78 | 0.20 | 7.00 | 100.0 | 99.7 |
| QRROB/U | 0.308 | 0.976 | 11.85 | -9.80 | -2.36 | 10.99 | 86.1 |  |
| QRROB/R | 0.308 | 0.979 | 11.87 | -9.82 | 0.20 | 7.00 | 100.0 | 99.7 |

[^6]Table 4.5
Monte Carlo Results for Sampling From the MU281 Population

| Estimators | $\mathrm{VAR}_{\mathrm{M}}$ | MSE $_{\text {M }}$ | $\mathrm{CV}_{\text {M }}$ | $\mathrm{BR}_{\text {M }}$ | MIN | MAX | CARV ${ }^{1}$ | CONV |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| $\underline{n=30}$ |  |  |  |  |  |  |  |  |
| GREG/U | 17,33 | 17,35 | 7,84 | -0,26 | -38,97 | 34,56 | 86,0 |  |
| GREG/R | 17,40 | 17,41 | 7,86 | -0,24 | 0,20 | 25,00 | 100,0 | 99,8 |
| CHAM | 13,23 | 13,26 | 6,86 | -0,33 | -47,09 | 39,08 | 56,9 |  |
| MODEL | 11,30 | 11,91 | 6,50 | 1,47 | -66,22 | 41,43 | 47,9 |  |
| LEE25/U | 11,21 | 11,60 | 6,41 | 1,17 | -59,75 | 37,03 | 53,3 |  |
| LEE25/R | 11,26 | 11,73 | 6.45 | 1,29 | 0,20 | 25,00 | 100,0 | 99,8 |
| QRROB/U | 12,92 | 13,29 | 6,86 | 1,15 | -54,14 | 39,73 | 70,8 |  |
| QRROB/R | 12,94 | 13,34 | 6,88 | 1,20 | 0,20 | 25,00 | 100,0 | 99,8 |
| $n=60$ |  |  |  |  |  |  |  |  |
| GREG/U | 7,57 | 7,57 | 5,18 | -0,10 | -12,77 | 15,34 | 86,4 |  |
| GREG/R | 7,58 | 7,58 | 5,18 | -0,09 | 0,20 | 9,00 | 100,0 | 99,9 |
| CHAM | 5,85 | 5,90 | 4.57 | -0,43 | -22,97 | 11,49 | 51,4 |  |
| MODEL | 4,53 | 5,23 | 4,30 | 1,57 | -24,02 | 14,58 | 38,7 |  |
| LEE2S/U | 4,55 | 5,18 | 4,28 | 1,49 | -23,74 | 14,41 | 41,2 |  |
| LEE25/R | 4,50 | 5,21 | 4,30 | 1,58 | 0,20 | 9,00 | 100,0 | 99,9 |
| QRROB/U | 5,40 | 6,16 | 4,67 | 1,64 | -21,08 | 21,07 | 68,6 |  |
| QRROB/R | 5,39 | 6,17 | 4,67 | 1,66 | 0,20 | 9,00 | 100,0 | 99,9 |

${ }^{1}$ The limits for the CARVs are $[0.20,25]$ for $n=30$ and $[0.20,9]$ for $n=60$.

## 5. CONCLUSION

The goal of this paper has been to introduce calibration estimators having good properties of robustness. Traditional calibration estimators are easy to use, since it is sufficient to have a set of starting weights, usually the sampling weights $d_{k}$, which are transformed into calibrated weights. The steps used in this paper have been the same, i.e., the robust default weights $r_{k}$ have been transformed into calibrated weights, and the constants $q_{k}$ have been chosen such that $\hat{B}_{q}$ is a robust estimator. The proposed choice of $r_{k}$ is given by the formula (3.9), with $a=9$, $b=1 / 4$. There remains to develop a theory for the optimal choice of $r_{k}$. The suggestion is made for applications to vary constant $a$, between say 6 and 12 , in order to determine the influence of the constant on the estimation. The limits $L$ and $U$ can be used to limit the weights, e.g., to make them all positive. We suggest the general use of $L=0.2$, $U=k N / n$, where $k$ is about 3 .

Note that robust calibration estimators are not meant to replace the GREG estimator, but to be used in conjunction with it. Thus, if the robust estimator and the GREG estimator are very different, a more in-depth analysis might help determine the reason. The proposed estimators could be useful as diagnostic tools.

It would be interesting to pursue the empirical studies of section 4 , by examining for example the effect of sampling design on the proposed procedures. Another important area
of development is the estimation of variance. Multipurpose surveys are yet another area of interest. In fact, for applications, there is rarely a single variable of interest, and methodologists would like to use a single set of weights for all the variables of interest. In terms of robustness, a solution has been proposed in the conclusion of a paper by Gwet and Rivest (1992), where robust weights were calculated for each variable of interest $y^{(i)}, i=1, \ldots, I$. For one unit, the final weight corresponds to the minimum weight among the weights obtained. Alternately, to obtain robust and calibrated estimators, we could calculate robust default weights for each variable of interest, providing a set of $r_{k}\left(y^{(i)}\right)$, and assume $r_{k}=\min r_{k}\left(y^{(i)}\right)$, where the minimum is on $i=1, \ldots, I$. These weights could then be transformed into calibrated weights. This procedure should be assessed in greater detail.

## ACKNOWLEDGEMENTS

I wish to thank Carl-Erik Särndal for introducing me to sampling theory and for suggesting that I consider the problem of outliers in sampling theory. I also thank Roch Roy and Christian Léger for helping me during various stages of development of this paper. My sincere thanks go to the Associate Editor, the Assistant Editor and two referees for comments which led to significant improvements in both content and layout.

## APPENDIX A PROOF OF PROPOSITION 1

Let $\Delta_{h}(u ; q, r)=r+q u-h(u ; q, r)$ and $z_{k}$ be a variable of interest. We assume the following conditions
$C_{1} \cdot N^{-1} \sum_{s} q_{k} z_{k}=O_{p}(1)$.
$C_{2} . N^{-1} \sum_{s} \Delta_{h}\left(x_{k}^{\prime} \lambda_{\infty} ; q_{k}, r_{k}\right) z_{k}=o_{p}\left(n^{-1 / 2}\right)$,
where $\lambda_{\infty}$ is a solution to equation (2.5).
Note that $\sum_{s}\left(r_{k}+q_{k} x_{k}^{\prime} \lambda_{1}\right) x_{k}=T_{x}$, where $\lambda_{1}=$ $-\left(\sum_{s} q_{k} x_{k} x_{k}^{\prime}\right)^{-1}\left(\hat{T}_{x r}-T_{x}\right)$, and also that $\sum_{s} h\left(x_{k}^{\prime} \lambda_{\infty}\right.$; $\left.q_{k}, r_{k}\right) x_{k}=T_{x}$. Thus, in using $C_{2}$, we find that

$$
N^{-1} \sum_{s} q_{k} x_{k} x_{k}^{\prime}\left(\lambda_{1}-\lambda_{\infty}\right)=o_{p}\left(n^{-1 / 2}\right),
$$

$\lambda_{1}-\lambda_{\infty}=o_{p}\left(n^{-1 / 2}\right)$, and therefore using $C_{1}$, with $z_{k}=x_{k} x_{k}^{\prime}$. It is also easily shown that

$$
\begin{aligned}
& N^{-1}\left(\hat{T}_{\mathrm{QR}}-\hat{T}_{\mathrm{RQR}}\right) \\
& =N^{-1} \sum_{s}\left(r_{k}+q_{k} x_{k}^{\prime} \lambda_{1}\right) y_{k}-N^{-1} \sum_{s} h\left(x_{k}^{\prime} \lambda_{\infty} ; q_{k}, r_{k}\right) y_{k} \\
& =N^{-1} \sum_{s} q_{k} x_{k}^{\prime} y_{k}\left(\lambda_{1}-\lambda_{\infty}\right)+N^{-1} \sum_{s} \Delta_{h}\left(x_{k}^{\prime} \lambda_{\infty} ; q_{k}, r_{k}\right) y_{k} \\
& =o_{p}\left(n^{-1 / 2}\right) .
\end{aligned}
$$

## APPENDIX B LIST OF ABBREVIATIONS

ADU: Asymptotically Design Unbiased.
BLUP: Best Linear Unbiased Predictor (Royall 1970).
CARV: Constraints applicable to the range of values for the weights $w_{k}$, by requiring for example that all the $w_{k} \in[L, U]$.
CE: Calibration constraints, $\sum_{s} w_{k} x_{k}=T_{x}$, where $T_{x}=\sum_{U} x_{k}$.
CH: Robust estimator proposed by Coakley and Hettmansperger (1993), a single-step GMestimator that is robust and efficient.
CHAM: Robust Chambers $(1982,1986)$ estimator.
GM: Generalized $M$ estimators, derived from robustness theory (see for example Hampel et al., 1986).
GREG: Generalized regression estimator proposed by Cassel et al., (1976).
HT: Horvitz-Thompson estimator $\sum_{s} d_{k} y_{k}$, where $d_{k}=\pi_{k}^{-1}$.
QR: Wright (1983) estimators, in the form $T_{x}^{\prime} \hat{B}_{q}+\sum_{s} r_{k} e_{k}$.
RQR: Generalization of the Wright (1983) estimators, obtained using a general metric as well as constraints on the weights.

## APPENDIX C LIST OF THE PRINCIPAL CONSTANTS

$c_{k}$ : Factor capable of accounting for heteroscedasticity problems.
$d_{k}$ : Sampling weights.
$g_{k}$ : $g$-weight defined by $w_{k} / d_{k}$.
$h_{k}$ : Quantity used to reduce the influence of outlier auxiliary information in $\hat{\boldsymbol{B}}_{q}$.
$\pi_{k}, \pi_{k l}$ : Inclusion probabilities of first and second order, respectively.
$q_{k}, r_{k}$ : Quantities defining an estimator QR . The $q_{k}$ are used to build the regression coefficients involved in the first part, $T_{x}^{\prime} \hat{B}_{q}$ the $r_{k}$ are used for the second part, $\sum_{s} r_{k} e_{k}$.
$u_{k}, \hat{u}_{k}$ : Weights used to build $\hat{B}_{q}$ in a robust way.
$u_{k}^{*}$ : Weights used to consider a robust correction factor $\sum_{s} r_{k} e_{k}$.
$w_{k}$ : Calibrated weight attributed to $y_{k}$ to form $\sum_{s} w_{k} y_{k}$.

## REFERENCES

BERSHAD, M.A. (1960). Some Observations on Outliers. Unpublished memorandum, Statistical Research Division, U.S. Bureau of the Census.
BOLFARINE, H., and ZACKS, S. (1992). Prediction Theory for Finite Population. New York: Springer-Verlag.

BREWER, K.R.W. (1994). Survey sampling inference: Some past perspectives and present prospects. Pakistan Journal of Statistics 10, 213-233.

BRUCE, A.G. (1991). Robust Estimation and Diagnostics for Repeated Sample Surveys. Mathematical Statistics Working Paper 1991/1, Statistics New Zealand.
CASSEL, C.M., SÄRNDAL, C.-E., and WRETMAN, J.H. (1976). Some results on generalized difference estimation and generalized regression estimation for finite population. Biometrika 63, 615620.

CHAMBERS, R.L. (1982). Robust Finite Population Estimation. Ph. D. thesis, Johns Hopkins University, Dept. of Biostatistics.
CHAMBERS, R.L. (1986). Outlier robust finite population estimation. Journal of the American Statistical Association, 81, 1063-1069.

CHAMBERS, R.L., and KOKIC, P.N. (1993). An integrated approach for the treatment of outliers in sub-annual surveys. Proceedings on the $49^{\boldsymbol{h}}$ Session, International Statistical Institute.
COAKLEY, C.W., and HETTMANSPERGER, T.P. (1993). A bounded influence, high breakdown, efficient regression estimator. Journal of the American Statistical Association 88, 872-880.

DALÉN, J. (1987). Practical Estimators of a Population Total Which Reduce the Impact of Large Observations. R \& D Report, Statistics Sweden.

DEVILLE, J.-C., and SÄRNDAL, C.E. (1992). Calibration estimators in survey sampling. Journal of the American Statistical Association, 87, 376-382.
DONOHO, D.L., and HUBER, P.J. (1983). The notion of breakdown point. In A Festschrift for Erich Lehmann, (Eds. P.J. Bickel, K.A. Doksum and J.L. Hodges). Belmont, CA: Wadsworth.
FULLER, W.A., LOUGHIN, M.M., and BAKER, H.D. (1994). Regression weighting in the presence of nonresponse with application to the 1987-1988 Nationwide Food Consumption Survey. Survey Methodology, 20, 75-85.
GAMBINO, J. (1987). Dealing With Outliers: A Look at Some Methods Used at Statistics Canada. Technical report, Business Survey Division, Statistics Canada.
GROSS, W.F., BODE, G., TAYLOR, J.M., and LLOYD-SMITH, C.W. (1986). Some finite population estimators which reduce the contribution of outliers. Proceedings of the Pacific Statistical Congress. Aucklaud, New Zealand, 20-24 May 1985.
GWET, J.-P., and RIVEST, L.-P. (1992). Outlier resistant alternatives to the ratio estimator. Journal of the American Statistical Association, 87, 1174-1182.
HAMPEL, F.R., RONCHETTI, E.M., ROUSSEEUW, P.J., and STAHEL, W.A. (1986). Robust Statistics: The Approach Based on Influence Functions. New York: Wiley.
HIDIROGLOU, M.A., and SRINATH, K.P. (1981). Some estimators of the population total from simple random samples containing large units. Journal of the American Statistical Association, 76, 690-695.
HUBER, P.J. (1981). Robust Statistics. New York: Wiley.
HULLIGER, B. (1995). Outlier robust Horvitz-Thompson estimators. Survey Methodology, 21, 79-87.
KISH, L. (1965). Survey Sampling. New York: Wiley.
LEE, H. (1991). Model-based estimators that are robust to outliers. Proceedings of the 1991 Annual Research Conference. U.S. Bureau of Census.
LEE, H. (1995). Outliers in business surveys. In Business Survey Methods, (Eds. B.G. Cox, D.A. Binder, B.N. Chinnappa, A. Christianson, M.J. Colledge and P.S. Kott). New York: Wiley.

LEE, H., GHANGURDE, P.D., MACH, L., and YUNG, W. (1992). Outiers in Sample Surveys. Methodology Branch Working Paper BSMD-92-008E, Statistics Canada.
MOSTELLER, F., and TUKEY, J.W. (1977). Data Analysis and Regression, a Second Course in Statistics. Redding, MA: Addison-Wesley.

RIVEST, L.P., and ROUILLARD, E. (1991). M-estimators and outlier resistant alternatives to the ratio estimator. Proceedings: Symposium 90, Measurement and Improvement of Data Quality, Statistics Canada, 245-257.
ROUSSEEUW, P.J., and LEROY, A.M. (1987). Robust Regression and Outlier Detection. New York: Wiley.
ROYALL, R.M. (1970). On finite population sampling under certain linear regression models. Biometrika 57, 377-387.
SÄRNDAL, C.-E. (1996). Efficient estimators with simple variance in unequal probability sampling. Journal of the American Statistical Association, 91, 1289-1300.

SÄRNDAL, C.E., 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.
SÄRNDAL, C.-E., SWENSSON, B., and WRETMAN, J.H. (1992). Model Assisted Survey Sampling. New York: Springer-Verlag.
SIMPSON, D.G., and CHANG, Y.-C.I. (1997). Reweighted approximate GM-estimators: asymptotics and residual-based graphics. Journal of Statistical Planning and Inference, 57, 273293.

SINGH, A.C., and MOHL, C.A. (1996). Understanding calibration estimators in survey sampling. Survey Methodology, 22, 107-115.

SINGH, D., and CHAUDHARY, F.S. (1986). Theory and Analysis of Sample Survey Designs. New York: Wiley.
STATISTICAL SCIENCES, INC. (1991). S-PLUS Reference Manual. Seattle: Statistical Science, Inc.
STUKEL, D.M., HIDIROGLOU, M.A., and SÄRNDAL, C.-E. (1996). Variance estimation for calibration estimators: A comparison of jackknifing versus Taylor linearizarion. Survey Methodology, 22, 117-125.
TAMBAY, J.-L. (1988). An integrated approach for the treatment of outliers in sub-annual surveys. Proceedings on the Section on Survey Research Methods, American Statistical Association, 229234.

WELSH, A.H., and RONCHETTI, E. (1998). Bias-calibrated estimation from sample surveys containing outliers. Journal of the Royal Statistical Society, Series B, 60, 413-428.
WRIGHT, R.L. (1983). Finite population sampling with multivariate auxiliary information. Journal of the American Statistical Association, 78, 879-884.

# Estimation in Surveys Using Conditional Inclusion Probabilities: Complex Design 

YVES TILLÉ ${ }^{1}$


#### Abstract

This paper investigates a repeated sampling approach to take into account auxiliary information in order to improve the precision of estimators. The objective is to build an estimator with a small conditional bias by weighting the observed values by the inverses of the conditional inclusion probabilities. A general approximation is proposed in cases when the auxiliary statistic is a vector of Horvitz-Thompson estimators. This approximation is quite close to the optimal estimator discussed by Fuller and Isaki (1981), Montanari (1987, 1997), Deville (1992) and Rao (1994, 1997). Next, the optimal estimator is applied to a stratified sampling design and it is shown that the optimal estimator can be viewed as an generalised regression estimator for which the stratification indicator variables are also used at the estimation stage. Finally, the application field of this estimator is discussed in the general context of the use of auxiliary information.


KEY WORDS: Conditional estimation; Weighted observation; Generalised regression estimator; Complex survey.

## 1. INTRODUCTION

At the estimation stage, practitioners of survey sampling often have auxiliary information available. This information can be the knowledge of a set of population means or totals. Sometimes, the available information is detailed, for instance when the values taken by a variable on all the units of the population are known. This information can be used to improve the precision of the estimators.

Our aim is to dealt with the use of auxiliary information based on a conditional principle. Conditional inference has been largely studied in the survey sampling literature. Indeed, the optimal estimator was discussed by Fuller and Isaki (1981), Montanari (1987, 1997), Deville (1992) and Rao (1994, 1997). The conditional properties of the poststratified estimators has been studied by Casady and Valliant (1993). In an earlier paper (Tillé 1998), a general technique that allows to build a mean or total estimator that has a small conditional bias has been proposed for simple random sampling. This technique is based on the use of conditional inclusion probabilities and allows one to take into account auxiliary information without any reference to a superpopulation model.

In this paper the use of conditional inclusion probabilities is generalised to any sampling design. It is shown that this technique allows to construct an estimator very similar to the optimal estimator discussed by Montanari (1987), Deville (1992) and Rao (1994). This family of estimators provides a valid conditional inference and can also be viewed as the optimal linear estimator. Next, these estimators are applied in the stratification case and are compared to the GREG-estimator. The GREG-estimator is generally conditionally biased. Nevertheless, it is shown that, in regression, the optimal estimator is a particularly
case of the GREG-estimator. Indeed, when the stratification variables are re-used as auxiliary variables in the GREGestimator, it is equal to the optimal estimator. Next, a set of simulations is given that shows the interest of the optimal estimator in stratification. The gain of precision can be very important when the stratification variables are very correlated to the interest variable. Finally we discuss the general estimation problem in survey sampling that can be viewed as a third-order problem where three sets of variables interact: the planning variables, the calibration variables and the interest variables.

The paper is organised as follows. In section 2 the notation is defined. In section 3, the problem of conditional inference is presented. In section 4, an approximation of the SCW-estimator is given for complex designs under technical hypotheses. These hypotheses are discussed in section 5. In section 6 the optimal estimator and the SCWestimator are compared to the generalised regression (GREG) estimator in the stratification framework. It is shown that the optimal estimator can be viewed as a GREG-estimator for which the stratification indicator variables are also used a posteriori. Next a set of simulations is presented in section 7 in order to compare the discussed estimators. Finally, the problem of interaction between the design and the auxiliary variables is discussed in section 8 .

## 2. PROBLEM AND NOTATION

Consider a finite population $U=\{1, \ldots, k, \ldots, N\}$ and suppose that a random sample $S$ is drawn without replacement from this population following a sampling design $p$ (.). The probability of selecting the sample $s$ is $\operatorname{Pr}(S=s)=p(s)$,

[^7]for all $s \subset U$. The indicator variables $I_{k}$ take the value 1 if unit $k$ is in the sample and 0 otherwise, for all $k \in U$. The inclusion probability of unit $k$ is $\pi_{k}=E\left(I_{k}\right)$, where symbol $E($.$) is the expectation with respect to the sampling design.$ The joint inclusion probability for unit $k$ and $l$ is $\pi_{k l}=$ $E\left(I_{k} I_{l}\right)$. Let $y_{k}$ denote the value of the variable $y$ for the $k$-th unit of the population. The aim is to estimate the population mean of $y$ :
$$
\bar{y}=\frac{1}{N} \sum_{k \in U} y_{k} .
$$

If $\pi_{k}>0$, for all $k \in U$ the Horvitz-Thompson estimator (1952) given by

$$
\hat{\bar{y}}=\frac{1}{N} \sum_{k \in S} \frac{y_{k}}{\pi_{k}}
$$

provides an unbiased estimator of $\bar{y}$.
Let $T$ be a statistic. The objective is to estimate $\bar{y}$ with a conditional bias as small as possible with respect to statistic $T$. Define the first-order conditional inclusion probabilities to be $\pi_{k \mid T}=E\left(I_{k} \mid T\right)$ for all $k \in U$ and the conditional joint inclusion probabilities to be $\pi_{k| | T}=$ $E\left(I_{k} I_{l} \mid T\right)$ for all $k \in U, l \in U, k \neq l$. The simple conditionally weighted estimator (SCW) is defined by

$$
\begin{equation*}
\hat{\bar{y}}_{\mid T}=\frac{1}{N} \sum_{k \in S} \frac{y_{k}}{\pi_{k \mid T}} \tag{2.1}
\end{equation*}
$$

This estimator is not exactly conditionally unbiased. Indeed, a conditionally unbiased estimator exists if and only if $\pi_{k \mid T}>0$ for all $k \in U$. For this reason, it is useful to enlarge the definition of conditional unbiasedness: an estimator is said to be virtually conditionally unbiased (VCU) if the conditional bias only depends on the units having null conditional inclusion probabilities. The SCW-estimator is VCU, indeed:

$$
B\left(\hat{\bar{y}}_{\mid T} \mid T\right)=E\left(\hat{\bar{y}}_{\mid T} \mid T\right)-\bar{y}=-\frac{1}{N} \sum_{k \mid \pi_{k \mid T}=0} y_{k}
$$

This estimator generalises some classic results (see Tille 1998) like post-stratification. Moreover, it allows us to build an original estimator for a contingency table when the population marginal totals are known. Unfortunately, the computation of the $\pi_{k \mid T}$ becomes very difficult in complex sampling designs. A general approximation for the SCWestimator will however be given when using a vector of Horvitz-Thompson estimators as auxiliary statistic.

## 3. USE OF A COMPLEX AUXILIARY STATISTIC

Suppose that the auxiliary information is represented by the vector $\mathbf{x}_{k}=\left(x_{k 1}, \ldots, x_{k j}, \ldots, x_{k j}\right)^{\prime}$ of values taken by the $J$ auxiliary variables on the $k$-th unit of $U$. In a first step, it is
supposed that the $\mathbf{x}_{k}$ are known for each unit of the population. Later, it will be considered the more restrictive case where only one function of the $\mathbf{x}_{k}$ such as

$$
\overline{\mathbf{x}}=\frac{1}{N} \sum_{k \in U} \mathbf{x}_{k}
$$

is known. Consider also the Horvitz-Thompson estimator of $\overline{\mathbf{x}}$ given by

$$
\hat{\mathbf{x}}_{\pi}=\frac{1}{N} \sum_{k \in S} \frac{\mathbf{x}_{k}}{\pi_{k}}
$$

If $\pi_{k}>0$, for all $k \in U, \hat{\mathbf{x}}_{\boldsymbol{\pi}}$ is an unbiased estimator of $\overline{\mathbf{x}}$

$$
\begin{equation*}
E\left(\hat{\hat{\mathbf{x}}_{\pi}}\right)=\overline{\mathbf{x}} \tag{3.2}
\end{equation*}
$$

The variance of $\hat{\bar{x}}_{\boldsymbol{\pi}}$ is given by

$$
\begin{gather*}
\mathbf{\Sigma}=\operatorname{Var}\left(\hat{\mathbf{x}}_{\pi}\right)=\frac{1}{N^{2}} \sum_{l \in U} \mathbf{x}_{l} \frac{\mathbf{x}_{l}}{\pi_{l}}\left(1-\pi_{l}\right)+ \\
\frac{1}{N^{2}} \sum_{l \in U} \sum_{\substack{m \in U \\
m \neq l}} \frac{\mathbf{x}_{l} \mathbf{x}_{m}^{\prime}}{\pi_{l} \pi_{m}}\left(\pi_{l m}-\pi_{l} \pi_{m}\right) . \tag{3.3}
\end{gather*}
$$

Suppose now that vector ( $\left.\hat{\bar{y}}_{\pi} \hat{\mathbf{x}}_{\pi}^{\prime}\right)^{\prime}$ has a multinormal distribution. Under this hypothesis, it can be derived a conditional unbiased estimator (see for instance, Deville 1992). First the conditional bias is computed:

$$
B\left(\hat{\bar{y}}_{\pi} \mid \hat{\mathbf{x}}_{n}\right)=E\left(\hat{\bar{y}}_{\pi} \mid \hat{\mathbf{x}}_{n}\right)-\bar{y}=\left(\hat{\mathbf{x}}_{\pi}-\overline{\mathbf{x}}\right) \operatorname{Var}\left(\hat{\mathbf{x}}_{\pi}\right)^{-1} \operatorname{Cov}\left(\hat{\mathbf{x}}_{\pi} \hat{\bar{y}}_{\pi}\right)
$$

If an estimator of $B\left(\hat{\bar{y}}_{\pi} \mid \hat{\mathbf{x}}_{\pi}\right)$ is available, the HorvitzThompson estimator can be corrected in the following way:

$$
\begin{aligned}
\hat{\bar{y}}_{c} & =\hat{\bar{y}}_{\pi}-\hat{B}\left(\hat{\bar{y}}_{\pi} \mid \hat{\bar{x}}_{\pi}\right) \\
& =\hat{\bar{y}}_{\pi}+\left(\overline{\mathbf{x}}-\hat{\mathbf{x}}_{\pi}\right) \widehat{\operatorname{Var}\left(\hat{\mathbf{x}}_{\pi}\right)^{-1} \widehat{\operatorname{Cov}}\left(\hat{\mathbf{x}}_{\pi}, \hat{\bar{y}}_{\pi}\right) .}
\end{aligned}
$$

This estimator is related to the optimal linear estimator discussed by Fuller and Isaki (1981), Montanari (1987) and Rao (1994). Indeed, Montanari showed that the best estimator in the sense of the smallest mean square error (MSE) of the form

$$
\begin{equation*}
\hat{\bar{y}}_{\beta}=\hat{\bar{y}}_{\pi}+\left(\overline{\mathbf{x}}-\hat{\mathbf{x}}_{\pi}\right)^{\prime} \beta \tag{3.4}
\end{equation*}
$$

occurs when $\beta$ takes the value:

$$
\beta_{\mathrm{OPT}}=\Sigma^{-1} \operatorname{Cov}\left(\hat{\overline{\mathbf{x}}}_{\pi}, \hat{\bar{y}}_{\pi}\right)
$$

The optimal linear estimator presented by Montanari leads thus to a very similar result to the conditional approach, although Montanari did not start with a conditional point of view. In Montanari's approach, the optimal estimator is found in a class of linear estimators defined by (3.4) without any reference to conditional properties. Nevertheless, Rao (1994) has pointed out that this estimator leads to valid conditional inference. The general problem
of the optimal estimator is that $\beta_{\text {opt }}$ is not known and must thus be estimated. By estimating $\beta_{\text {opr }}$, the optimal properties of the estimator are lost.

In order to estimate $\beta_{\text {opt }}$, (or $B\left(\hat{\bar{y}}_{\pi} \mid \hat{\overline{\mathbf{x}}}_{\pi}\right)$ ) two cases can be distinguished. In the first one, the values taken by the auxiliary variable on all the units of the population are known. In this case, $\sum$ is thus known and

$$
\begin{aligned}
\operatorname{Cov}\left(\hat{\mathbf{x}}_{\pi}, \hat{\bar{y}}_{\pi}\right)= & \frac{1}{N^{2}} \sum_{k \in U} \frac{\mathbf{x}_{k} y_{k}}{\pi_{k}}\left(1-\pi_{k}\right)+ \\
& \frac{1}{N^{2}} \sum_{k \in U} \sum_{\substack{l \in U \\
l * k}} \frac{\mathbf{x}_{k} y_{l}}{\pi_{k} \pi_{l}}\left(\pi_{k l}-\pi_{k} \pi_{l}\right) \\
= & \frac{1}{N} \sum_{k \in U}\left(\overline{\mathbf{x}}_{\mid k}-\overline{\mathbf{x}}\right) y_{k}
\end{aligned}
$$

can be unbiasedly estimated by

$$
\begin{equation*}
\widehat{\operatorname{Cov}}_{1}\left(\hat{\mathbf{x}}_{\pi}, \hat{\bar{y}}_{\pi}\right)=\frac{1}{N} \sum_{k \in S} \frac{\left(\overline{\mathbf{x}}_{\mid k}-\overline{\mathbf{x}}\right) y_{k}}{\pi_{k}} \tag{3.5}
\end{equation*}
$$

where

$$
\begin{equation*}
\overline{\mathbf{x}}_{\mid k}=E\left(\hat{\overline{\mathbf{x}}}_{\pi} \mid k \in S\right)=\frac{1}{N} \sum_{\substack{l \in U \\ l \pm \downarrow}} \frac{\mathbf{x}_{l} \pi_{k l}}{\pi_{k} \pi_{l}}+\frac{\mathbf{x}_{k}}{\pi_{k} N} . \tag{3.6}
\end{equation*}
$$

By using (3.5), a first asymptotically optimal estimator can be constructed

$$
\begin{equation*}
\hat{\bar{y}}_{\mathrm{AOPT} 1}=\hat{\bar{y}}_{\pi}+\left(\overline{\mathbf{x}}-\hat{\mathbf{x}}_{\pi}\right)^{\prime} \Sigma^{-1} \frac{1}{N} \sum_{k \in S} \frac{\overline{\mathbf{x}}_{\mid k}-\overline{\mathbf{x}}}{\pi_{k}} y_{k} . \tag{3.7}
\end{equation*}
$$

In the second case, only the population mean $\overline{\mathbf{x}}$ is known, $\boldsymbol{\Sigma}$ must thus be estimated and $\operatorname{Cov}\left(\hat{\mathbf{x}}_{\pi}, \hat{\bar{y}}_{\pi}\right)$ can not be estimated using (3.5). Montanari proposes to estimate $\Sigma$ and $\operatorname{Cov}\left(\hat{\mathbf{x}}_{\pi}, \hat{\bar{y}}_{\pi}\right)$ by the classic Horvitz-Thompson estimator:

$$
\begin{aligned}
& \hat{\boldsymbol{\Sigma}}=\frac{1}{N^{2}} \sum_{k \in S} \frac{\mathbf{x}_{k} \mathbf{x}_{k}^{\prime}}{\pi_{k}^{2}}\left(1-\pi_{k}\right)+ \\
& \frac{1}{N^{2}} \sum_{k \in S} \sum_{\substack{l \in S \\
l * k}} \frac{\mathbf{x}_{k} \mathbf{x}_{l}^{\prime}}{\pi_{k} \pi_{l}} \frac{\pi_{k l}-\pi_{k} \pi_{l}}{\pi_{k l}}
\end{aligned}
$$

and

$$
\begin{aligned}
& \widehat{\operatorname{Cov}_{2}}\left(\hat{\mathbf{x}}_{n}, \hat{\bar{y}}_{n}\right)=\frac{1}{N^{2}} \sum_{k \in S} \frac{\mathbf{x}_{k} y_{k}}{\pi_{k}^{2}}\left(1-\pi_{k}\right)+ \\
& \frac{1}{N^{2}} \sum_{k \in S} \sum_{l \in S} \frac{\mathbf{x}_{k} y_{l}}{\pi_{k} \pi_{l}} \frac{\pi_{k l}-\pi_{k} \pi_{l}}{\pi_{k l}} .
\end{aligned}
$$

$\beta_{\text {OPT }}$ is thus estimated by ( $\hat{\boldsymbol{\Sigma}}$ is supposed non singular)

$$
\hat{\boldsymbol{\beta}}_{\text {OPT }}=\hat{\boldsymbol{\Sigma}}^{-1} \operatorname{Cov}_{2}\left(\hat{\overline{\mathbf{x}}}_{\pi}, \hat{\bar{y}}_{\pi}\right) .
$$

By estimating $\beta_{\text {opt }}$, another asymptotically optimal (AOPT) estimator can be given:

$$
\begin{equation*}
\hat{\bar{y}}_{\mathrm{AOPT} 2}=\hat{\bar{y}}_{\pi}+\left(\hat{\overline{\mathbf{x}}}_{\pi}-\overline{\mathbf{x}}\right)^{\prime} \hat{\mathrm{B}}_{\mathrm{OPT}} . \tag{3.8}
\end{equation*}
$$

The difference between the AOPT1 and AOPT2 estimator is the way we estimate $\operatorname{Cov}\left(\hat{\mathbf{x}}_{\pi}, \hat{\hat{y}}_{\pi}\right)$ and $\boldsymbol{\Sigma}$. However, the AOPT1-estimator needs more complete auxiliary information.

The generalised regression (GREG) estimator defined by Cassel, Särndal and Wretman (1976), Wright (1983), Särndal, Swensson and Wretman (1992, p. 225) is also an estimator of the linear class given by expression (3.4). For the GREG-estimator $\beta$ is defined by

$$
\beta_{\text {GREG }}=\left[\sum_{k \in U} \frac{\mathbf{x}_{k} \mathbf{x}_{k}^{\prime}}{c_{k}}\right]^{-1} \sum_{k \in U} \frac{\mathbf{x}_{k} y_{k}}{c_{k}}
$$

and can be estimated by

$$
\hat{\beta}_{\mathrm{GREG}}=\left[\sum_{k \in S} \frac{\mathbf{x}_{k} \mathbf{x}_{k}^{\prime}}{\pi_{k} c_{k}}\right]^{-1} \sum_{k \in S} \frac{\mathbf{x}_{k} y_{k}}{\pi_{k} c_{k}},
$$

where quantities $c_{k}>0, k \in U$, are weights defined for all the population units. The GREG-estimator does not have good conditional properties. It is generally conditionally biased (Rao 1994).

## 4. APPROXIMATION OF THE SCW-ESTIMATOR

Another way to construct a conditionally unbiased estimator is to find an approximation of the SCW-estimator given in (2.1). Indeed this estimator has good unbiasedness properties because it is VCU. If $\hat{\mathbf{x}}_{\boldsymbol{\pi}}$ is used as an auxiliary statistic, we shall seek an approximation of

$$
\pi_{k \mid T}=E\left(I_{k} \mid \hat{\mathbf{x}_{\pi}}\right) .
$$

If the random vector $\hat{\mathbf{x}}_{\boldsymbol{\pi}}$ takes for instance the value $z$, we get by Bayes's theorem that

$$
E\left(I_{k} \mid \hat{\mathbf{x}}_{\pi}=\mathbf{z}\right)=\operatorname{Pr}\left(k \in S \mid \hat{\mathbf{x}}_{\pi}=\mathbf{z}\right)=\pi_{k} \frac{\operatorname{Pr}\left(\hat{\mathbf{x}}_{\pi}=\mathbf{z} \mid k \in S\right)}{\operatorname{Pr}\left(\hat{\mathbf{x}}_{\pi}=\mathbf{z}\right)}
$$

In order to compute the conditional inclusion probabilities, it is thus necessary to know the probability distribution of $\hat{\mathbf{x}}_{\pi}$ unconditionally and conditionally on the presence of each unit in the sample. Except for some particular case, this probability distribution is very complex; for this reason an approximation will be constructed.

It is possible to derive the means and variances of $\hat{\mathbf{x}}_{\pi}$ unconditionally and conditionally on the presence of each unit in the sample. Indeed $E\left(\hat{\overline{\mathbf{x}}}_{\boldsymbol{\pi}}\right)$ is given in (3.2), $\operatorname{Var}\left(\hat{\mathbf{x}}_{\pi}\right)$ in (3.3), $E\left(\hat{\overline{\mathbf{x}}}_{\pi} \mid k \in S\right.$ ) in (3.6), and

$$
\begin{align*}
\boldsymbol{\Sigma}_{k}= & \operatorname{Var}\left(\hat{\mathbf{x}}_{n} \mid k \in S\right) \\
= & \frac{1}{N^{2}} \sum_{\substack{l \in U \\
l \neq k}} \frac{\mathbf{x}_{l} \mathbf{x}_{l}^{\prime} \pi_{k l}}{\pi_{k} \pi_{l}^{2}}\left(1-\frac{\pi_{k l}}{\pi_{k}}\right) \\
& +\frac{1}{N^{2}} \sum_{\substack{l \in U \\
l \neq k}} \sum_{\substack{m \in U \\
m \neq l \\
m \neq k}} \frac{\mathbf{x}_{l} \mathbf{x}_{m}^{\prime}}{\pi_{k} \pi_{l} \pi_{m}}\left(\pi_{k l m}-\frac{\pi_{k l} \pi_{k m}}{\pi_{k}}\right) \tag{4.9}
\end{align*}
$$

where $\pi_{k l m}$ is the third-order inclusion probability. Matrixes $\boldsymbol{\Sigma}$ and $\boldsymbol{\Sigma}_{k}$ are assumed to be non singular ${ }_{\dot{\boldsymbol{\Sigma}}}$

As the probability distribution of $\dot{\hat{\mathbf{x}}}_{\pi}$ is generally unknown, the following three assumptions will be used to construct an approximation of conditional inclusion probabilities.
(i) If the sample size $n$ is large, $\hat{\mathbf{x}}_{\boldsymbol{\pi}}$ has a multivariate normal distribution unconditionally and conditionally on the presence of each unit in the sample.
(ii) $\mathbf{R}_{k}^{-1}-\mathbf{R}^{-1}=O_{J \times J}\left(n^{-1}\right)$ for all $k \in U$ where $\mathbf{R}=\mathbf{V}^{-1 / 2}$ $\boldsymbol{\Sigma}^{-1 / 2}, \mathbf{R}_{k}=\mathbf{V}^{-1 / 2} \boldsymbol{\Sigma}_{k} \mathbf{V}^{-1 / 2}, \mathbf{V}$ denotes a $J \times J$ diagonal matrix having the elements of the diagonal of $\Sigma$ on its diagonal and $O_{J \times \prime}\left(n^{-a}\right)$ denotes a matrix of quantities that when multiplied by $n^{\alpha}$ remains bounded as $n \rightarrow \infty$.
(iii) $\gamma_{k}=\mathbf{V}^{-1 / 2}\left(\overline{\mathbf{x}}_{\mid k}-\overline{\mathbf{x}}\right)=O_{J}\left(n^{-1 / 2}\right)$ where $O_{J}\left(n^{-a}\right)$ denotes a vector of quantities that when multiplied by $n^{\alpha}$ remains bounded as $n \rightarrow \infty$.

These three hypotheses are made on the sample size. It is thus supposed that when $n$ increases, $N$ increases at least as quickly as $n$. Nevertheless, no hypothesis are made on $f=n / N$. Assuming that the hypotheses given in section 3 are verified, the following result gives an approximation of the SCW-estimator:
Result 1: Assuming (i), (ii) and (iii), and if the auxiliary statistic used is $\hat{\mathbf{x}}_{\pi}$, then

$$
\begin{align*}
\hat{\bar{y}}_{\mid T}=\hat{\bar{y}}_{\pi}+\left(\overline{\mathbf{x}}-\hat{\mathbf{x}}_{\pi}\right)^{\prime} \Sigma^{-1} & \frac{1}{N} \sum_{k \in S} \frac{\overline{\mathbf{x}}_{l k}-\overline{\mathbf{x}}}{\pi_{k}} y_{k} \\
& +O_{p}\left(n^{-1}\right) \approx \hat{\bar{y}}_{\text {AOPTI }} \tag{4.10}
\end{align*}
$$

where $n \times O_{p}\left(n^{-1}\right)$ is a quantity bounded in probability. Proof of Result 1 is given in the appendix.

## 5. DISCUSSION ABOUT THE HYPOTHESES

These three hypotheses are verified for simple random sampling without replacement when only one auxiliary variable is available. Indeed, in this case, we have $J=1$, $\mathbf{x}_{k}=x_{k}, \overline{\mathbf{x}}=\bar{x}, \overline{\mathbf{x}}_{\mid k}=\bar{x}_{\mid k} \hat{\mathbf{x}}_{\pi}=\hat{\bar{x}}_{\pi}$. We get
$\pi_{k}=\frac{n}{N}, \pi_{k l}=\frac{n}{N} \frac{n-1}{N-1}$ and $\pi_{k l m}=\frac{n}{N} \frac{n-1}{N-1} \frac{n-2}{N-2}$.
By (3.6), (3.3), (4.9),

$$
\begin{equation*}
\bar{x}_{\mid k}=\bar{x}+\frac{N-n}{N-1} \frac{x_{k}-\bar{x}}{n} \tag{5.11}
\end{equation*}
$$

$$
\begin{equation*}
\operatorname{Var}\left(\hat{\bar{x}}_{\pi}\right)=\frac{N-n}{N-1} \frac{\sigma_{x}^{2}}{n}, \tag{5.12}
\end{equation*}
$$

$\operatorname{Var}\left(\hat{\bar{x}}_{\pi} \mid k \in S\right)=\frac{N(N-n)(n-1)}{(N-2)(N-1) n^{2}}\left\{\sigma_{x}^{2}-\frac{\left(x_{k}-\bar{x}\right)^{2}}{N-1}\right\}$
where

$$
\sigma_{x}^{2}=\frac{1}{N} \sum_{k \in U}\left(x_{k}-\bar{x}\right)^{2}
$$

Now, consider the three hypotheses for this particular case.

- Hypothesis (i) was proved by Madow (1948) under some conditions.
- Hypothesis (ii) becomes

$$
\frac{\operatorname{Var}\left(\hat{\bar{x}}_{\pi} \mid k \in S\right)}{\operatorname{Var}\left(\hat{\bar{x}}_{\pi}\right)}-1=O\left(n^{-1}\right)
$$

By (5.12) and (5.13), we get

$$
\begin{gathered}
\frac{\operatorname{Var}\left(\hat{\bar{x}}_{\pi} \mid k \in S\right)}{\operatorname{Var}\left(\hat{\bar{x}}_{\pi}\right)}=\frac{N(n-1)}{(N-2) n}\left\{1-\frac{\left(x_{k}-\bar{x}\right)^{2}}{(N-1) \sigma_{x}^{2}}\right\} \\
=1-\frac{1}{n}\left\{\frac{N-2 n}{(N-2)}+\frac{N(n-1)}{(N-2)} \frac{\left(x_{k}-\bar{x}\right)^{2}}{(N-1) \sigma_{x}^{2}}\right\} \\
=1+O\left(\frac{1}{n}\right) .
\end{gathered}
$$

- Hypothesis (iii) becomes

$$
\frac{\bar{x}_{\mid k}-\bar{x}}{\sqrt{\operatorname{Var}\left(\hat{\bar{x}}_{\pi}\right)}}=O\left(n^{1 / 2}\right)
$$

By (5.11) (5.12), we get

$$
\gamma_{k}=\frac{\bar{x}_{\mid k}-\bar{x}}{\sqrt{\operatorname{Var}\left(\hat{\bar{x}}_{\pi}\right)}}=\frac{\sqrt{N-n}}{N-1} \frac{x_{k}-\bar{x}}{\sigma_{x} \sqrt{n}}=O\left(\frac{1}{\sqrt{n}}\right) .
$$

In simple random sampling, these hypotheses can better be interpreted. Hypothesis (i) is the classic assumption of normality that was also needed for the construction of the optimal estimator. In simple random sampling, it is easy to verify that Hypothesis (iii) implies hypothesis (ii). Both technical hypotheses simply imply that a particular unit cannot take a $\left|x_{k}-\bar{x}\right|$ value much more important than the other ones.

The three hypotheses are thus valid under simple random sampling when only one variable is available. This result can also be extended to stratified sampling when the number of strata is fixed and the sample size within each stratum is large. In cluster sampling, when the number of clusters is large and the clusters are selected with a simple random sampling design, these hypotheses are still applicable. Hypothesis (i) was also partially showed by Rosén (1972) for sampling with unequal probabilities. Actually, the proof of Rosén is restricted to a rejective sampling design.

The proposed hypotheses, are generally less restrictive than a superpopulation model. Indeed, a superpopulation model is a set of hypotheses on the interest variables while the three hypotheses presented only affect the auxiliary variables. In a superpopulation model, the relation between the interest variable and the auxiliary variables are the most extensive contribution of the model. In the conditional approach, no hypothesis is made on the interest variable. If the hypotheses presented are debatable, it is thus clear that a superpopulation model is a set of hypotheses much more restrictive than those used in the conditional approach.

## 6. APPLICATION TO STRATIFIED SAMPLING

### 6.1 The Problem

In stratification, auxiliary information is used it a priori to improve the estimation. In this case, three sets of variables interact: the stratification variables, the auxiliary variables used a posteriori and the interest variable. Suppose that the population is partitioned into $H$ strata $U_{h}, h=1, \ldots, H$, of size $N_{h}, h=1, \ldots, H$. The population means of the strata are denoted $\bar{y}_{h}=N_{h}^{-1} \sum_{k \in U_{h}} y_{k}$ and $\overline{\mathbf{x}}_{h}=N_{h}^{-1} \sum_{k \in U_{h}} \mathbf{x}_{k}$. A simple random sample $S_{h}$ of fixed size $n_{h}\left(\sum_{h=1}^{H} n_{h}=n\right)$ is selected without replacement independently in each stratum. From the general theory of stratification (see for instance Särndal, Swensson and Wretman 1992, p. 100), we get

$$
\hat{\bar{y}}_{\pi}=\frac{1}{N} \sum_{h=1}^{H} N_{h} \hat{\bar{y}}_{h} \text { and } \hat{\overline{\mathbf{x}}}_{\pi}=\frac{1}{N} \sum_{h=1}^{H} N_{h} \hat{\mathbf{x}}_{h}
$$

where

$$
\hat{\bar{y}}_{h}=\frac{1}{n_{h}} \sum_{k \in S_{h}} y_{k} \text { and } \hat{\mathbf{x}}_{h}=\frac{1}{n_{h}} \sum_{k \in \mathcal{S}_{h}} \mathbf{x}_{k} \text {. }
$$

Moreover, we have that

$$
\begin{aligned}
& \operatorname{Cov}\left(\hat{\mathbf{x}}_{\pi}, \hat{\bar{y}}_{\pi}\right)= \\
& \quad \frac{1}{N^{2}} \sum_{h=1}^{H} N_{h}^{2} \frac{1-f_{h}}{n_{h}} \frac{1}{N_{h}-1} \sum_{k \in U_{h}}\left(\mathbf{x}_{k}-\overline{\mathbf{x}}_{h}\right)\left(y_{k}-\bar{y}_{h}\right)^{\prime}
\end{aligned}
$$

and

$$
\begin{aligned}
\Sigma= & \frac{1}{N^{2}} \sum_{h=1}^{H} N_{h}^{2} \\
& \frac{1-f_{h}}{n_{h}} \frac{1}{N_{h}-1} \sum_{k \in U_{h}}\left(\mathbf{x}_{k}-\overline{\mathbf{x}}_{h}\right)\left(\mathbf{x}_{k}-\overline{\mathbf{x}}_{h}\right)^{\prime}
\end{aligned}
$$

where $f_{h}=n_{h} / N_{h}, h=1, \ldots, H$.

### 6.2 AOPT1-estimator

If $k \in U_{\alpha}$, by extending expression (3.6) to stratified sampling, we get

$$
\overline{\mathbf{x}}_{\mid k}=E\left(\hat{\mathbf{x}}_{\pi} \mid k \in S\right)=\overline{\mathbf{x}}+\frac{N_{\alpha}^{2}}{N\left(N_{\alpha}-1\right)} \frac{1-f_{\alpha}}{n_{\alpha}}\left(\mathbf{x}_{k}-\overline{\mathbf{x}}_{\alpha}\right)
$$

and

$$
\begin{aligned}
& \frac{1}{N} \sum_{k \in S} \frac{\left(\overline{\mathbf{x}}_{1 k}-\overline{\mathbf{x}}\right) y_{k}}{\pi_{k}}= \\
& \quad \frac{1}{N^{2}} \sum_{h=1}^{H} \frac{N_{h}^{3}}{N_{h}-1} \frac{1-f_{h}}{n_{h}} \frac{1}{n_{h}} \sum_{k \in S_{h}}\left(\mathbf{x}_{k}-\overline{\mathbf{x}}_{h}\right) y_{k} .
\end{aligned}
$$

From (3.7) the AOPT1-estimator of can be derived as

$$
\begin{aligned}
& \hat{\bar{y}}_{\text {AOPT } 1}=\hat{\bar{y}}_{\pi}+\left(\overline{\mathbf{x}}-\hat{\mathbf{x}}_{\pi}\right)^{\prime} \\
& \left\{\sum_{h=1}^{H} N_{h}^{2} \frac{1-f_{h}}{n_{h}} \frac{1}{N_{h}-1} \sum_{k \in U_{h}}\left(\mathbf{x}_{k}-\overline{\mathbf{x}}_{h}\right)\left(\mathbf{x}_{k}-\overline{\mathbf{x}}_{h}\right)^{\prime}\right\}^{-1} \\
& \quad \times \sum_{h=1}^{H} \frac{N_{h}^{3}}{N_{h}-1} \frac{1-f_{h}}{n_{h}} \frac{1}{n_{h}} \sum_{k \in S_{h}}\left(\mathbf{x}_{k}-\overline{\mathbf{x}}_{h}\right) y_{k} .
\end{aligned}
$$

The use of this estimator requires the knowledge of very substantial auxiliary information. The population means $\overline{\mathbf{x}}_{h}$ of the auxiliary variables must be known for each stratum as well as the stratum sizes $N_{h}$. Moreover, the values taken by the auxiliary variables must be known for each unit of $U$.

However $\hat{\bar{y}}_{\text {AOPT1 }}$ has in stratification an important drawback: it is not calibrated on the strata size $N_{h}$ i.e., when the objective consists in estimating the strata sizes $N_{h}$, generally $\hat{N}_{\text {AOPT1 }} \neq N_{h}$. This drawback can easily be overcome by centring the interest variable. We thus get:

$$
\begin{aligned}
& \hat{\bar{y}}_{\text {AOPTIC }}=\hat{\bar{y}}_{\pi}+\left(\overline{\mathbf{x}}-\hat{\mathbf{x}}_{\pi}\right)^{\prime} \\
& \left\{\sum_{h=1}^{H} N_{h}^{2} \frac{1-f_{h}}{n_{h}} \frac{1}{N_{h}-1} \sum_{k \in U_{h}}\left(\mathbf{x}_{k}-\overline{\mathbf{x}}_{h}\right)\left(\mathbf{x}_{k}-\overline{\mathbf{x}}_{h}\right)^{\prime}\right\}^{-1} \\
& \times \sum_{h=1}^{H} \frac{N_{h}^{3}}{N_{h}-1} \frac{1-f_{h}}{n_{h}} \frac{1}{n_{h}} \sum_{k \in S_{h}}\left(\mathbf{x}_{k}-\overline{\mathbf{x}}_{h}\right)\left(y_{k}-\hat{\bar{y}}_{h}\right) .
\end{aligned}
$$

### 6.3 AOPT2-Estimator

The AOPT2-estimator can also be used in stratification. In this case, from (3.8) we get

$$
\begin{aligned}
& \hat{\bar{y}}_{\mathrm{AOPT} 2}=\hat{\bar{y}}_{\pi}+\left(\overline{\mathbf{x}}-\hat{\mathbf{x}}_{\pi}\right)^{\prime} \\
& \left\{\sum_{h=1}^{H} N_{h}^{2} \frac{1-f_{h}}{n_{h}} \frac{1}{n_{h}-1} \sum_{k \in S_{h}}\left(\mathbf{x}_{k}-\hat{\mathbf{x}}_{h}\right)\left(\mathbf{x}_{k}-\hat{\mathbf{x}}_{h}\right)^{\prime}\right\}^{-1} \\
& \times \sum_{h=1}^{H} N_{h}^{2} \frac{1-f_{h}}{n_{h}} \frac{1}{n_{h}-1} \sum_{k \in S_{h}}\left(\mathbf{x}_{k}-\hat{\overline{\mathbf{x}}}_{h}\right)\left(y_{k}-\hat{\bar{y}}_{h}\right) .
\end{aligned}
$$

The AOPT2-estimator only needs the knowledge of the population mean vector $\overline{\mathbf{x}}$ and of the stratum sizes $N_{h}$. It has however a drawback, the $\overrightarrow{\mathbf{x}}_{h}$ are estimated and thus $J \times H$ degrees of freedom are lost. If the number of strata is large, this loss of degrees of freedom could increase the instability of this estimator when $J \times H$ is large.

### 6.4 GREG-Estimator

The GREG-estimator does not take into account the joint inclusion probabilities. It is given by

$$
\begin{aligned}
\hat{\bar{y}}_{\text {GREG }}=\hat{\bar{y}}_{\pi}+\left(\overline{\mathbf{x}}-\hat{\bar{x}}_{\pi}\right)^{\prime}\{ & \left\{\sum_{h=1}^{H} \frac{N_{h}}{n_{h}} \sum_{k \in S_{h}} \frac{\mathbf{x}_{k} \mathbf{x}_{k}{ }^{\prime}}{c_{k}}\right\}^{-1} \\
& \sum_{h=1}^{H} \frac{N_{h}}{n_{h}} \sum_{k \in S_{h}} \frac{\mathbf{x}_{k} y_{k}}{c_{k}}
\end{aligned}
$$

Although this estimator is more stable, it is conditionally biased. Moreover, if we want to estimate the stratum sizes $N_{h}$ by the GREG-estimator, we do not find exactly $N_{h}$. Indeed, if $y_{k}=1$ when $k \in U_{h}$ and $y_{k}=0$ when $k \notin U_{h}$, then

$$
\begin{aligned}
& \hat{N}_{\text {GREG }}= N \hat{\bar{y}}_{\text {GREG }} \\
&= \hat{N}_{h \pi}+N\left(\overline{\mathbf{x}}-\hat{\bar{x}}_{\pi}\right)^{\prime} \\
&\left\{\sum_{h=1}^{H} \frac{N_{h}}{n_{h}} \sum_{k \in S_{h}} \frac{\mathbf{x}_{k} \mathbf{x}_{k}^{\prime}}{c_{k}}\right\}^{-1} \\
& \times \sum_{h=1}^{H} \frac{N_{h}}{n_{h}} \sum_{k \in S_{h}} \frac{\mathbf{x}_{k} y_{k}}{c_{k}}
\end{aligned}
$$

Since, in stratified sampling, $\hat{N}_{h \pi}=N_{h}$, we get

$$
\begin{align*}
& \hat{N}_{\text {GREG }}=N_{h}+N\left(\overline{\mathbf{x}}-\hat{\mathbf{x}}_{\pi}\right)^{\prime}\left\{\sum_{h=1}^{H} \frac{N_{h}}{n_{h}} \sum_{k \in S_{h}} \frac{\mathbf{x}_{k} \mathbf{x}_{k}^{\prime}}{c_{k}}\right\}^{-1} \\
& \times \sum_{h=1}^{H} \frac{N_{h}}{n_{h}} \sum_{k \in S_{h}} \frac{\mathbf{x}_{k}}{c_{k}} \tag{6.14}
\end{align*}
$$

Expression (6.14) shows that generally $\hat{N}_{\text {GREG }} \neq N_{h}$. Thus, the GREG-estimator destroys the stratification effect because it does not take the stratification into account. Indeed, the stratification is represented by the joint inclusion probabilities. In the GREG-estimator, only the first-order inclusion probabilities are used. On the other hand, it is easy to verify that the AOPT1 and AOPT2estimator of $N_{h}$ are exactly equal to $N_{h}$. The AOPTestimators is thus calibrated on the $N_{h}$.

We propose to use the GREG-estimator in the following cases: when the sample size is small or if the number of strata is large and when the stratification gives poor auxiliary information on the interest variable. Indeed, in this case, the loss of precision due to the loss of degrees of freedom, will be more important than the precision benefit due to the optimality of the estimator. An interesting analysis of the benefit due to the optimal estimator is also given in Montanari (1998).

### 6.5 GREG-Estimator With use of the Stratification Variables

A variant of use of the GREG-estimator consists in re-using the stratification variables at the estimation stage. Consider the column vector

$$
\mathbf{w}_{k}=\left(z_{k 1} \ldots, z_{k h}, \ldots, z_{k(H)} \mathbf{x}_{k}^{\prime}\right)^{\prime}
$$

where $z_{k h}=1$ if $k \in U_{h}$ and 0 if not. This vector is thus composed of the values taken by the indicator variables of the presence of unit $k$ in the $H$ strata and of the values taken by the $x$-auxiliary variables.

Now if $\overline{\mathbf{w}}$ denotes the population mean of vectors $\mathbf{w}_{k}$ and $\hat{\overline{\mathbf{w}}}_{\pi}$ its Horvitz-Thompson estimator, the GREGestimator using the auxiliary information $\bar{w}_{\pi}$ is given by

$$
\begin{align*}
& \hat{\bar{y}}_{\text {GREGw }}=\hat{\bar{y}}_{\pi}+\left(\overline{\mathbf{w}}-\hat{\mathbf{w}}_{\pi}\right)^{\prime} \\
& \left\{\sum_{h=1}^{H} \frac{N_{h}}{n_{h}} \sum_{k \in S_{h}} \frac{\mathbf{w}_{k} \mathbf{w}_{k}^{\prime}}{c_{k}}\right\}^{-1} \sum_{h=1}^{H} \frac{N_{h}}{n_{h}} \sum_{k \in S_{h}} \frac{\mathbf{w}_{k} y_{k}}{c_{k}} . \tag{6.15}
\end{align*}
$$

The presentation of expression (6.15) can be simplified. Indeed, the following result was proved by Tille (1994) and generalised by Särndal (1996):

## Result 2:

When the stratification variables are re-used at the estimation stage, and if the $c_{k}$ are equal into the strata ( $c_{k}=c_{h}, k \in U_{h}$ ) the GREG-estimator can be written

$$
\begin{align*}
& \hat{\bar{y}}_{\text {GREG } w}= \\
& \hat{\bar{y}}_{\pi}+\left(\overline{\mathbf{x}}-\hat{\mathbf{x}}_{\pi}\right)^{\prime}\left\{\sum_{h=1}^{H} \frac{N_{h}}{n_{h} c_{h}} \sum_{k \in S_{h}}\left(\mathbf{x}_{k}-\hat{\mathbf{x}}_{h}\right)\left(\mathbf{x}_{k}-\hat{\mathbf{x}}_{h}\right)^{\prime}\right\}^{-1} \\
& \times \sum_{h=1}^{H} \frac{N_{h}}{n_{h} c_{h}} \sum_{k \in S_{h}}\left(\mathbf{x}_{k}-\hat{\mathbf{x}}_{h}\right)\left(y_{k}-\hat{\bar{y}}_{h}\right) . \tag{6.16}
\end{align*}
$$

A proof of Result 2 is given in the Appendix. Note that expression (6.16) is equal to the AOPT2-estimator when

$$
c_{h}=C \frac{n_{h}}{N_{h}} \frac{1-1 / n_{h}}{1-f_{h}},
$$

for $h=1, \ldots, H$, where $C>0$ is a constant. When the $f_{h}$ are small and the $n_{h}$ are large and proportional to the $N_{h}$ both estimators are equivalent. This result shows that with the conditional approach, the fact that the sampling design is stratified is automatically taken into account in the estimation method. The GREG-estimator does not take into account the stratification effect and thus it is necessary to reintroduce the stratification variables at the estimation stage so as not to lose the stratification effect.

## 7. SIMULATIONS

A set of simulations was carried out in order to compare the four following estimators: $\hat{\hat{y}}_{\pi}, \hat{\hat{y}}_{\text {AOPTIC, }} \hat{\hat{y}}_{\text {AOPT2, }}$ $\hat{\hat{y}}_{\text {GREG }}$. The population is made up of 4 strata of 250 units ( $N=1,000$ ). A stratified sampling design is applied with proportional allocation. For each simulation, 10,000 samples of size $n=100$ are selected and the following ratios has been estimated:

$$
\begin{aligned}
& M_{1}=\operatorname{MSE}\left(\hat{\bar{y}}_{\pi}\right) / \operatorname{MSE}\left(\hat{\bar{y}}_{\pi}\right)=1, \\
& M_{2}=\operatorname{MSE}\left(\hat{\bar{y}}_{\text {GREG }}\right) / \operatorname{MSE}\left(\hat{\bar{y}}_{\pi}\right), \\
& M_{3}=\operatorname{MSE}\left(\hat{\bar{y}}_{\mathrm{AOPTIC}}\right) / \operatorname{MSE}\left(\hat{\hat{y}}_{\pi}\right), \\
& M_{4}=\operatorname{MSE}\left(\hat{\bar{y}}_{\mathrm{AOPT} 2}\right) \operatorname{MSE}\left(\hat{\bar{y}}_{\pi}\right) .
\end{aligned}
$$

The populations are generated by means of the following models: $\mathcal{P}_{1}: x_{k}=a_{k}, y_{k}=e_{k}, k \in U$, (total independence), $\mathcal{P}_{2}: x_{k}=a_{k}, y_{k}=3 x_{k}+e_{k}, k \in U$, (dependence between $x$ and y), $\mathcal{P}_{3}: x_{k}=a_{k}, y_{k}=x_{k}+2 h(k)+e_{k}, k \in U$, (dependence between $x, y$ and the strata), $\mathcal{P}_{4}: x_{k}=a_{k}, y_{k}=\exp \left(10+2 x_{k}\right.$ $\left(+10 h(k)+e_{k}\right), k \in U$, non-linearity and dependence between $x, y$ and the strata), $\mathcal{P}_{s}: x_{k}=a_{k}, y_{k}=\exp \left(e_{k}+3 x_{k}\right)+3 h(k)$, $k \in U$, (non-linearity and dependence between $x, y$ and the strata), $\mathcal{P}_{6}: x_{k}=a_{k}, y_{k}=3 h(k)+e_{k}, k \in U$, (strong dependence between $y$ and the strata), $\mathcal{P}_{7}: x_{k}=a_{k}, y_{k}=$ $50 h(k)+e_{k}, k \in U$, (very strong dependence between $y$ and the strata), where $a_{k}$ and $e_{k}$ are independent normal variable with mean equal to 0 and variance equal to 1 , and $h(k)$ is the number of the stratum of unit $k$. Results of the simulations is given in Table 1.

Table 1
Results of 10,000 Simulations

|  | $\mathcal{P}_{1}$ | $\mathcal{P}_{2}$ | $\mathcal{P}_{3}$ | $\mathcal{P}_{4}$ | $\mathcal{P}_{5}$ | $\mathcal{P}_{6}$ | $\mathcal{P}_{7}$ |
| :--- | :--- | :--- | :--- | :--- | :--- | :--- | :--- |
| $M_{1}$ | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 |
| $M_{2}$ | 1.0070 | 0.0906 | 0.5180 | 0.9261 | 0.9263 | 1.1047 | 38.5104 |
| $M_{3}$ | 1.0069 | 0.0906 | 0.4835 | 0.9277 | 0.9269 | 1.0015 | 1.0123 |
| $M_{4}$ | 1.0060 | 0.0936 | 0.4850 | 0.9257 | 0.9239 | 1.0006 | 1.0111 |

Table 1 shows that the GREG-estimator provides a good estimation when the stratification variables are not correlated to the interest variable. Nevertheless, the more is the dependence between the stratification variable and the interest variable, the more is the gain of precision of $\hat{\hat{y}}_{\text {AOPT1C }}$ and $\hat{\hat{y}}_{\text {AOPT2 }}$. The loss of degrees of freedom of the optimal estimator does not seem to affect the precision for this sample size. Moreover, the gain obtained by the knowledge of the population stratum is not significant for this sample size. For all these cases, the optimal estimator is thus clearly preferable to the GREG-estimator.

## 8. A THIRD - ORDER PROBLEM

The complexity of determining the conditional weights is not a specific problem of the SCW-estimator. It is due to the general problem of estimation with auxiliary information used a posteriori when an auxiliary variable is already used a priori in the sampling design. This problem can be presented as a third-order interaction problem among

- the interest variables;
- the sampling design and thus the auxiliary variables used a priori;
- the auxiliary variables used a posteriori.

Indeed, the use of auxiliary information at the estimation stage leads to the following problem: how do these auxiliary variables used a posteriori interact with the interest variable through a given sampling design? The problem being complex, we have to take into account the relationships between each set of variables above as well as the third-order interactions among these three sets of variables.

It is very difficult to find a really operational estimator which uses the three second-order interactions and the third-order interaction. For this reason, one can attempt to simplify the problem. The neutralisation of one of the aspects of this problem significantly simplifies the research of an estimator. Most of the possible simplifications have already been studied. We can cite some of these:

- If no auxiliary information is used a posteriori (except the population size $N$ ) we can only construct the Horvitz-Thompson estimator or Hájek's ratio (1971).
- Searching general solutions using auxiliary information for simple random sampling does not pose major problems. In this case, no auxiliary information is used a priori.
- Using a superpopulation model allows one to fix a relation existing between the interest variable and the auxiliary variables used a posteriori. In this case, it is possible to determine the optimal estimator (under the model).
- For the GREG-estimator and also for the calibration methods see Deville and Särndal 1992), in the designbased inference framework, only the first-order probabilities are retained from the sampling design. A simple random sampling is thus treated in the same way as a stratified design for which the first-order inclusion probabilities are all equal. For this reason, a regression estimator applied to a stratified design generally destroys the calibration on the stratum frequencies given by the a priori stratification. In this case, the simplification arises because all the contributions of the auxiliary variables used a priori to the sampling design can be described only by the first-order probabilities.
- Finally, for the optimal linear estimator, it is implicitly supposed that the dependence between HorvitzThompson estimators of the variables $\mathbf{x}$ and $y$ is linear. Obviously, these estimators neglect the non-linear dependence between the estimators. Nevertheless, it takes into account the joint inclusion probabilities. When the sampling design is stratified, the estimator remains calibrated on the population stratum frequencies.

The CW-estimator takes into account this third-order interaction. Moreover, in this case, auxiliary information does not necessarily intervene in a linear way. The weights depend on both the sampling design and the auxiliary statistic. These weights applied to the values taken by the
interest variable take into account all the interactions between the three variable groups.

The methods using conditional inclusion probabilities are interesting for different reasons: they give a general frameallowing to search and conceive estimators using auxiliary information without reference to a superpopulation model and lead to valid conditional inference. They bring into prominence all the complexity of the estimation problem with auxiliary nformation. According to the known auxiliary information, we can find either known results (as for example post-stratification) or very complex and not really operational estimators. However, a first approximation leads to a known result, i.e. the optimal linear estimator.

## ACKNOWLEDGEMENTS

The author is grateful to two anonymous referees and an Associate Editor for constructive suggestions and to Professor Carl Särndal for interesting comments on a previous version of this paper.

## APPENDIX: PROOF OF RESULT 1 AND 2

Lemma 1 will be used in the proof of Result 1.
Lemma 1 If $\mathbf{R}_{k}^{-1}-\mathbf{R}^{-1}=O_{f \times J}\left(n^{-1}\right)$, then $\left|\mathbf{R}_{k}\right|^{-1}|\mathbf{R}|=$ $1+O\left(n^{-1}\right)$, where $\mathbf{R}$ and $\mathbf{R}_{k}$ are defined as in hypothesis (ii).

Proof

$$
\left[\mathbf{R}_{k}^{-1}-\mathbf{R}^{-1}\right] \mathbf{R}=O_{J_{\times J}}\left(n^{-1}\right) \mathbf{R}
$$

and thus

$$
\begin{aligned}
\left|\mathbf{R}_{k}^{-1} \mathbf{R}\right|= & \left|I+O_{J \times J}\left(n^{-1}\right) \mathbf{R}\right|= \\
& {\left[1+O\left(n^{-1}\right)\right]^{J}+O\left(n^{-1}\right)=1+O\left(n^{-1}\right) }
\end{aligned}
$$

where $\mathbf{I}$ is a $J \times J$ identity matrix. Thus,

$$
\frac{\left|\mathbf{R}_{k}\right|}{|\mathbf{R}|}=\frac{1}{1+O\left(n^{-1}\right)}=1+O\left(n^{-1}\right) .
$$

Note that lemma 1 is a consequence of hypothesis (ii) .

## Proof of Result 1

If we define

$$
d_{k}=\frac{n}{N \pi_{k}}, \text { for all } k \in U .
$$

by hypothesis (i), we get:

$$
a_{k}\left(\hat{\mathbf{x}_{\pi}}\right)=\frac{n}{\pi_{k \mid T} N}=\frac{n \operatorname{Pr}\left(\hat{\mathbf{x}}_{\pi}\right)}{N \pi_{k} \operatorname{Pr}\left(\hat{\mathbf{x}}_{\pi} \mid k \in S\right)}=d_{k} \frac{f\left(\hat{\mathbf{x}}_{\pi}\right)}{f_{k}\left(\hat{\mathbf{x}}_{\pi}\right)}
$$

where $f$ (resp. $f_{k}$ ) is the density function of a multivariate normal variable with mean $\overline{\mathbf{x}}$ (resp. $\overline{\mathbf{x}}_{\mid k}$ ) and variancecovariance matrix $\boldsymbol{\Sigma}$ (resp. $\boldsymbol{\Sigma}_{k}$ ). Thus,

$$
\begin{equation*}
a_{k}\left(\hat{\overline{\mathbf{x}}}_{\pi}\right)=d_{k} \frac{\frac{1}{|\Sigma|^{1 / 2}} \exp -\frac{1}{2}\left(\hat{\mathbf{x}}_{\pi}-\overline{\mathbf{x}}\right)^{\prime} \Sigma^{-1}\left(\hat{\mathbf{x}}_{\pi}-\overline{\mathbf{x}}\right)}{\frac{1}{\left|\Sigma_{k}\right|^{1 / 2}} \exp -\frac{1}{2}\left(\hat{\mathbf{x}}_{\pi}-\overline{\mathbf{x}}_{\mid k}\right)^{\prime} \Sigma_{k}^{-1}\left(\hat{\mathbf{x}}_{\pi}-\overline{\mathbf{x}}_{\mid k}\right)} \tag{8.17}
\end{equation*}
$$

If we also note

$$
\begin{aligned}
\mathbf{R} & =\mathbf{V}^{-1 / 2} \boldsymbol{\Sigma} \mathbf{V}^{-1 / 2}, \\
\mathbf{R}_{k} & =\mathbf{V}^{-1 / 2} \boldsymbol{\Sigma}_{k} \mathbf{V}^{-1 / 2}, \\
\hat{\mathbf{x}}_{\boldsymbol{\pi}}^{c} & =\mathbf{V}^{-1 / 2}\left(\hat{\mathbf{x}}_{\pi}-\overline{\mathbf{x}}\right), \\
\gamma_{k} & =\mathbf{V}^{-1 / 2}\left(\overline{\mathbf{x}}_{\mid k}-\overline{\mathbf{x}}\right),
\end{aligned}
$$

and

$$
\begin{equation*}
c_{k}=d_{k} \frac{\left|\mathbf{R}_{k}\right|^{1 / 2}}{|\mathbf{R}|^{1 / 2}} \exp \frac{1}{2} \hat{\mathbf{x}}_{\pi}^{c^{\prime}}\left(\mathbf{R}_{k}^{-1}-\mathbf{R}^{-1}\right) \hat{\mathbf{x}}_{\pi}^{c}, \tag{8.18}
\end{equation*}
$$

we get

$$
\begin{align*}
a_{k}\left(\hat{\mathbf{x}}_{\pi}\right) & =d_{k} \frac{|\mathbf{R}|^{-1 / 2} \exp -\frac{1}{2} \hat{\mathbf{x}}_{\pi}^{c^{\prime}} \mathbf{R}^{-1} \hat{\mathbf{x}}_{\pi}^{c}}{\left|\mathbf{R}_{k}\right|^{-1 / 2} \exp -\frac{1}{2}\left(\hat{\overline{\mathbf{x}}}_{\pi}^{c}-\gamma_{k}\right)^{\prime} \mathbf{R}_{k}^{-1}\left(\hat{\overline{\mathbf{x}}}_{\pi}^{c}-\gamma_{k}\right)} \\
& =c_{k} \exp \frac{1}{2} \gamma_{k} \mathbf{R}_{k}^{-1}\left(\gamma_{k}-2 \hat{\mathbf{x}}_{\pi}^{c}\right) \tag{8.19}
\end{align*}
$$

By using a Taylor development for the vector $\gamma_{k}$ of (8.19), we get

$$
\begin{equation*}
a_{k}\left(\hat{\overline{\mathbf{x}}}_{\pi}\right)=c_{k}\left(1-\gamma_{k} \mathbf{R}_{k}^{-1} \hat{\overline{\mathbf{x}}}_{\pi}^{c}\right)+R\left(\gamma_{k}^{(0)}\right) \tag{8.20}
\end{equation*}
$$

where

$$
\begin{aligned}
R\left(\gamma_{k}^{(0)}\right) & =c_{k}\left[\exp \frac{1}{2} \gamma_{k}^{(0)} \mathbf{R}_{k}^{-1}\left(\gamma_{k}^{(0)}-2 \hat{\mathbf{x}}_{\pi}^{c}\right)\right] \\
& \left.\times \gamma_{k}^{\prime}\left\{\left[\mathbf{R}_{k}^{-1}\left(\gamma_{k}^{(0)}-\hat{\mathbf{x}}_{\pi}^{c}\right)\right]\left[\mathbf{R}_{k}^{-1}\left(\gamma_{k}^{(0)}-\hat{\mathbf{x}}_{\pi}^{c}\right)\right]\right]^{\prime}-\mathbf{R}_{k}^{-1}\right\} \gamma_{k}
\end{aligned}
$$

and $\gamma_{k}^{(0)}$ is a vector whose elements are included between the correspondent elements of $\gamma_{k}$ and 0 . By hypothesis (iii), we directly get

$$
R\left(\gamma_{k}^{(0)}\right)=O_{p}\left(n^{-1}\right)
$$

On the other hand, we have by hypothesis (ii), lemma 1 and (8.18) that

$$
\begin{align*}
c_{k} & =d_{k}\left[1+O\left(n^{-1}\right)\right]^{1 / 2} \exp \left\{\frac{1}{2} \hat{\mathbf{x}}_{\pi}^{c^{\prime}} O_{J \times J}\left(n^{-1}\right) \hat{\mathbf{x}}_{\pi}^{c}\right\} \\
& =d_{k}\left[1+O_{p}\left(n^{-1}\right)\right] . \tag{8.21}
\end{align*}
$$

By (8.20) and (8.21), we get

$$
\begin{aligned}
a\left(\hat{\mathbf{x}}_{\pi}\right)= & d_{k}\left[1+O_{p}^{-1}\left(n^{-1}\right)\right] \\
& \left\{1-\gamma_{k}\left(\mathbf{R}^{-1}+O_{j \times j}\left(n^{-1}\right)\right) \hat{\mathbf{x}}_{\pi}^{c}+O_{p}\left(n^{-1}\right)\right\} \\
= & d_{k}\left\{1-\gamma_{k}^{\prime} \mathbf{R}^{-1} \hat{\mathbf{x}}_{\pi}^{c}+O_{p}\left(n^{-1}\right)\right\} \\
= & d_{k}\left\{1+\left(\overline{\mathbf{x}}-\hat{\mathbf{x}}_{\pi}\right)^{\prime} \sum^{-1}\left(\overline{\mathbf{x}}_{\mid k}-\overline{\mathbf{x}}\right)+O_{p}\left(n_{-1}\right)\right\} .
\end{aligned}
$$

Finally, we get

$$
\begin{aligned}
\hat{\bar{y}}_{\mid T} & =\frac{1}{n} \sum_{k \in S} a_{k}\left(\hat{\overline{\mathbf{x}}}_{\pi}\right) y_{k} \\
& =\hat{y}_{\pi}+\left(\overline{\mathbf{x}}-\hat{\mathbf{x}}_{\pi}\right)^{\prime} \sum^{-1} \frac{1}{N} \sum_{k \in S} \frac{\left(\overline{\mathbf{x}}_{\mid k}-\overline{\mathbf{x}}\right) y_{k}}{\pi_{k}}+O_{p}\left(n^{-1}\right)
\end{aligned}
$$

## Proof of Result 2

In Särndal (1980), we see that the GREG-estimator presented in (6.15) can also be written:

$$
\begin{aligned}
& \hat{\bar{y}}_{\text {GREG } w}=\hat{\bar{y}}_{\boldsymbol{\pi}}+N^{-1}\left(\mathbf{1}_{N}^{\prime} \mathbf{W}_{N}-\mathbf{1}_{S}^{\prime} \boldsymbol{\Pi}_{S}^{-1} \mathbf{W}_{S}\right)^{\prime} \\
&\left(\mathbf{W}_{s}^{\prime} \mathbf{C}_{S}^{-1} \mathbf{\Pi}_{s}^{-1} \mathbf{W}_{S}\right)^{-1} \mathbf{W}_{s}^{\prime} \mathbf{C}_{S}^{-1} \mathbf{\Pi}_{s}^{-1} \mathbf{y}_{S}
\end{aligned}
$$

where $\mathbf{1}_{N}\left(\right.$ resp. $\left.\mathbf{1}_{S}\right)$ is a column vector composed of $N$ (resp. $n$ ) ones, $\Pi_{N}$ (resp. $\Pi_{s}$ ) is a diagonal matrix having the inclusion probabilities of the population (resp. sample) units on its diagonal, $\mathbf{C}_{s}$ is a diagonal matrix having the $c_{k}$ of the sample units on its diagonal, $\mathbf{y}_{S}$ is a column vector composed of the values taken by the interest variable $y$ in the sample,

$$
\mathbf{W}_{N}=\left(\begin{array}{cccc}
z_{11} & \cdots & z_{1 H} & \mathbf{x}_{1}^{\prime} \\
\vdots & & \vdots & \vdots \\
z_{N 1} & \cdots & z_{N H} & \mathbf{x}_{N}^{\prime}
\end{array}\right)
$$

and $\mathbf{W}_{S}$ is a $n \times(H+J)$ matrix composed of the $n$ rows of $\mathbf{W}_{N}$ corresponding to the units selected in the sample.

The matrix to invert can be partitioned into four parts:

$$
\left(\mathbf{W}_{s}^{\prime} \mathbf{C}_{s}^{-1} \boldsymbol{\Pi}_{s}^{-1} \mathbf{W}_{s}\right)=\left[\begin{array}{cc}
\mathbf{A} & \mathbf{D} \\
\mathbf{D}^{\prime} & \mathbf{B}
\end{array}\right]
$$

where $\mathbf{A}$ is an $H \times H$ matrix having $N_{h} / c_{h}, h=1, \ldots, H$, on its diagonal,

$$
\mathbf{B}=\sum_{h=1}^{H} \frac{N_{h}}{n_{h} c_{h}} \sum_{k \in S_{h}} \mathbf{x}_{k} \mathbf{x}_{k}^{\prime}
$$

and

$$
\mathbf{D}^{\prime}=\left[\frac{N_{1} \hat{\mathbf{x}}_{1}}{c_{1}} \ldots \frac{N_{H} \hat{\mathbf{x}}_{H}}{c_{H}}\right] .
$$

By using the technique of matrix inversion by partition, we get

$$
\left(\mathbf{W}_{s}^{\prime} \mathbf{C}_{s}^{-1} \mathbf{\Pi}_{s}^{-1} \mathbf{W}_{s}\right)^{-1}=\left[\begin{array}{cc}
\left(\mathbf{A}-\mathbf{D B}^{-1} \mathbf{D}^{\prime}\right)^{-1} & -\mathbf{A}^{-1} \mathbf{D} \mathbf{Q} \\
-\mathbf{Q} \mathbf{D}^{\prime} \mathbf{A}^{-1} & \mathbf{Q}
\end{array}\right]
$$

where $\mathbf{Q}=\left(\mathbf{B}-\mathbf{D}^{\prime} \mathbf{A}^{-1} \mathbf{D}\right)^{-1}$. Since

$$
\left(\mathbf{1}_{N}^{\prime} \mathbf{W}_{N}-\mathbf{1}_{s}^{\prime} \boldsymbol{\Pi}_{S}^{-1} \mathbf{W}_{S}\right)=\left[\begin{array}{c}
\mathbf{0}_{H} \\
\hat{\mathbf{x}}_{\boldsymbol{\pi}}-\overline{\mathbf{x}}
\end{array}\right]
$$

where $0_{H}$ is a column vector composed of $H$ zeros, we get

$$
\begin{aligned}
&\left(\mathbf{1}_{N}^{\prime} \mathbf{W}_{N}-\mathbf{1}_{s}^{\prime} \boldsymbol{\Pi}_{S}^{-1} \mathbf{W}_{S}\right)^{\prime}\left(\mathbf{W}_{S}^{\prime} \mathbf{C}_{s}^{-1} \boldsymbol{\Pi}_{S}^{-1} \mathbf{W}_{S}\right)^{-1} \\
&=\left(\hat{\mathbf{x}}_{\boldsymbol{\pi}}-\overline{\mathbf{x}}\right)^{\prime} \mathbf{Q}\left[-\mathbf{D}^{\prime} \mathbf{A}^{-1} \mathbf{I}_{(J \times J)}\right]
\end{aligned}
$$

where $\mathbf{I}_{(J \times J)}$ is a $J \times J$ identity matrix. Since

$$
\begin{aligned}
\mathbf{Q}= & \left\{\sum_{h=1}^{H} \frac{N_{h}}{n_{h} c_{h}} \sum_{k \in S_{h}}\left(\mathbf{x}_{k}-\hat{\mathbf{x}}_{h}\right)\left(\mathbf{x}_{k}-\hat{\mathbf{x}}_{h}\right)^{\prime}\right\}^{-1}, \\
& {\left[-\mathbf{D A}^{-1} \boldsymbol{I}_{(J \times J)}\right]=\left[-\hat{\mathbf{x}} \ldots-\hat{\overline{\mathbf{x}}}_{H} \mathbf{J}_{(J \times J)}\right] }
\end{aligned}
$$

and

$$
\mathbf{W}_{s}^{\prime} \mathbf{C}_{s}^{-1} \boldsymbol{\Pi}_{s}^{-1} \mathbf{y}_{s}=\left[\begin{array}{c}
\frac{N_{1} \hat{y}_{1}}{c_{1}}  \tag{8.23}\\
\vdots \\
\frac{N_{H} \hat{\bar{y}}_{H}}{c_{H}} \\
\sum_{h=1}^{H} \frac{N_{h}}{n_{h} c_{h}} \sum_{k \in S_{h}} x_{k} y_{k}
\end{array}\right],
$$

we get Result 2 by multiplication of (8.22) and (8.23).

## REFERENCES

CASADY, R.J., and VALLIANT, R. (1993). Conditional properties of post-stratified estimators under normal theory. Survey Methoddology, 19, 183-192.
CASSEL, C.-M., SÄRNDAL, C.-E., and WRETMAN, J. H. (1976). Some results on generalized difference estimation and generalized regression estimation for finite population. Biometrika, 63, 615-620.

DEVILLE J.-C. (1992). Constrained samples, conditional inference, weighting: three aspects of the utilisation of auxiliary information. Proceeding of the Workshop: Auxiliary Information in Surveys, Örebro.
DEVILLE J.-C., and SÄRNDAL, C.-E. (1992). Calibration estimators in survey sampling. Journal of the American Statistical Association, 376-382.

FULLER, W.A., and ISAKI, C.T. (1981). Survey design under superpopulation models. In Current Topics in Survey Sampling. (Eds. D. Krewski, R. Platek, J.N.K. Rao, and M.P. Singh). New York: Academic Press, 196-226.
HÁ JEK, J. (1971). Comment on a essay of D. Basu. Foundations of statistical inference. (Éds. V.P. Godambe, et D.A. Sprott), Toronto: Holt, Rinehart and Winston, 236.
HORVITZ, D.G., and THOMPSON, D.J. (1952).A generalization of sampling without replacement from a finite universe. Journal of the American Statistical Association, 7, 663-685.
MADOW, W.G. (1948). On the limiting distributions of estimates based on samples from finite universes. Annals of Mathematical Statistics, 19, 535-545.
MONTANARI, G. E. (1987). Post sampling efficient QR-prediction in large sample survey. International Statistical Review, 55, 191-202.
MONTANARI, G.E. (1997). On conditional properties of finite population mean estimators. Proceeding of the 51st Session of the International Statistical Institute, Contributed paper, 351-352.

MONTANARI, G. E. (1998). On regression estimation of finite population means. Survey Methodology, 24, 69-77.
RAO, J.N.K. (1994). Estimating totals and distribution functions using auxiliary information at the estimation stage. Journal of Official Statistics, 10, 153-165.
RAO, J.N.K. (1997). Development in sample survey theory: an appraisal. Canadian Journal of Statistics, 25, 1-21.
ROSÉN, B. (1972). Asymptotic theory for successive sampling with varying probabilities without replacement. Annals of Mathematical Statistics, 43, 373-397.
SÄRNDAL, C.-E. (1980). On $\pi$-inverse weighting versus best linear unbiased weighting in probability sampling. Biometrika, 67, 639-650.
SÄRNDAL, C.-E (1996). Efficient estimators with simple variance in unequal probability sampling. Journal of the American Statistical Association, 91, 1289-1300.
SÄRNDAL, C.-E., SWENSSON, B., and WRETMAN (1992). Model Assisted Survey Sampling. New York: Springer Verlag.
TILLÉ, Y. (1994). Utilisation d'information auxiliaire en théorie des sondages sans reférence à un modèle. Ph.D Thesis, Université Libre de Bruxelles, Institut de Statistique.
TLLLÉ, Y. (1998). Estimation in surveys using conditional inclusion probabilities: simple random sampling. International Statistical Review, 66, 303-322.
WRIGHT, R.L. (1983). Finite population with multivariate auxiliary information. Journal of the American Statistical Association, 78, 879-883.

# On Robust Small Area Estimation Using a Simple Random Effects Model 

N.G.N. PRASAD and J.N.K. RAO ${ }^{1}$


#### Abstract

Robust small area estimation is studied under a simple random effects model consisting of a basic (or fixed effects) model and a linking model that treats the fixed effects as realizations of a random variable. Under this model a model-assisted estimator of a small area mean is obtained. This estimator depends on the survey weights and remains design-consistent. A model-based estimator of its mean squared error (MSE) is also obtained. Simulation results suggest that the proposed estimator and Kott's (1989) model-assisted estimator are equally efficient, and that the proposed MSE estimator is often much more stable than Kott's MSE estimator, even under moderate deviations of the linking model. The method is also extended to nested error regression models.


KEY WORDS: Design consistent; Linking model; Mean squared error; Survey weights.

## 1. INTRODUCTION

Unit-level random effects models are often used in small area estimation to obtain efficient model-based estimators of small area means. Such estimators typically do not make use of the survey weights (e.g., Ghosh and Meeden 1986; Battese, Harter and Fuller 1988; Prasad and Rao 1990). As a result , the estimators are not design consistent unless the sampling design is self-weighting within areas. We refer the reader to Ghosh and Rao (1994) for an appraisal of small area estimation methods.

Kott (1989) advocated the use of design-consistent model-based estimators (i.e., model assisted estimators) because such estimators provide protection against model failure as the small area sample size increases. He derived a design-consistent estimator of a small area mean under a simple random effects model. This model has two components: the basic (or fixed effects) model and the linking model. The basic model is given by

$$
\begin{equation*}
y_{i j}=\theta_{i}+e_{i j}, j=1,2, \ldots, N_{i} ; i=1,2, \ldots, m \tag{1}
\end{equation*}
$$

where the $y_{i j}$ are the population values and the $e_{i j}$ are uncorrelated random errors with mean zero and variance $\sigma_{i}^{2}$ for each small area $i(=1,2, \ldots, m)$. For simplicity, we take $\theta_{i}$ as the small area mean $\bar{Y}_{i}=\sum_{j} y_{i i j} / N_{i}$, where $N_{i}$ is the number of population units in the $i$-th area. Note that $\bar{Y}_{i}=\theta_{i}+\bar{E}_{i}$ and $\bar{E}_{i}=\sum_{j} e_{i j} / N_{i} \approx 0$ if $N_{i}$ is large.

The linking model assumes that $\theta_{i}$ is a realization of a random variable satisfying the model

$$
\begin{equation*}
\theta_{i}=\mu+v_{i} \tag{2}
\end{equation*}
$$

where the $v_{i}$ are uncorrelated random variables with mean zero and variance $\sigma_{v}^{2}$. Further, $\left\{v_{i}\right\}$ and $\left\{e_{i j}\right\}$ are assumed to be uncorrelated.

Assuming that the model (1) also holds for the sample $\left\{y_{i j} ; j=1,2, \ldots, n_{i} ; i=1,2, \ldots, m\right\}$ and combining the sample model with the linking model, Kott (1989) obtained the familiar unit-level random effects model

$$
\begin{equation*}
y_{i j}=\mu+v_{i}+e_{i j}, j=1,2, \ldots, n_{i} ; i=1,2, \ldots, m, \tag{3}
\end{equation*}
$$

also called the components-of-variance model. It is customary to assume equal variances $\sigma_{i}^{2}=\sigma^{2}$, although the case of random error variances has also been studied (Kleffe and Rao 1992; Arora and Lahiri 1997).

Assuming $\sigma_{i}^{2}=\sigma^{2}$, Kott (1989) derived an efficient estimator $\hat{\theta}_{i K}$ of $\theta_{i}$ which is both model-unbiased under (3) and design-consistent. He also proposed an estimator of its mean squared error (MSE) which is model unbiased under the basic model (1) as well as design-consistent. But this MSE estimator can be quite unstable and can even take negative values, as noted by Kott (1989) in his empirical example. Kott (1989) used his MSE estimators mainly to compare the overall reduction in MSE from using $\hat{\theta}_{i K}$ in place of a direct design-based estimator $\overline{\mathbf{y}}_{\text {iw }}$ given by (4) below. He remarked that more stable MSE estimators are needed.

The main purpose of this paper is to obtain a pseudo empirical best linear unbiased prediction (EBLUP) estimator of $\theta_{i}$ which depends on the survey weights and is designconsistent (section 2). A stable model-based MSE estimator is also obtained (section 3). Results of a simulation study in section 4 show that the proposed MSE estimator is often much more stable than the MSE estimator of Kott, as measured by their coefficient of variation, even under moderate deviations of the linking model (2). Results under the simple model (3) are also extended to a nested error regression model (section 5).

[^8]
## 2. PSEUDO EBLUP ESTIMATOR

Suppose $\tilde{w}_{i j}$ denotes the basic design weight attached to the $j$-th sample unit ( $j=1,2, \ldots, n_{i}$ ) in the $i$-th area ( $i=1,2, \ldots, m$ ). A direct design-based estimator of $\theta_{i}$ is then given by the ratio estimator

$$
\begin{equation*}
\bar{y}_{i w}=\sum_{j} \tilde{w}_{i j} y_{i j} / \sum_{j} \tilde{w}_{i j}=\sum_{j} w_{i j} y_{i j} \tag{4}
\end{equation*}
$$

where $w_{i j}=\tilde{w}_{i j} / \sum_{j} \tilde{w}_{i j}$. The direct estimator $\bar{y}_{i w}$ is designconsistent but fails to borrow strength from the other areas.

To get a more efficient estimator, we consider the following reduced model obtained from the combined model (3) with $\sigma_{i}^{2}=\sigma^{2}$ :

$$
\begin{align*}
\bar{y}_{i w} & =\sum_{j} w_{i j}\left(\mu+v_{i}+e_{i j}\right) \\
& =\mu+v_{i}+\bar{e}_{i w}, \tag{5}
\end{align*}
$$

where the $\bar{e}_{i w}$ are uncorrelated random variables with mean zero and variance $\delta_{i}=\sigma^{2} \sum_{j} w_{i j}^{2}$. The reduced model (5) is an area-level model similar to the well-known Fay-Herriot model (Fay and Herriot 1979). It now follows from the standard best linear unbiased prediction (BLUP) theory (e.g., Prasad and Rao 1990) that the BLUP estimator of $\theta_{i}=\mu+v_{i}$ for the reduced model (5) is given by

$$
\begin{equation*}
\tilde{\theta}_{i}=\tilde{\mu}_{w}+\tilde{v}_{i} \tag{6}
\end{equation*}
$$

where

$$
\tilde{v}_{i}=\gamma_{i v}\left(\bar{y}_{i w}-\tilde{\mu}_{w}\right)
$$

with $\tilde{\boldsymbol{\mu}}_{w}=\sum_{i} \gamma_{i w} \bar{y}_{i w} / \sum_{i} \gamma_{i v}$ and $\gamma_{i w}=\sigma_{v}^{2} /\left(\sigma_{v}^{2}+\delta_{i}\right)$. Note that $\tilde{\theta}_{i}$ is different from the BLUP estimator under the full model (3). We therefore denote $\tilde{\theta}_{i}$ as a pseudo-BLUP estimator. The estimator (6) may also be written as a convex combination of the direct estimator $\bar{y}_{i w}$ and $\tilde{\mu}_{w}$ :

$$
\begin{equation*}
\tilde{\theta}_{i}=\gamma_{i w} \bar{y}_{i w}+\left(1-\gamma_{i w}\right) \tilde{\mu}_{w} . \tag{7}
\end{equation*}
$$

The estimator $\tilde{\theta}_{i}$ depends on the parameters $\sigma_{v}^{2}$ and $\sigma^{2}$ which are generally unknown in practice. We therefore replace $\sigma_{v}^{2}$ and $\sigma^{2}$ in (7) by model-consistent estimators $\hat{\sigma}_{v}^{2}$ and $\hat{\sigma}^{2}$ under the original unit-level model (3) to obtain the estimator

$$
\begin{equation*}
\hat{\theta}_{i}=\hat{\gamma}_{i w} \bar{y}_{i w}+\left(1-\hat{\gamma}_{i w}\right) \hat{\mu}_{w} \tag{8}
\end{equation*}
$$

where

$$
\hat{\gamma}_{i w}=\hat{\sigma}_{v}^{2} /\left(\hat{\sigma}_{v}^{2}+\hat{\sigma}^{2} \sum_{j} w_{i j}^{2}\right)
$$

and

$$
\hat{\mu}_{w}=\sum_{i} \hat{\gamma}_{i w} \bar{y}_{i w} / \sum_{i} \hat{\gamma}_{i w^{*}}
$$

The estimator $\hat{\theta}_{i}$ will be referred to as pseudo-EBLUP estimator. We use standard estimators of $\sigma_{v}^{2}$ and $\sigma^{2}$, based on the within-area sums of squares

$$
Q_{w}=\sum_{i} \sum_{j}\left(y_{i j}-\bar{y}_{i}\right)^{2}
$$

and the between-area sums of squares

$$
Q_{b}=\sum_{i} n_{i}\left(\bar{y}_{i}-\bar{y}\right)^{2}
$$

where $\bar{y}=\sum_{i} n_{i} \bar{y}_{i} / \sum_{i} n_{i}$ is the overall sample mean. We have

$$
\hat{\sigma}^{2}=Q_{w} /\left(\sum_{i} n_{i}-m\right)
$$

and $\hat{\sigma}_{v}^{2}=\max \left(\tilde{\sigma}_{v}^{2}, 0\right)$ where

$$
\tilde{\sigma}_{v}^{2}=\left[Q_{b}-(m-1) \hat{\sigma}^{2}\right] / n^{*}
$$

with

$$
n^{*}=\sum_{i} n_{i}-\sum_{i} n_{i}^{2} / \sum_{i} n_{i}
$$

It may be noted that $\sigma_{v}^{2}$ and $\sigma^{2}$ are either not estimable or poorly estimated from the reduced model (5) due to identifiability problems. Following Kackar and Harville (1984), it can be shown that the pseudo-EBLUP estimator $\theta_{i}$ is model-unbiased for $\theta_{i}$ under the original model (3) for symmetrically distributed errors $\left\{v_{i}\right\}$ and $\left\{e_{i j}\right\}$, not necessarily normal. It is also design consistent, assuming that $n_{i} \sum_{j} w_{i j}^{2}$ is bounded as $n_{i}$ increases, because $\hat{\gamma}_{i w}$ converges in probability to 1 as $n_{i} \rightarrow \infty$ regardless of the validity of the model (3), assuming $\hat{\sigma}_{v}^{2}$ and $\hat{\sigma}^{2}$ converge in probability to some values, say, $\hat{\sigma}_{v}^{* 2}$ and $\sigma^{* 2}$.

Kott's (1989) model-based estimator of $\theta_{i}$ is obtained by taking a weighted combination of $\bar{y}_{i w}$ and $\sum_{l * i} c_{l}^{(i)} \bar{y}_{l}$, that is,

$$
f_{i}\left(\alpha_{i}, c^{(i)}\right)=\left(1-\alpha_{i}\right) \bar{y}_{i w}+\alpha_{i} \sum_{l \neq i} c_{l}^{(i)} \bar{y}_{l},
$$

and then minimizing the model mean squared error (MSE) of $f_{i}\left(\alpha_{i}, c^{(i)}\right)$ with respect to $\alpha_{i}$ and $c_{l}^{(i)}$ subject to modelunbiasedness condition: $\sum_{l \neq i} c_{l}^{(i)}=1$. This leads to

$$
\begin{equation*}
\hat{\theta}_{i K}=f_{i}\left(\hat{\alpha}_{i}, \hat{c}^{(i)}\right) \tag{9}
\end{equation*}
$$

with

$$
\begin{aligned}
& \hat{\alpha}_{i}= \\
& \sum_{j}\left[w_{i j}^{2} /\left\{\sum_{j} w_{i j}^{2}+\sum_{l * i} \hat{c}_{l}^{(i)^{2}} / n_{i}+\left(1+\sum_{l \neq i} \hat{c}_{l}^{(i)^{2}}\right)\left(\hat{\sigma}_{v}^{2} / \hat{\sigma}^{2}\right)\right\}\right]
\end{aligned}
$$

and

$$
\hat{c}_{l}^{(i)}=\left[\left(\hat{\sigma}_{v}^{2} / \hat{\sigma}^{2}\right)+n_{l}^{-1}\right] / \sum_{h * i}\left[\left(\hat{\sigma}_{v}^{2} / \hat{\sigma}^{2}\right)+n_{h}^{-1}\right] .
$$

The estimator $\hat{\theta}_{i K}$ is also model-unbiased and designconsistent. In a previous version of this paper, we proposed an estimator similar to (9). It uses the best estimators of $\mu$ under the unit-level model, based on the unweighted means $\bar{y}_{i}$, rather than $\hat{\mu}_{w}$, the best estimator of $\mu$ under the reduced model (4), based on the survey-weighted means $\bar{y}_{i w}$.

## 3. ESTIMATORS OF MSE

It is straightforward to derive the MSE of the pseudoBLUP estimator $\tilde{\theta}_{i}$ under the unit level model (3). We have

$$
\begin{equation*}
\operatorname{MSE}\left(\tilde{\theta}_{i}\right)=E\left(\tilde{\theta}_{i}-\theta_{i}\right)^{2}=g_{1 i}\left(\sigma_{v}^{2}, \sigma^{2}\right)+g_{2 i}\left(\sigma_{v}^{2}, \sigma^{2}\right) \tag{10}
\end{equation*}
$$

with

$$
g_{1 i}\left(\sigma_{v}^{2}, \sigma^{2}\right)=\left(1-\gamma_{i v}\right) \sigma_{v}^{2}
$$

and

$$
g_{2 i}\left(\sigma_{v}^{2}, \sigma^{2}\right)=\sigma_{v}^{2}\left(1-\gamma_{i w}\right)^{2} / \sum_{i} \gamma_{i w}
$$

The leading term, $g_{1 i}\left(\sigma_{v}^{2}, \sigma^{2}\right)$ is of order $O(1)$, while the second term, $g_{2 i}\left(\sigma_{v}^{2}, \sigma^{2}\right)$, due to estimation of $\mu$ is of order $O\left(m^{-1}\right)$ for large $m$.

A naive MSE estimator of the pseudo-EBLUP estimator $\hat{\theta}_{i}$ is obtained by estimating $\operatorname{MSE}\left(\hat{\theta}_{i}\right)$ given by (10):

$$
\begin{equation*}
\operatorname{mse}_{\mathrm{N}}\left(\hat{\theta}_{i}\right)=g_{1 i}\left(\hat{\sigma}_{v}^{2}, \hat{\sigma}^{2}\right)+g_{2 i}\left(\hat{\sigma}_{v}^{2}, \hat{\sigma}^{2}\right) \tag{11}
\end{equation*}
$$

But (11) could lead to significant underestimation of $\operatorname{MSE}\left(\hat{\theta}_{i}\right)$ because it ignores the uncertainty associated with $\hat{\sigma}_{v}^{2}$ and $\hat{\boldsymbol{\sigma}}^{2}$. Note that

$$
\begin{equation*}
\operatorname{MSE}\left(\hat{\theta}_{i}\right)=\operatorname{MSE}\left(\tilde{\theta}_{i}\right)+E\left(\hat{\theta}_{i}-\tilde{\theta}_{i}\right)^{2} \tag{12}
\end{equation*}
$$

under normality of the errors $\left\{v_{j}\right\}$ and $\left\{e_{i j}\right\}$ so that $\operatorname{MSE}\left(\tilde{\theta}_{i}\right)$ is always smaller than $\operatorname{MSE}\left(\hat{\theta}_{i}\right)$; see Kackar and Harville (1984).

To get a "correct" estimator of $\operatorname{MSE}\left(\hat{\theta}_{i}\right)$, we first approximate the second order term $E\left(\hat{\theta}_{i}-\tilde{\theta}_{i}\right)^{2}$ in (12) for large $m$, assuming that $\left\{v_{i}\right\}$ and $\left\{e_{i j}\right\}$ are normally distributed. Following Prasad and Rao (1990), we have

$$
\begin{equation*}
E\left(\hat{\theta}_{i}-\tilde{\theta}_{i}\right)^{2} \approx g_{3 i}\left(\sigma_{\nu}^{2}, \sigma^{2}\right) \tag{13}
\end{equation*}
$$

where the neglected terms are of lower order than $m^{-1}$, and

$$
\begin{align*}
& g_{3 i}\left(\sigma_{v}^{2}, \sigma^{2}\right)=\gamma_{i w}\left(1-\gamma_{i w}\right)^{2} \sigma_{v}^{-2} \\
& \left\{V\left(\tilde{\sigma}_{v}^{2}\right)-2\left(\sigma_{v}^{2} / \sigma^{2}\right) \operatorname{Cov}\left(\tilde{\sigma}_{v}^{2}, \hat{\sigma}^{2}\right)+\right. \\
& \left.\left(\sigma_{v}^{2} / \sigma^{2}\right)^{2} \operatorname{Var}\left(\hat{\sigma}^{2}\right)\right\} ; \tag{14}
\end{align*}
$$

see Appendix 1. The variances and covariances of $\hat{\sigma}_{v}^{2}$ and $\hat{\boldsymbol{\sigma}}^{2}$ are also given in the Appendix 1. It can be shown that $g_{1 i}\left(\hat{\sigma}_{v}^{2}, \hat{\sigma}^{2}\right)+g_{3 i}\left(\hat{\sigma}_{v}^{2}, \hat{\sigma}^{2}\right)$ is approximately unbiased for $g_{1 i}\left(\sigma_{v}^{2}, \sigma^{2}\right)$ in the sense that its bias is of lower order than $m^{-1}$ (see Appendix 2). Similarly, $g_{2 i}\left(\hat{\sigma}_{v}^{2}, \hat{\sigma}^{2}\right)$ and $g_{3 i}\left(\hat{\sigma}_{v}^{2}, \hat{\sigma}_{2}^{2}\right)$ are approximately unbiased for $g_{2 i}\left(\sigma_{v}^{2}, \sigma^{2}\right)$ and $g_{3 i}\left(\sigma_{v}^{2}, \sigma^{2}\right)$, respectively. It now follows that an approximately model-unbiased estimator of $\operatorname{MSE}\left(\hat{\theta}_{i}\right)$ is given by

$$
\begin{equation*}
\operatorname{mse}\left(\hat{\theta}_{i}\right)=g_{1 i}\left(\hat{\sigma}_{v}^{2}, \hat{\sigma}^{2}\right)+g_{2 i}\left(\hat{\sigma}_{v}^{2}, \hat{\sigma}^{2}\right)+2 g_{3 i}\left(\hat{\sigma}_{v}^{2}, \hat{\sigma}^{2}\right) \tag{15}
\end{equation*}
$$

For the estimator $\hat{\theta}_{i K}$ given by (9), Kott (1989) proposed an estimator of MSE as

$$
\begin{equation*}
\operatorname{mse}\left(\hat{\theta}_{i K}\right)=\left(1-2 \hat{\alpha}_{i}\right) v^{*}\left(\bar{y}_{i w}\right)+\hat{\alpha}_{i}^{2}\left(\bar{y}_{i w}-\sum_{l \neq i} c_{l}^{(i)} \bar{y}_{l}\right)^{2}, \tag{16}
\end{equation*}
$$

where $v^{*}\left(\bar{y}_{i w}\right)$ is both a design-consistent estimator of the design-MSE of $\bar{y}_{i w}$ and a model-unbiased estimator of the model-variance of $\bar{y}_{i w}$ under the basic model (1). Since $\hat{\alpha}_{i}$ converges in probability to zero as $n_{i} \rightarrow \infty$, it follows from (16) that mse $\left(\hat{\theta}_{i K}\right)$ is also both design-consistent and model unbiased assuming only the basic model (1). However, mse $\left(\hat{\theta}_{i K}\right)$ is unstable and can even take negative values when $\hat{\alpha}_{i}$ exceeds 0.5 , as noted by Kott (1989).

Note that our MSE estimator, $\operatorname{mse}\left(\hat{\theta}_{i}\right)$ is based on the full model (3) obtained by combining the basic model (1) with the linking model (2). However, our simulation results in section 4 show that it may perform well even under moderate deviations from the linking model.

## 4. SIMULATION STUDY

We conducted a limited simulation study to evaluate the performances of the proposed estimator $\hat{\theta}_{i}$, given by ( 8 ), and its estimator of MSE, given by (15), relative to Kott's estimator $\hat{\theta}_{i K}$, given by (9), and its estimator of MSE, given by (16). We studied the performances under two different approaches: (i) For each simulation run, a finite population of $m=30$ small areas with $N_{i}=200$ population units in each area is generated from the assumed unit-level model and then a PPS (probability proportional to size) sample within each small area is drawn independently, using $n_{i}=20$. (ii) A fixed finite population is first generated from the assumed unit-level model and then for each simulation run a PPS sample within each small area is drawn independently, employing the fixed finite population. Approach (i) refers to both the design and the linking model whereas approach (ii) is design-based in the sense that it refers only to the design. The errors $\left\{v_{i}\right\}$ and $\left\{e_{i j}\right\}$ are assumed to be normally distributed in generating the finite populations $\left\{y_{i j}, i=1,2, \ldots, 30 ; j=1,2, \ldots, 200\right\}$. We considered two cases: (1) The linking model (2) is true with $\mu=50$. (2) The linking model is violated by letting $\mu$ vary across areas: $\mu_{i}=50, i=1,2, \ldots, 10 ; \mu_{i}=55, i=11,12, \ldots, 20 ; \mu_{i}=60$,
$i=21,22, \ldots, 30$. To implement PPS sampling within each area, size measures $z_{i j}(i=1,2, \ldots, 30 ; j=1,2, \ldots, 200)$ were generated from an exponential distribution with mean 200. Using these $z$-values, we computed selection probabilities $p_{i j}=z_{i j} / \sum_{i j} z_{i j}$ for each area $i$ and then used them to select PPS with replacement samples of sizes $n_{i}=n$, by taking $n=20$, and the associated sample values $\left\{y_{i j}\right\}$ were observed.

The basic design weights are given by $\tilde{w}_{i j}=n^{-1} p_{i j}^{-1}$ so that $w_{i j}=p_{i j}^{-1} / \sum_{j} p_{i j}^{-1}$. Using these weights and the associated sample values $y_{i j}$ we computed estimates $\hat{\theta}_{i}$ and $\hat{\theta}_{i K}$ and associated estimates of MSE, and also the ratio estimate $\bar{y}_{i w}$ for each simulation run; the formula for $v^{*}\left(\bar{y}_{i w}\right)$ under PPS sampling is given in Appendix 3. This process was repeated $R=10,000$ times to get from each run $r(=1,2, \ldots, R) \hat{\theta}_{i}(r)$ and $\hat{\theta}_{i K}(r)$ and associated MSE estimates $\mathrm{mse}_{\mathrm{i}}\left(\hat{\theta}_{i}(r)\right)$ and $\operatorname{mse}_{\mathrm{i}}\left(\hat{\theta}_{i K}(r)\right)$ and also the direct estimate $\bar{y}_{i w}(r)$. Using these values, empirical relative efficiencies (RE) of $\hat{\theta}_{i}$ and $\hat{\theta}_{i K}$ over $\bar{y}_{i w}$ were computed as

$$
\operatorname{RE}\left(\hat{\theta}_{i}\right)=\operatorname{MSE}_{.}\left(\bar{y}_{i w}\right) / \operatorname{MSE}_{*}\left(\hat{\theta}_{i}\right)
$$

and

$$
\operatorname{RE}\left(\hat{\theta}_{i K}\right)=\operatorname{MSE}_{*}\left(\bar{y}_{i w}\right) / \operatorname{MSE}_{*}\left(\hat{\theta}_{i K}\right)
$$

where MSE. denotes the MSE over $R=10,000$ runs. For $\overline{\bar{Y}}^{2}(r)$ MSE, $\left(\hat{\theta}_{i}\right)=\sum_{r}\left[\hat{\theta}_{i}(r)-\bar{Y}_{i}(r)\right]^{2} / R$, where $\bar{Y}_{i}(r)$ is the $i$-th area population mean for the $r$-th run. Note that $\bar{Y}_{i}(r)$ remains the same over the runs $r$ under the design-based approach because the finite population is fixed over the simulation runs.

Similarly, the relative biases of the MSE estimators were computed as

$$
\operatorname{RB}\left[\operatorname{mse}\left(\hat{\theta}_{i}\right)\right]=\left[\operatorname{MSE}_{*}\left(\hat{\theta}_{i}\right)-E_{*} \operatorname{mse}\left(\hat{\theta}_{i}\right)\right] / \operatorname{MSE}_{.}\left(\hat{\theta}_{i}\right)
$$

and

$$
\begin{aligned}
\operatorname{RB}\left[\operatorname{mse}\left(\hat{\theta}_{i K}\right)\right]= & {\left[\operatorname{MSE}_{*}\left(\hat{\theta}_{i K}\right)-E \operatorname{mse}\left(\hat{\theta}_{i K}\right)\right] / } \\
& \operatorname{MSE}_{*}\left(\hat{\theta}_{i K}\right)
\end{aligned}
$$

where $E$. denotes the expectation over $R=10,000$ runs. For example, $E$ mse $\left(\hat{\theta}_{i}\right)=\sum_{r} \operatorname{mse}\left(\hat{\theta}_{i}(r)\right) / R$. Finally, the empirical coefficient of variation (CV) of the MSE estimators were computed as

$$
\mathrm{CV}\left[\operatorname{mse}\left(\hat{\theta}_{i}\right)\right]=\left[\operatorname{MSE}_{.}\left\{\operatorname{mse}\left(\hat{\theta}_{i}\right)\right\}\right]^{1 / 2} / \operatorname{MSE}_{*}\left(\hat{\theta}_{i}\right)
$$

and

$$
\operatorname{CV}\left[\operatorname{mse}\left(\hat{\theta}_{i K}\right)\right]=\left[\operatorname{MSE} .\left\{\operatorname{mse}\left(\hat{\theta}_{i K}\right)\right\}\right]^{1 / 2} / \operatorname{MSE}_{*}\left(\hat{\theta}_{i K}\right)
$$

Note that $\operatorname{MSE}_{*}\left[\operatorname{mse}\left(\hat{\theta}_{i}\right)\right]=\sum_{r}\left[\operatorname{mse}\left(\hat{\theta}_{i}(r)\right)-\operatorname{MSE}_{*}\left(\hat{\theta}_{i}\right)\right]^{2} / R$ and a similar expression for MSE, [mse $\left.\left(\hat{\theta}_{i K}\right)\right]$.

Table 1 reports summary measures of the values of percent RE, |RB| and CV for cases (1) and (2) under approach (i). Summary measures under approach (ii) are reported in Table 2. Summary measures considered are the mean and the median (med) over the small areas $i=1,2, \ldots, 30$.

Table 1
Relative Efficiency (RE) of Estimators, Absolute Relative Bias ( $\mathbb{R B I}$ ) and Coefficient of Variation (CV) of MSE estimators ( $\sigma=5.0, n=20$ ): Approach (i)


Case 1: $\mu_{i}=50, i=1,2, \ldots, 30 ;$ Case 2: $\mu_{i}=50, i=1,2, \ldots, 10$; $\mu_{i}=55, \mathrm{i}=11,12, \ldots, 20 ; \mu_{i}=60, i=21,22, \ldots, 30$.

It is clear from Tables 1 and 2 that $\hat{\theta}_{i K}$ and $\hat{\theta}_{i}$ perform similarly with respect to RE which decreases as $\sigma_{v} / \sigma$ increases. Under approach (ii), RE is large for both cases 1 and 2 when $\sigma_{v} / \sigma \leq 0.4$, whereas it decreases significantly under approach (i) if the linking model is violated (case 2); the direct estimator $\bar{y}_{i w}$ is quite unstable under approach (ii).

Turning to the performance of MSE estimators under approach (i), Table 1 shows that IRBI of mse ( $\hat{\theta}_{i}$ ) is negligible ( $<4 \%$ ) when the linking model holds (Case 1) and that it is small ( $<10 \%$ ) even when the linking model is violated, although it increases. The estimator mse $\left(\hat{\theta}_{i K}\right)$ has a larger IRBI but it is less than $15 \%$. The CV of mse $\left(\hat{\theta}_{i}\right)$ is much smaller than the CV of mse ( $\hat{\theta}_{i K}$ ) for both Cases 1 and 2 . For example, when the model holds (Case 1) the median CV is $25 \%$ for mse $\left(\hat{\theta}_{i}\right)$ compared to $148 \%$ for mse $\left(\hat{\theta}_{i K}\right)$ when $\sigma_{v}=1$; the median CV decreases to $8 \%$ for $\mathrm{mse}\left(\hat{\theta}_{i}\right)$ compared to $48 \%$ for mse $\left(\hat{\theta}_{i K}\right)$ when $\sigma_{v}=2$. This pattern is retained when the model is violated (Case 2). It may be noted that the probability of $\operatorname{mse}\left(\hat{\theta}_{i K}\right)$ taking a negative value is quite large ( $>0.3$ ) when $\sigma_{\nu} / \sigma \leq 0.4$.

Under approach (ii), Table 2 shows that IRBI of mse ( $\hat{\theta}_{i}$ ) is larger than the value under approach (i) and ranges from $15 \%$ to $25 \%$. On the other hand, IRBI of mse ( $\hat{\theta}_{i K}$ ) is smaller and ranges from $4 \%$ to $15 \%$. The CV of mse $\left(\hat{\theta}_{i K}\right)$, how-
ever, is much larger than under approach (i). For example, the median CV for Case 1 is $295 \%$ compared to $38 \%$ for mse $\left(\hat{\theta}_{i}\right)$ when $\sigma_{\nu}=1$ which decreases to $122 \%$ compared to $23 \%$ when $\sigma_{v}=2$. A similar pattern holds for case 2 where the fixed finite population is generated from the model with varying means.

Table 2
Relative Efficiency (RE) of Estimators, Absolute Relative Bias (IRB1) and Coefficient of Variation (CV) of MSE estimators ( $\sigma=5.0, n=20$ ): Approach (ii)

| $\sigma_{v}$ |  | RE\% | \|RB1\% |  | CV\% |  |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: |
|  | $\hat{\theta}_{i K} \hat{\theta}_{i} \operatorname{mse}\left(\hat{\theta}_{i k}\right) \operatorname{mse}\left(\hat{\theta}_{i}\right) \operatorname{mse}\left(\hat{\theta}_{i K}\right) \operatorname{mse}\left(\hat{\theta}_{i}\right)$ |  |  |  |  |  |
|  | Case 1 |  |  |  |  |  |
| 1 | Mean | 283281 | 14.2 | 25.4 | 289 | 39 |
|  | Med | 275279 | 15.0 | 24.7 | 295 | 38 |
| 2 | Mean | 180182 | 7.3 | 19.2 | 115 | 24 |
|  | Med | 177181 | 6.9 | 18.7 | 122 | 23 |
| 3 | Mean | 129129 | 4.8 | 14.8 | 68 | 24 |
|  | Med | 129128 | 4.2 | 13.9 | 65 | 24 |
| Case 2 |  |  |  |  |  |  |
| 1 | Mean | 278276 | 15.7 | 26.8 | 291 | 41 |
|  | Med | 271275 | 16.6 | 26.2 | 297 | 40 |
| 2 | Mean | 175177 | 8.8 | 20.7 | 117 | 26 |
|  | Med | 173177 | 8.5 | 20.3 | 124 | 25 |
| 3 | Mean | 124124 | 6.3 | 16.2 | 70 | 25 |
|  | Med | 125124 | 6.8 | 15.5 | 67 | 26 |

Case 1: $\mu_{i}=50, i=1,2, \ldots, 30$; Case 2: $\mu_{i}=50, i=1,2, \ldots, 10$; $\mu_{i}=55, i=11,12, \ldots, 20 ; \mu_{i}=60, i=21,22, \ldots, 30$.

To reduce $\mathbb{R B} \mid$ of mse $\left(\hat{\theta}_{i}\right)$ under approach (ii), one could combine it with mse $\left(\hat{\theta}_{i K}\right)$ by taking a weighted average, but it appears difficult to chose the appropriate weights. The weighted average will be more stable than $\operatorname{mse}\left(\hat{\theta}_{i K}\right)$.

## 5. NESTED ERROR REGRESSION MODEL

The results in sections 2 and 3 can be extended to nested error regression models

$$
\begin{equation*}
y_{i j}=x_{i j}^{\prime} \beta+v_{i}+e_{i j}, j=1,2, \ldots, n_{i} ; i=1,2, \ldots, m \tag{17}
\end{equation*}
$$

using the results of Prasad and Rao (1990), where $x_{i j}$ is a $p$-vector of auxiliary variables with known population mean $\bar{X}_{i}$ and related to $y_{i j}$, and $\beta$ is the $p$-vector of regression coefficients. The reduced model is given by

$$
\begin{equation*}
\bar{y}_{i w}=\bar{x}_{i w}^{\prime} \beta+v_{i}+\bar{e}_{i w} \tag{18}
\end{equation*}
$$

with $\bar{x}_{i w}^{\prime}=\sum_{j} w_{i j} x_{i j}$. Model-consistent estimates $\hat{\sigma}_{v}^{2}$ and $\hat{\sigma}^{2}$ are obtained from the unit-level model (17), employing either the method of fitting constants (Prasad and Rao 1990) or REML (restricted maximum likelihood) estimation (Datta and Lahiri 1997).

The pseudo-EBLUP of $\theta_{i}=\bar{X}_{i}^{\prime} \beta+v_{i}$ is given by

$$
\begin{equation*}
\hat{\theta}_{i}=\hat{\gamma}_{i w} \bar{y}_{i w}+\left(1-\hat{\gamma}_{i w}\right) \bar{X}_{i}^{\prime} \hat{\beta}_{w}, \tag{19}
\end{equation*}
$$

where

$$
\hat{\beta}_{w}=\left(\sum_{i} \hat{\gamma}_{i w} \bar{x}_{i w} \bar{x}_{i w}^{\prime}\right)^{-1}\left(\sum_{i} \hat{\gamma}_{i w} \bar{x}_{i w} \bar{y}_{i w}\right) .
$$

An approximate model-unbiased estimator of $\operatorname{MSE}\left(\hat{\theta}_{i}\right)$ is given by (15) with

$$
g_{1 i}\left(\hat{\sigma}_{v}^{2}, \hat{\sigma}^{2}\right)=\left(1-\hat{\gamma}_{i w}\right) \hat{\sigma}_{v}^{2}
$$

as before,

$$
\begin{aligned}
& g_{2 i}\left(\hat{\sigma}_{v}^{2}, \hat{\sigma}^{2}\right)= \\
& \quad \hat{\sigma}_{v}^{2}\left(\bar{X}_{i}-\hat{\gamma}_{i w} \bar{x}_{i w}\right)^{\prime}\left(\sum_{i} \hat{\gamma}_{i w} \bar{x}_{i w} \bar{x}_{i w}^{\prime}\right)^{-1}\left(\bar{X}_{i}-\hat{\gamma}_{i w} x_{i w} \bar{x}_{i w}\right)
\end{aligned}
$$

and $g_{3 i}\left(\hat{\sigma}_{v}^{2}, \hat{\sigma}^{2}\right)$, obtained from (14), involves the estimated variances and covariances of $\tilde{\sigma}_{v}^{2}$ and $\hat{\sigma}^{2}$. The latter can be obtained from Prasad and Rao (1990) for the method of fitting constants and from Datta and Lahiri (1997) for REML.

## 6. CONCLUSION

We have proposed a model-assisted estimator of a small area mean under a simple unit-level random effects model. This estimator depends on the survey weights and is designconsistent. We have also obtained a model-based MSE estimator. Results of our simulation study have shown that the proposed MSE estimator performs well, even under moderate deviations of the linking model. The proposed approach is also extended to a nested error regression model.

## ACKNOWLEDGEMENTS

This work was supported by research grants from the Natural Sciences and Engineering Research Council of Canada. We are thankful to the Associate Editor and the referee for constructive comments and suggestions.

## APPENDIX 1

## Proof of (13):

From general results (Prasad and Rao 1990) we have

$$
E\left(\hat{\theta}_{i}-\tilde{\theta}_{i}\right)^{2} \approx \operatorname{tr}\left[A_{i}\left(\sigma_{v}^{2}, \sigma^{2}\right) B_{i}\left(\sigma_{v}^{2}, \sigma^{2}\right)\right],
$$

where $B_{i}\left(\sigma_{v}^{2}, \sigma^{2}\right)$ is the $2 \times 2$ covariance matrix of $\tilde{\sigma}_{v}^{2}$ and $\hat{\sigma}^{2}$, and $A_{i}\left(\sigma_{v}^{2}, \sigma^{2}\right)$ is the $2 \times 2$ covariance matrix of

$$
\left(\frac{\partial \theta_{i}^{*}}{\partial \sigma_{v}^{2}}, \frac{\partial \theta_{i}^{*}}{\partial \sigma^{2}}\right) .
$$

Now, noting that

$$
\begin{aligned}
& \frac{\partial \theta_{i}^{*}}{\partial \sigma_{v}^{2}}=\frac{\partial \gamma_{i w}}{\partial \sigma_{v}^{2}} \bar{y}_{i w}=\left[\frac{\gamma_{i w}\left(1-\gamma_{i w}\right)}{\sigma_{v}^{2}}\right] \bar{y}_{i w} \\
& \frac{\partial \theta_{i}^{*}}{\partial \sigma^{2}}=\frac{\partial \gamma_{i w}}{\partial \sigma^{2}} \bar{y}_{i w}=-\left[\frac{\gamma_{i w}\left(1-\gamma_{i w}\right)}{\sigma^{2}}\right] \bar{y}_{i w}
\end{aligned}
$$

and $V\left(\bar{y}_{i w}\right)=\sigma_{v}^{2}+\delta_{i}=\sigma_{v}^{2} / \gamma_{i w}$, we get

$$
A_{i}\left(\sigma_{v}^{2}, \sigma^{2}\right)=\left[\gamma_{i w}\left(1-\gamma_{i w}\right)^{2} \sigma_{v}^{-2}\right]\left[\begin{array}{cc}
1 & -\sigma_{v}^{2} / \sigma^{2} \\
-\sigma_{v}^{2} / \sigma^{2} & \left(\sigma_{v}^{2} / \sigma\right)^{2}
\end{array}\right],
$$

and hence the result (14).
Covariance matrix of $\tilde{\boldsymbol{\sigma}}_{v}^{\mathbf{2}}$ and $\hat{\boldsymbol{\sigma}}^{\mathbf{2}}$ :
Under normality, we have

$$
\begin{aligned}
& V\left(\hat{\sigma}^{2}\right)=2 \sigma^{4} /\left(\sum_{i} n_{i}-m\right), \\
& V\left(\tilde{\sigma}_{v}^{2}\right)=2 n_{*}^{-2} \\
& {\left[\sigma^{4}(m-1)\left(\sum n_{i}-1\right)\left(\sum n_{i}-m\right)^{-1}+2 n_{*} \sigma^{2} \sigma_{v}^{2}+n_{* *} \sigma_{v}^{4}\right]}
\end{aligned}
$$

and

$$
\operatorname{Cov}\left(\hat{\sigma}^{2}, \tilde{\sigma}_{v}^{2}\right)=-(m-1) n_{*}^{-1} V\left(\hat{\sigma}^{2}\right),
$$

where

$$
n_{* *}=\sum n_{i}^{2}-2 \sum n_{i}^{3} / \sum n_{i}+\left(\sum n_{i}^{2}\right)^{2} /\left(\sum n_{i}\right)^{2}
$$

see Searle, Casella and McCulloch (1992, p. 428).

## APPENDIX 2

Proof of $E\left[g_{1 i}\left(\hat{\sigma}_{v}^{2}, \hat{\sigma}^{2}\right)+g_{z i}\left(\hat{\sigma}_{v}^{2}, \hat{\sigma}^{2}\right)\right] \approx g_{1 i}\left(\sigma_{v}^{2}, \sigma^{2}\right)$ :
By a Taylor expansion of $g_{1 i}\left(\hat{\sigma}_{v}^{2}, \hat{\sigma}^{2}\right)$ around $\left(\sigma_{v}^{2}, \sigma^{2}\right)$ to second order and noting that $E\left(\hat{\sigma}^{2}-\sigma^{2}\right)=0$ and $E\left(\hat{\sigma}_{v}^{2}-\sigma_{v}^{2}\right) \approx 0$, we get

$$
\begin{aligned}
E\left[g_{1 i}\left(\hat{\sigma}_{v}^{2}, \hat{\sigma}^{2}\right)-g_{1 i}\left(\sigma_{v}^{2}, \sigma^{2}\right)\right] & \\
& \approx \frac{1}{2} \operatorname{tr}\left[D_{i}\left(\sigma_{v}^{2}, \sigma^{2}\right) B_{i}\left(\sigma_{v}^{2}, \sigma^{2}\right)\right]
\end{aligned}
$$

where $D_{i}\left(\sigma_{v}^{2}, \sigma^{2}\right)$ is the $2 \times 2$ matrix of second order derivatives of $g_{1 i}\left(\sigma_{v}^{2}, \sigma^{2}\right)$ with respect to $\sigma_{v}^{2}$ and $\sigma^{2}$. It is easy to verify that

$$
\frac{1}{2} \operatorname{tr}\left[D_{i}\left(\sigma_{v}^{2}, \sigma^{2}\right) B_{i}\left(\sigma_{v}^{2}, \sigma^{2}\right)\right]=g_{3 i}\left(\sigma_{v}^{2}, \sigma^{2}\right) .
$$

Now, noting that $E\left[g_{3 i}\left(\hat{\sigma}_{v}^{2}, \hat{\sigma}^{2}\right)\right] \approx g_{3 i}\left(\sigma_{v}^{2}, \sigma^{2}\right)$ we get the desired result.

## APPENDIX 3

The design-based estimator of variance of $\bar{y}_{i v}$ under PPS sampling is given by

$$
v\left(\bar{y}_{i w}\right)=\frac{m}{m-1} \sum_{j} w_{i j}^{2}\left(y_{i j}-\bar{y}_{i w}\right)^{2} .
$$

Kott (1989) model-assisted variance estimator is

$$
\begin{aligned}
v^{*}\left(\bar{y}_{i w}\right) & =\left\{V\left(\bar{y}_{i w}\right) / E v\left(\bar{y}_{i w}\right)\right\} v\left(\bar{y}_{i w}\right) \\
& =\frac{\left(\sum_{j} w_{i j}^{2}\right) \sum_{j} w_{i j}^{2}\left(y_{i j}-\bar{y}_{i w}\right)^{2}}{\sum_{j} w_{i j}^{2}\left(1-2 w_{i j}+\sum_{j} w_{i j}^{2}\right)},
\end{aligned}
$$

where $E$ and $V$ denote expectation and variance with respect to the basic model (1).

## REFERENCES

ARORA,V., and LAHIRI, P. (1997). On the superiority of the Bayesian method over the BLUP in small area estimation problems. Statistica Sinica, 7, 1053-1063.

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

DATTA, G.S., and LAHIRI, P. (1997). A Unified Measure of Uncertainty of Estimated Best Linear Unbiased Predictor in Small-Area Estimation Problems. Technical Report, University of Nebraska-Lincoln.

FAY, R.E., and HERRIOT, R.A. (1979). Estimates of income for small places: an application of James-Stein procedures to census data. Journal of the American Statistical Association, 74, 269-277.
GHOSH, M., and MEEDEN, G. (1986). Empirical Bayes estimation in finite population sampling. Journal of the American Statistical Association, 81, 1058-1069.

GHOSH, M., and RAO, J.N.K. (1994). Small area estimation: an appraisal. Statistical Science, 9, 55-93.
KACKAR, R.N., and HARVILLE, D.A. (1984). Approximations for standard errors of estimators for fixed and random effects models. Journal of the American Statistical Association, 79, 853-862.

KLEFFE, J., and RAO, J.N.K. (1992). Estimation of mean square error of empirical best linear unbiased predictors under a random error variance linear model. Journal of Multivariate Analysis, 43, 1-15.

KOTT, P. (1989). Robust small domain estimation using random effects modelling. Survey Methodology, 15, 3-12.
PRASAD, N.G.N., and RAO, J.N.K. (1990). The estimation of mean squared errors of small-area estimators. Journal of the American Statistical Association, 85, 163-171.

SEARLE, S.R., CASELLA, G., and McCULLOCH, C.E. (1992). Variance Components. New York: John Wiley and Sons.

# Small Area Estimation Using Multilevel Models 

FERNANDO A.S. MOURA and DAVID HOLT ${ }^{1}$


#### Abstract

In this paper a general multilevel model framework is used to provide estimates for small areas using survey data. This class of models allows for variation between areas because of: (i) differences in the distributions of unit level variables between areas, (ii) differences in the distribution of area levcl variables between areas and (iii) area specific components of variance which make provision for additional local variation which cannot be explained by unit-level or area-level covariates. Small area estimators are derived for this multilevel model formulation and an approximation to the mean square error (MSE) of each small area estimate for this general class of mixed models is provided together with an estimator of this MSE. Both the approximations to the MSE and the estimator of MSE take into account three sources of variation: (i) the prediction MSE assuming that both the fixed and components of variance terms in the multilevel model are known, (ii) the additional component due to the fact that the fixed coefficients must be estimated, and (iii) the further component due to the fact that the components of variance in the model must be estimated. The proposed methods are estimated using a large data set as a basis for numerical investigation. The results confirm that the extra components of variance contained in multilevel models as well as small area covariates can improve small area estimates and that the MSE approximation and estimator are satisfactory.


KEY WORDS: Small area estimation; Mixed models; Multilevel models; EBLUE.

## 1. INTRODUCTION

The need for small area (and small domain) estimates from survey data has long been recognized. The difficulty with the production of such estimates is that for most, if not all, small areas, the sample size achieved by a survey designed for national purposes is too small for direct estimates to be made with acceptable precision. Early attempts to tackle this problem using methods such as synthetic estimation (Gonzalez 1973) involved the use of auxiliary information and the pooling of information across small areas. An excellent review and bibliography are given by Ghosh and Rao (1994).

Empirical studies show that such methods made too little provision for local variation and consequently the resulting small area estimates were shrunk too far towards a predicted mean. More recent approaches (e.g., Battese and Fuller 1981 and Battese, Harter and Fuller 1988) use some components of variance model, or equivalent, to provide for local variation. Empirical studies show the superiority of this approach (e.g., Prasad and Rao 1990).

This paper proposes a general multilevel model framework for small area estimation. This involves the potential to use auxiliary information at both the unit and small area level. In addition any of the regression parameters, rather than just the intercept as proposed by Battese and Fuller (1981), may be treated as varying randomly between small areas. The local variation is provided for by using differences between the means of unit level auxiliary variables, the small area level variables, and the various components of variance which allow variation between areas.

For this general model, the small area predictor is obtained. In addition, an approximation to the mean square error (MSE) of each separate small area prediction and an estimator of this MSE are developed.

The numerical study, based on a large data set from Brazil shows that such models may be useful for predicting small area estimates. The robustness of the approach to misspecification of the variance-covariance matrix of the small area random effects and misspecification of small area covariates are also investigated. Further numerical results demonstrate the success of the MSE approximation and its estimator.

## 2. THE MULTILEVEL MODEL FRAMEWORK

### 2.1 Introduction

We consider the following multilevel model for predicting the small area means:

$$
\begin{align*}
& Y_{i}=X_{i} \beta_{i}+\varepsilon_{i} \\
& \beta_{i}=Z_{i} \gamma+v_{i} \quad i=1, \ldots, m \tag{2.1}
\end{align*}
$$

where $Y_{i}$ is the vector of length $n_{i}$ for the characteristic of interest for the sample units in the $i$-th small area, $i=1, \ldots, m ; X_{i}$ is the matrix of explanatory variables at sample unit level; $Z_{i}$ is the design matrix of small area variables; $\gamma$ is the vector of length $q$ of fixed coefficients and $v_{i}=\left(v_{i 0}, \ldots, v_{i p}\right)^{T}$ is the vector of length $(p+1)$ of random effects for the $i$-th small area. We assume the

[^9]following about the distribution of the random vectors: (a) the $\nu_{i}$ are independent between small areas and have a joint distribution within each small area with $E\left(v_{i}\right)=0$ and (b) $V\left(v_{i}\right)=\Omega$ (b) The $\varepsilon_{i} s$ and $v_{i} s$ are independent and $V\left(\varepsilon_{i}\right)=\sigma^{2} I$.

For the whole population (2.1) applies with $n_{i}$ replaced by $N_{i}$, the small area population sizes.

The set of $m$ equations in.(2.1) can be concisely written by stacking them as

$$
\begin{equation*}
Y=X Z_{\gamma}+X v+\varepsilon . \tag{2.2}
\end{equation*}
$$

It is worth noting that the random intercept model (see section 2.3) can be regarded as a special case of the model (2.1) where $Z_{i}$ is equal to the identity matrix for each small area and $\Omega$ has all terms constrained to be zero excèpt the one corresponding to the variance of the intercept term. Other intermediate models exist, for instance, when $\Omega$ is diagonal so that the small area regression coefficients are random but uncorrelated between covariates.

Holt (in Ghosh 1994, page 82) observes that the advantage of the model (2.1) over other competitors is that it effectively integrates the use of unit level and area level covariates into a single model. Besides the use of extra random effects for the regression coefficients gives greater flexibility in situations where it is not appropriate to assume the same slope coefficients apply for all small areas.

### 2.2 Fixed and Component of Variance Parameter Estimates

The fixed and components of variance parameters in the model (2.1) are $\gamma$ and $\theta=\left([\operatorname{Vech}(\Omega)]^{T}, \sigma^{2}\right)^{T}$ respectively. Various methods for estimating these model parameters in the case of a general mixed linear model are available. Most of them, based on iterative algorithms, lead to the maximum likelihood estimator (MLE) or the restricted maximum likelihood estimator (RMLE) under certain regularity conditions.

Goldstein (1986) shows how consistent estimators can be obtained by applying iterative generalised least squares procedures (IGLS). He also proved its equivalence to the maximum likelihood estimator under normality. Later Goldstein (1989) proposed a slight modification of his algorithm (namely, restricted iterative generalised least squares (RIGLS)) which is equivalent to RMLE under normality. Unlike the IGLS estimates, the RIGLS estimation procedures provide unbiased estimates of the component of variance parameters by taking into account the loss in degrees of freedom resulting from estimating the fixed parameters.

This work is confined to the RIGLS approach as in Goldstein (1989). The RIGLS procedure is described in details in Appendix A.

### 2.3 The Estimator of the Small Area Mean

Assuming the model (2.1) and considering that the population size $N_{i}$ in the $i$-th small area is large, we can write the mean for the $i$-th small area as

$$
\begin{equation*}
\mu_{i}=\bar{X}_{i}^{T} Z_{i} \gamma+\bar{X}_{i}^{T} v_{i} \tag{2.3}
\end{equation*}
$$

where $\bar{X}_{i}$ is the $(p+1)$ population mean vector for the $i$-th small area.

An estimator of $\mu_{i}$ may be obtained by plugging the RIGLS estimators of $\gamma$ and $\theta$ in the respective terms of equation (2.3), where the predictor of the $i$-th small area random effect $v_{i}$ is given by $\hat{v}_{i}=\hat{\Omega} X_{i}^{T} \hat{V}_{i}^{-1}\left(Y_{i}-X_{i} Z_{i} \hat{\gamma}\right)$ where $\hat{V}_{i}^{-1}=\hat{\sigma}^{-2} I-\hat{\sigma}^{-4} X_{i} \hat{\Omega} \hat{G}_{i}^{-1} X_{i}^{T} \quad$ and $\quad \hat{G}_{i}^{-1}=$ $\left(I+\hat{\sigma}^{-2} X_{i}{ }^{i} X_{i} \hat{\Omega}\right)^{-1}$.

This estimator of $\mu_{i}$ is known as Empirical Best Linear Unbiased Estimator (EBLUE)

$$
\begin{equation*}
\hat{\mu}_{i}=\bar{X}_{i}^{T} Z_{i} \hat{\gamma}+\bar{X}_{i}^{T} \hat{v}_{i} . \tag{2.4}
\end{equation*}
$$

Battese et al., $(1981,1988)$ propose and apply a random intercept model to provide small area estimates. In this case, the Empirical Best Linear Unbiased Estimator is

$$
\hat{\mu}_{i(\mathrm{RN})}=\bar{X}_{i}^{T} \hat{\beta}+\hat{v}_{i 0} .
$$

We use the label (RI) to imply a random intercept model since only the intercept of each small area is random while the other components of $\beta$ remain fixed.

### 2.4 Approximation to the Mean Square Error (MSE)

Kackar and Harville (1984) show that, if $\hat{\theta}$ is a translation invariant estimator of $\theta$ and the random terms are normally distributed, the mean square error of a predictor of a linear combination of a fixed and random effect can be decomposed into two terms. The first one is due to the variability in estimating the fixed parameters when the components of variance are known, the second term comes from estimating the components of variance.

Since under normality the RIGLS estimator is equivalent to the RMLE estimator and the RMLE is translationinvariant, Kackar and Harville's (1984) results can be applied to the small area means estimators $\hat{\mu}_{i}, i=1, \ldots, m$ :

$$
\begin{equation*}
\operatorname{MSE}\left(\hat{\mu}_{i}\right)=E\left[\hat{\mu}_{i}-\mu_{i}\right]^{2}=E\left[\tilde{\mu}_{i}-\mu_{i}\right]^{2}+E\left[\hat{\mu}_{i}-\tilde{\mu}_{i}\right]^{2} \tag{2.5}
\end{equation*}
$$

where $\bar{\mu}_{i}$ is the BLUE of $\mu_{i}$.
The first term of (2.5), that is MSE[ $\left.\tilde{\mu}_{i}\right]$, can be obtained by direct calculation as

$$
\begin{align*}
& \operatorname{MSE}\left(\tilde{\mu}_{i}\right)=\bar{X}_{i}^{T}\left(G_{i}^{-1}\right)^{T} \Omega \bar{X}_{i}+ \\
& \sigma^{2} \bar{X}_{i}\left(G_{i}^{-1}\right)^{T} Z_{i}\left(\sum_{i=1}^{m} Z_{i}^{T} G_{i}^{-1} X_{i}^{T} X Z_{i}\right)^{-1} Z_{i}^{T} G_{i}^{-1} \bar{X}_{i} \tag{2.6}
\end{align*}
$$

where $G_{i}=I+\sigma^{2} X_{i}^{T} X_{i} \Omega$. Kackar and Harville (1984) point out that the second term of (2.6) is not tractable, except for special cases, and propose an approximation to
it. Prasad and Rao (1990) propose an approximation to this second term and work out the details of their approximation for three particular cases: the random intercept model, random regression coefficient model and the Fay-Herriot model. They also give some regularity conditions for their approximation to be of the second order, and prove that their MSE approximation for the Fay-Herriot model is of the second order. Nevertheless, it seems to be more difficult to give general conditions for more complex models such as model (2.1).

Applying Prasad and Rao's approach, an approximation to the second term of (2.5) is developed in Appendix B.

It is worth noting that the MSE approximation of ( $\hat{\mu}_{i}$ ) can be decomposed into three terms:

$$
\begin{equation*}
\operatorname{MSE}\left(\hat{\mu}_{i}\right) \approx T_{1}+T_{2}+T_{3} \tag{2.7}
\end{equation*}
$$

where $T_{1}$ and $T_{2}$ are respectively the first and the second term of equation (2.6) and $T_{3}$ is described in Appendix B.

The term $T_{1}$ is the variability of $\hat{\mu}_{i}$ when all parameters are known, the second term $T_{2}$ is due to estimating the fixed effects and the third term $T_{3}$ comes from estimating the components of variance.

When sampling fractions are not negligible, estimators of the small area means can be built in the spirit of the finite population approach by predicting specifically for the nonsampled units:

$$
\begin{equation*}
\hat{\mu}_{i}^{F}=f_{i} \bar{y}_{i}+\left(\bar{X}_{i}-f_{i} \bar{x}_{i}\right)^{T}\left(Z_{i} \hat{\gamma}+\hat{v}_{i}\right) \tag{2.8}
\end{equation*}
$$

where the superscript $F$ indicates that a correction for the finite population sampling fraction $f_{i}$ was used; $\bar{x}_{i}$ is the $(p+1)$ vector of sample means.

The $\operatorname{MSE}\left(\hat{\mu}_{i}^{F}\right)$ can be obtained by noting that

$$
\hat{\mu}_{i}^{F}-\bar{Y}_{i}=\left(1-f_{i}\right)\left[\left(\bar{X}_{i}^{C}\right)^{T}\left(Z_{i}(\hat{\gamma}-\gamma)+\hat{v}_{i}-v_{i}-\bar{\varepsilon}_{i}^{c}\right)\right]
$$

where $\bar{X}_{i}^{c}=\left(1-f_{i}\right)^{-1}\left(\bar{X}_{i}-f_{i} \bar{x}_{i}\right)$ and $\bar{\varepsilon}_{i}^{c}$ is the mean of $\varepsilon_{i j}$ for the non-sampled units in the $i$-th small area. Therefore

$$
\begin{equation*}
\operatorname{MSE}\left(\hat{\mu}_{i}^{F}\right)=\left(1-f_{i}\right)^{2}\left[\operatorname{MSE} \cdot\left(\hat{\mu}_{i}\right)+N_{i}^{-1}\left(1-f_{i}\right)^{-1} \sigma^{2}\right] \tag{2.9}
\end{equation*}
$$

${ }^{\text {where }}{ }^{C}$ MSE ${ }^{\text {* }} \hat{\mu}_{i}$ ) is the equation (2.7) with $\bar{X}_{i}$ replaced by $\bar{X}_{i}^{C}$.

### 2.5 Estimation of Mean Square Error

It is common practice to estimate the MSE of a linear combination of the fixed and random effects in a mixed model as in (2.1) by replacing estimates of the components of variance respectively in the expression of MSE. This estimator ignores the contribution to MSE due to estimating the components of variance parameters. Several studies (see for example Singh, Stukel and Pfeffermann 1998 or

Harville and Jeske 1992) argue that this procedure tends to underestimate the MSE. Prasad and Rao (1990) reported a simulation study which showed that the use of this "naive" estimator leads to severe downwards bias. They also showed for the Fay-Herriot model (a special case of the model (2.1)), using "truncated Henderson" estimates for the variance components, that

$$
\begin{aligned}
E\left(\hat{T}_{1}\right)=T_{1}-T_{3}+o\left(m^{-1}\right) ; E\left(\hat{T}_{2}\right) & =T_{2}+o\left(m^{-1}\right) ; \\
E\left(\hat{T}_{3}\right) & =T_{3}+o\left(m^{-1}\right) .
\end{aligned}
$$

Harville and Jeske (1992) establish some conditions for the unbiasedness of Prasad and Rao's mean square error estimator. However, considering the more general model (2.1), again it seems more difficult to give general conditions for which the order of bias of Prasad and Rao's estimator is $o\left(m^{-1}\right)$, especially if iterative procedures as RIGLS are used to obtain the parameter estimates.

Nevertheless, motivated by the simulation study summarised in Section 3.4 and an extensive simulation study described in Moura (1994), we propose to use an estimator similar to Prasad and Rao's for $\operatorname{MSE}\left(\hat{\mu}_{i}\right)$ :

$$
\begin{equation*}
\mathbf{M S E}=\hat{T}_{1}+\hat{T}_{2}+2 \hat{T}_{3} . \tag{2.10}
\end{equation*}
$$

Where $\hat{T}_{i}$ are obtained from (2.5) by replacing $\sigma^{2}$ and $\Omega$ by their respective RIGLS estimators.

From equation (2.9) we can also obtain an estimator for $\operatorname{MSE}\left(\hat{\mu}_{i}^{F}\right)$ as follows:

$$
\begin{equation*}
\operatorname{MSE}\left(\hat{\mu}_{i}^{F}\right)=\left(1-f_{i}\right)^{2}\left[\operatorname{MSE}{ }^{*}\left(\hat{\mu}_{i}\right)+N_{i}^{-1}\left(1-f_{i}\right)^{-1} \hat{a}^{2}\right] \tag{2.11}
\end{equation*}
$$

$\bar{W}^{c}$ where $\operatorname{MSE}{ }^{*}\left(\hat{\mu}_{i}\right)$ is the equation (2.10) with $\bar{X}_{i}$ replaced by $\bar{X}_{i}^{c}$.

## 3. A MODEL-BASED NUMERICAL INVESTIGATION

### 3.1 Comparison of the Estimators

In order to investigate the properties of alternative estimators, data was used from 38,740 households in the enumeration districts in one county in Brazil. The Head of Household's income was treated as the dependent variable. Two unit level independent variables were identified as the educational attainment of the Head of Household (ordinal scale of 0-5) and the number of rooms in the household (1-11+).

The assumed model is

$$
\begin{align*}
& Y_{i j}=\beta_{i 0}+\beta_{i 1} x_{1 i j}+\beta_{i 2} x_{2 i j}+\varepsilon_{i j} \quad i=1, \ldots, m ; j=1, \ldots, N_{i} \\
& \beta_{i 0}=\gamma_{00}+v_{i j} ; \beta_{i 1}=\gamma_{10}+v_{i 1} ; \beta_{i 2}=\gamma_{20}+v_{i 2} \tag{3.1}
\end{align*}
$$

where $x_{1}$ and $x_{2}$ respectively represent the number of rooms and the educational attainment of the head of the
household (centred about their respective population means).

The parameter values for the fit model and their respective standard errors are
$\gamma_{00}=8.456(0.108)$
$\gamma_{10}=1.223(0.046) \gamma_{20}=2.596(0.086)$
$\sigma_{00}=1.385(0.194) \quad \sigma_{01}=0.354(0.66) \quad \sigma_{02}=0.492(0.117)$
$\sigma_{12}=0.333(0.054) \sigma_{11}=0.234(0.35) \quad \sigma_{22}=0.926(0.124)$
$\sigma^{2}=47.74(0.345)$
To carry out numerical investigations within the model-based framework a simulation was carried out keeping the enumeration district identifiers and the values of the two explanatory variables ( $X$ ) fixed. Initially the area population means $\bar{X}_{1 i}$ and $\bar{X}_{2 i}$ were calculated for the whole data set and a randomly selected subsample of $10 \%$ of records from each small area was identified. This same subset was retained throughout the simulations (the Simulation subset).

The data generation for the simulations was carried out in two stages using a data generation model which was the General Model (G), the Diagonal Model (D), the Random Intercept Model (RI) as appropriate. In the first case the parameter values were taken from the estimates mentioned earlier. In the second case the off-diagonal terms were set to zero, in the third case only $\sigma_{00}=1.385$ was non-zero.

The first stage of the data simulation process was to generate the level 2 random terms (that is, the non-zero elements of $v_{i 0}$ and $v_{i 1}$ and $v_{i 2}$ ) depending on the choice of the data generation model. These random terms were Normally distributed (jointly Normal in the case of the General Data Generation Model and the Diagonal Data Generation Model). At this stage the expected value of the mean for the $i$-th area conditional on the area level random effects generated by the model $m_{1}=\mathrm{G}, \mathrm{D}, \mathrm{RI}$ in the $r$-th simulation could be obtained:

$$
\mu_{i m_{1}}^{(r)}=\beta_{0 i}^{(r)}+\beta_{1 i}^{(r)} \bar{X}_{1 i}+\beta_{2 i}^{(r)} \bar{X}_{2 i}
$$

At the second stage of the data simulation process, unit values $\left(Y_{i j}\right)$ were created for each of the data generation models. Having generated the data for the simulation subset under one of the data generation models, all three of the estimation models (G, D and RI) could be fitted to the simulated data to obtain parameter estimates and predictors for the small area means.

For each data generation model $m_{1}=\mathrm{G}, \mathrm{D}, \mathrm{RI}$ the whole simulation process was repeated $\mathrm{R}=5000$ times to yield a set of small area means $\mu_{i m_{1}}^{(r)}$ and predicted means $\hat{\mu}_{i m_{1}, m_{2}}^{(r)}, r=1, \ldots, R$ for each small ${ }_{1}$ area, $i, i=1, \ldots, m$ and for the three estimation models: $m_{2}=$ G, D, RI. For each small area and for data generated under model $m_{1}=G, D, R I$, the Mean Square Error (MSE) of the prediction process for each estimation model $m_{2}$ may be defined as

$$
\operatorname{MSE}\left[\hat{\mu}_{i m_{1}, m_{2}}\right]=R^{-1} \sum_{r=1}^{R}\left(\hat{\mu}_{i m_{1}, m_{2}}^{(r)}-\mu_{i m_{1}}^{(r)}\right)^{2}
$$

and the absolute relative error (ARE) by

$$
\operatorname{ARE}\left[\hat{\mu}_{i m_{1}, m_{2}}\right]=R^{-1} \sum_{r=1}^{R}\left|\hat{\mu}_{i m_{1}, m_{2}}^{(r)}-\mu_{i m_{1}}^{(r)}\right| / \mu_{i m_{1}}^{(r)}
$$

For comparative purposes we contrast the properties of each estimator with those of the estimator which is the same as the data generation model. Hence we define the Ratio of Mean Square Error (RMSE):

$$
\begin{aligned}
\mathrm{RMSE}_{\mathrm{m}_{2}, \mathrm{~m}_{1}} & = \\
& \left\{\sum_{i=1}^{m} \operatorname{MSE}\left[\hat{\mu}_{i m_{1}, m_{2}}\right]\right\} /\left\{\sum_{i=1}^{m} \operatorname{MSE}\left[\hat{\mu}_{i m_{1}, m_{1}}\right]\right\} \times 100
\end{aligned}
$$

and the Ratio of Absolute Relative Error (RARE):


It will be seen that when the data are generated from a simpler model (e.g., RI) the more complex estimation procedures do not suffer any appreciable worsening of efficiency or bias. On the other hand when the data are generated from a more complex model the simpler estimators have inferior properties. However the difference between the Diagonal and General estimators is much less than between these and the Random Intercept Estimator. From Table 1 one would conclude that it is worth introducing additional random coefficients of some kind, beyond the simple Random Intercept model assumptions, but not necessarily the full General Model.

Table 1
Ratios of Mean Square Errors and Ratios of Absolute Relative Errors (in parentheses) for the three Estimators and Three Data Generation Models

| Estimator | Data Generation Model |  |  |
| :--- | :---: | :---: | :---: |
|  | G | D | RI |
| General (G) | 100.0 | 101.8 | 101.2 |
|  | $(100.0)$ | $(100.9)$ | $(100.6)$ |
| Diagonal (D) | 108.8 | 100.0 | 100.2 |
|  | $(82.6)$ | $(100.0)$ | $(100.1)$ |
| R. Intercept (RI) | 131.9 | 109.1 | 100.0 |
|  | $(176.9)$ | $(105.6)$ | $(100.0)$ |

The summary measures in Table 1 are average properties over all small areas. A careful analysis of the MSE performance of the estimators for each small area shows that there is a modest increase in the MSE for the Diagonal Estimator
compared to the General Estimator for all areas, whereas for the Random Intercept estimator a relatively small number of areas exhibit a substantial increase in MSE. A similar pattern occurs between the Diagonal and Random Intercept estimator when the Diagonal Data Generation Model is used.

### 3.2 Introducing a Small-Area Level Covariate

In this section an attempt is made to investigate the impact on small area estimates of introducing an area covariate $Z$. Unfortunately for the data set used, it was not possible to identify a single contextual area level covariate which had a substantial effect on the multilevel models. Nevertheless, the number of cars per household in each small area was a useful covariate for the random coefficients for the individual level random slopes coefficients for "Room" and "Edu", but not for the random intercept term. This was observed after some preliminary model fit analysis on the real data. Although the "numbers of cars" was the best small area level covariate found to explain between area variation, it was not as powerful at the individual level as "Room" and "Edu", the individual level covariates chosen.

The model above with the small area covariate $Z$ can be written as

$$
\begin{align*}
& Y_{i j}=\beta_{i 0}+\beta_{i 1} x_{1 i j}+\beta_{i 2} x_{2 i j}+\varepsilon_{i j} \quad i=1, \ldots, m ; j=1, \ldots, N_{i} \\
& \beta_{i 0}=\gamma_{00}+v_{i 0} ; \beta_{i 1}=\gamma_{10}+\gamma_{11} z_{i}+v_{i 1} ; \beta_{i 2}=\gamma_{20}+\gamma_{21} z_{i}+v_{i 2} . \tag{3.2}
\end{align*}
$$

The small area random effects were assumed uncorrelated in order to avoid convergence failure in the simulation study.

Table 2 reports the parameter estimates and their respective standard errors obtained by fitting the Diagonal Model with the $Z$ covariate (3.2) and without the $Z$ covariate (2.1). It is worth noting the significant reduction of all the components of variance estimates, except $\hat{\sigma}_{00}$ and $\hat{\sigma}^{2}$, after introducing the explanatory area covariate $Z$.

In order to investigate the effect of misspecification of the $Z$ variable, the model based simulation procedure described in section 3.1 was applied to the two models above, where the data generation was done according to the parameters presented in Table 2. Table 3 summarises the simulation results.

It is worth noting that in both cases there is a significant loss of efficiency by using an unsuitable estimator. It can also be seen from an individual analysis of MSE for each small area that a considerable gain in efficiency is achieved with the introduction of a small area covariate $Z$ over the diagonal model. For many small areas the MSE of the Diagonal with $Z$ is significantly less than the MSE of the corresponding estimator without $Z$. Even for those few areas in which the MSE of the Diagonal with $Z$ is unchanged or even slightly increased by the introduction of $Z$, the difference is not appreciable.

Table 2
Parameter Estimates and Standard Errors for General Model with Area Level Covariate: Demographic Data

| Parameter | Diagonal <br> Model with $Z$ | Diagonal Model |
| :---: | :---: | :---: |
| $\gamma_{00}$ | $8.442(0.112)$ | $8.688(0.136)$ |
| $\gamma_{10}$ | $0.451(0.179)$ | $1.321(0.085)$ |
| $\gamma_{20}$ | $0.744(0.272)$ | $2.636(0.134)$ |
| $\gamma_{11}$ | $3.779(0.507)$ | - |
| $\gamma_{22}$ | $1.659(0.323)$ | - |
| $\sigma_{00}$ | $0.745(0.308)$ | $0.637(0.303)$ |
| $\sigma_{11}$ | $0.237(0.083)$ | $0.471(0.116)$ |
| $\sigma_{22}$ | $0.700(0.197)$ | $1.472(0.295)$ |
| $\sigma^{2}$ | $44.00(1.05)$ | $44.01(1.05)$ |

Table 3
Ratios of Mean Square Errors and Ratios of Absolute Relative Errors (in parentheses) for the Diagonal and the Diagonal with $Z$ Estimators Under the Two Respective Data Generation Models

| Estimator | Data Generation Model |  |
| :--- | :---: | :---: |
|  | Diagonal | Diagonal with $Z$ |
| Diagonal | 100.0 | 110.3 |
|  | $(100.0)$ | $(125.4)$ |
| Diagonal with $Z$ | 126.2 | 100.0 |
|  | $(107.5)$ | $(100.0)$ |

### 3.3 Comparisons with Regression Estimator

One essential advantage of the multilevel models over regression models is to recognize that groups (here the small areas) share common features; they are not completely independent as could be assumed, for example by using separate linear regression model for each small area. Nevertheless, the relatively small intraclass correlation observed for the data set used plus the fact that each small area has on average 28 units, could make one think that in this case the use of the multilevel model would not result in great improvement in the small area estimators. However, it is gratifying to know that even in these circumstances the multilevel model small area estimator performs on average better than the synthetic separate regression estimator, under either the multilevel model or even under the regression model. Table 4 illustrates this finding.

The multilevel data generation model used was the General one with the parameters given in section 3.1. The parameters used in the data generation regression model were obtained by fitting a separate regression for each small area.

It can be seen from Table 4 that the Separate Regression estimator which does not explore the difference of small areas through small area random effects shows substantial loss of efficiency when compared with the General estimator.

Table 4
Ratios of Mean Square Errors and Ratios of Absolute Relative Errors (in parentheses) for the General and the Separate Regression Estimators Under the Two Respective Data Generation Models

| Estimator | Data Generation Model |  |
| :--- | :---: | :---: |
|  | General | Separate Regression |
| General | 100.0 | 88.1 |
|  | $(100.0)$ | $(83.1)$ |
| Separated | 247.6 | 100.0 |
| Regression | $(154.7)$ | $(100.0)$ |

Figure 1 illustrates this fact by showing a plot of the ratio of mean square error between the General estimator and the Separate Regression estimator for each small area. To demonstrate the effect of the small area sample size on the efficiency, the ratio of the MSEs is plotted against the sample size for each small area. It is clear from Figure 1 that the gain in efficiency tends to decrease as the sample size increases.


Figure 1. Model-based efficiencies of the general estimator compared with the separate regression estimator for each small area

### 3.4 An Evaluation of the MSE Approximation and the MSE Estimator

From the simulation results we may investigate the properties of the MSE approximation (2.7). If we consider the General estimator when the General Data Generation model is used the MSE approximation appears to be very good. The average underestimation of the MSE approximation was $0.31 \%$ of the MSE value with a range from the largest underestimate of $5.4 \%$ of the MSE value through to a largest overestimate of $4.8 \%$ of the MSE value. For the situation considered here $T_{1}$ contributed on average $94.6 \%$ of the total variation and $T_{3}$ a further $4.3 \%$. Given the large component of variance due to $\sigma^{2}$, these results are not unexpected. For individual areas the component $T_{1}$ varied between $87.4 \%$ to $99.1 \%$ of the total and $T_{3}$ varied between
$0.7 \%$ and $10.5 \%$ of the total. The component $T_{2}$ never contributed more than $2.2 \%$ of the total MSE for any area.

We also investigated the performance of the MSE estimator represented by equation (2.10) against the "naive" estimator of the MSE, which does not consider the last term of (2.10). The average Root Mean Squared Error of the proposed MSE estimator is $17.5 \%$ ranging from $4.7 \%$ to $32.3 \%$, while for the naive estimator the average is $20.9 \%$ ranging from $5.2 \%$ to $47.5 \%$. The MSE estimator is on average unbiased while the naive MSE estimator underestimates the MSE on average by $9.1 \%$, its relative bias ranging from $-23.5 \%$ to $-0.9 \%$. Our results agree with others, see Singh, Stukel and Pfeffermann (1998) and Prasad and Rao (1990), which show that the naive estimator can exhibit severe bias.

## 4. DISCUSSION

Prasad and Rao (1990) and Battese et al., (1981, 1988) have demonstrated that models which include small area specific components of variance can provide greatly improved small area estimators. Some of the numerical results in this paper show that within the model-based simulation framework even better estimators can be obtained by allowing the small area slopes as well as the intercept to be random.

The overall conclusions from this investigation for this set of parameter values are that: a component of variance model more complex than the Random Intercept estimator is beneficial; overspecification of the model (e.g., using the General estimator with data generated under the Random Intercept Model) does not lead to serious loss of efficiency; the use of small area covariates can also improve the small area estimates; and the use of multilevel models should be preferred rather than the Separate Regression Model. The simulation study confirms that the MSE approximation appears to be precise and the MSE estimation is approximately unbiased, reflecting the variation in MSE between areas, but further theoretical investigation about the exact order of the approximation should be done.

Clearly model fitting and diagnostics are crucial. If we apply a general mixed model in circumstances where it is only a poor fit to the data, then the results may be disappointing. Considerably more investigation is needed to understand what characteristics of specific small areas are likely to provide efficiency gains if general mixed models are used rather than simpler models.

## ACKNOWLEDGEMENTS

We would like to thank the referees and the Editor for their helpful comments on the earlier version of this paper.

## APPENDIX A: RESTRICTED ITERATIVE GENERALIZED PROCEDURE

The generalized least squares estimator of $\gamma$ in the model (2.1) is given by

$$
\begin{align*}
& \tilde{\gamma}=\left(Z^{T} X^{T} V^{-1} X Z\right)^{-1}\left(Z^{r} X^{T} V^{-1} Y\right)= \\
& \left(\sum_{i=1}^{m} Z_{i}^{T} X_{i}^{T} V_{i}^{-1} X_{i} Z_{i}\right)^{-1}\left(\sum_{i=1}^{m} Z_{i}^{T} X_{i}^{T} V_{i}^{-1} Y_{i}\right) \tag{A.1}
\end{align*}
$$

where $V=\operatorname{Diag}\left(V_{1}, \ldots, V_{m}\right)$ and $V_{i}=\sigma^{2} I+X^{T} \Omega X$ is the covariance matrix of $Y_{i}, i=1, \ldots, m$.

However, $V$ is assumed to be a function of unknown parameters, thus $\gamma$ cannot be estimated using (A.1). On the other hand, if $\gamma$ is known then

$$
\begin{equation*}
Y^{*}=\operatorname{vech}\left[(Y-X Z y)(Y-X Z \gamma)^{T}\right] \tag{A.2}
\end{equation*}
$$

is an unbiased estimator of $\operatorname{vech}(V)$. Furthermore vech $(V)$ is a linear function of $\theta$. Then we can consider the following linear model:

$$
\begin{equation*}
Y^{*}=F \theta+\xi \tag{A.3}
\end{equation*}
$$

Where $F=\partial \operatorname{vech}(V) / \partial \theta$ and $\xi$ is a random variable with mean $O=(0, \ldots, 0)$ and the covariance of $\xi$ is given by $V_{\xi}=2 \varphi_{n}(V \otimes V) \varphi_{n}^{T}$. The matrix $\varphi_{n}$ is any linear transformation of $\operatorname{vec}(A)$ into $\operatorname{vech}(A)$, and $A$ is any $n \times n$ matrix such that $\operatorname{vech}(A)=\varphi_{n} \operatorname{vec}(A)$, see Fuller(1987) for further details. Then, assuming that $F$ has full rank and $V_{\xi}$ is known and non-singular, it may be shown that the Generalized Least Square Estimator of $\theta$ is given by

$$
\begin{equation*}
\bar{\theta}_{a}=\operatorname{cov}\left(\bar{\theta}_{a}\right)\left(\frac{\partial \operatorname{vec}(V)}{\partial \theta}\right)^{T}\left(\frac{1}{2} V^{-1} \otimes V^{-1}\right) \operatorname{vec}\left(\tilde{Y} \tilde{Y}^{T}\right) \tag{A.4}
\end{equation*}
$$

where

$$
\operatorname{cov}\left(\tilde{\theta}_{a}\right)=\left[\left(\frac{\partial \mathrm{vec}(V)}{\partial \theta}\right)^{T}\left(\frac{1}{2} V^{-1} \otimes V^{-1}\right)\left(\frac{\partial \mathrm{vec}(V)}{\partial \theta}\right)\right]^{-1}
$$

and

$$
\bar{Y}=Y-X Z \gamma .
$$

Note that $\ddot{\theta}_{a}$ depends on $\theta$ and $\gamma$, so both may be iteratively estimated. The IGLS procedure starts with an initial estimate of $V$ (that is, setting initial values of $\theta$ ) which produces an estimate of $\gamma$. Hence replacing the initial estimate of $V$ together with the estimate of $\gamma$ in (A.1) provides an improved estimate of $\theta$. In most cases convergence is achieved after a few iterations between equations (A.1) and (A.4), although it is not always guaranteed.

The RIGLS approach is based on the fact that if $\gamma$ is estimated by using generalised least squares with $V$ known then

$$
E\left[(Y-X Z \hat{\gamma})(Y-X Z \hat{\gamma})^{T}\right]=V-X Z\left(Z^{T} X^{T} V^{-1} X Z\right)^{-1} Z^{T} X^{T}
$$

The equation above suggests that we use

$$
\begin{equation*}
(Y-X Z \gamma)(Y-X Z \gamma)^{T}+X Z\left(Z^{T} X^{T} V^{-1} X Z\right)^{-1} Z^{T} X^{T} \tag{A.5}
\end{equation*}
$$

instead of $(Y-X Z \gamma)(Y-X Z \gamma)^{T}$ at each iteration cycle described above in order to obtain an approximately unbiased estimator of $V$ and consequently of $\theta$.

As pointed out by Goldstein $(1986,1989)$, if we start with a consistent estimate of $\gamma$, say the ordinary least squares estimator, then the final estimates will be consistent providing finite fourth moments exist.

It is worth noting that it is possible for the above procedure to yield negative estimates of variances. This problem can be avoided by imposing constraints at each iteration. For further details on this issue see Goldstein (1986).

## APPENDIX B: AN APPROXIMATION TO $E\left[\hat{\mu}_{i}-\tilde{\mu}_{i}\right]^{2}$

Prasad and Rao (1990), based on Kachar and Harville (1984), developed a second order approximation to the second term of ( 2.5 ) under some regularity conditions:

$$
\begin{align*}
E\left[\hat{\mu}_{i}-\bar{\mu}_{i}\right]^{2} & \approx T_{3}= \\
& \operatorname{tr}\left[\left(\frac{\partial d_{i}}{\partial \theta}\right) V\left(\frac{\partial d_{i}}{\partial \theta}\right)^{T} E\left((\hat{\theta}-\theta)(\hat{\theta}-\theta)^{r}\right]\right] \tag{B.1}
\end{align*}
$$

where, for the model (2.1), $d_{i}=\bar{X}_{i}^{T} K_{i}(I \otimes \Omega) X^{T} V^{-1}$, $K_{i}=[0, \ldots, I, \ldots 0]$, is the $(p+1) \times(p+1) m$ matrix with the identity matrix $I$ of order $p+1$ in the i -th position and 0 as the null matrix of order $p+1$, and $\hat{\theta}$ is any translationinvariant estimator of $\theta=\left(\theta_{1}, \ldots, \theta_{s}\right)$ where $\theta_{s}=\sigma^{2}$ and $\theta_{k} ; k=1, \ldots, s-1$ are the distinct elements of $\Omega$. Goldstein (1989) proves that under normality of the random terms of model (2.1), the RIGLS estimator of $\theta$ is equivalent to the Restricted Maximum Likelihood Estimator (RMLE), which is translation invariant.

Let us approximate $E\left[(\hat{\theta}-\theta)(\hat{\theta}-\theta)^{\tau}\right]$ to the asymptotic covariance matrix of the RMLE estimator ( $B$ ). The jk -th element of $B^{-1}$ is given by(see Harville (1977))

$$
b_{j k}^{*}=T r\left(\sum_{i=1}^{m} P_{i} \frac{\partial V}{\partial \theta_{j}} P_{i} \frac{\partial V}{\partial \theta_{K}}\right)
$$

for $j$ and $k=1, \ldots, s$ where $P_{i}=V_{i}^{-1}-V_{i}^{-1} X_{i} Z_{i}$ ( $\sum_{i=1}^{m} Z_{i}^{T} X_{i}^{T} V_{i}^{-1} X_{i} Z_{i}$ ) $\quad Z_{i}^{T} X_{i}^{T} V_{i}^{-1}$. Let $b_{j, k}$ be $j k$-th element of $B$. After some matrix algebra, it can be shown
that

$$
\begin{gather*}
T_{3}=\bar{X}_{i}^{T}\left(G_{i}^{-1}\right)^{T}\left(\sum_{j=1}^{s-1} \sum_{k=1}^{s-1} b_{j k} \Delta_{j} C_{i} \Delta_{k}^{T}\right) G_{i}^{-1} \bar{X}_{i}- \\
2 \bar{X}_{i}^{T}\left(G_{i}^{-1}\right)^{T}\left(\sum_{j=1}^{s-1} b_{j, s} \Delta_{j}\right) R_{i} \Omega \bar{X}_{i}+b_{s s} \bar{X}_{i}^{T} \Omega S_{i} \Omega \bar{X}_{i} \tag{B.2}
\end{gather*}
$$

where $C_{i}=\sigma^{-2} G_{i}^{-1} X_{i}^{T} X_{i} ; R_{i}=\sigma^{-4} G_{i}^{-2} X_{i}^{T} X_{i}$; $S_{i}=\sigma^{-6} G_{i}^{-3} X_{i}^{T} X_{i}$; and

$$
\Delta_{k}=\frac{\partial \Omega}{\partial \Omega_{k}} k=1, \ldots, s-1
$$

is the s-1 square derivative matrix with respect to $\theta_{k} ; k=1, \ldots, s-1$.

## REFERENCES

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

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

FULLER, W.A. (1987). Measurement Error Models. Chichester: John Wiley.
GOLDSTEIN, H. (1986). Multilevel mixed linear model analysis using iterative generalised least squares estimation. Biometrika, 73, 43-56.

GOLDSTEIN, H. (1989). Restricted unbiased iterative generalised least squares estimation. Biometrika, 76, 622-623.

GONZALES, M.E. (1973). Use and evaluation of synthetic estimates. Proceedings of the Social Stalistics Section, American Statistical Association, 33-36.
GHOSH, M., and RAO, J.N.K. (1994). Small area estimation: an appraisal. Statistical.Science, 9, 55-93.
HARVILLE, D.A. (1977). Maximum likelihood approach to variance component estimation and related problems. Journal of the American Statistical Association, 72, 320-340.

HARVILLE, D.A., and JESKE, D.R. (1992). Mean squared error of estimation on prediction under a general linear model. Journal of the American Statistical Association, 87, 724-731.
HENDERSON, C.R. (1975). Best linear unbiased estimation and prediction under a selection model. Biometrics, 31, 423-447.
HOLT, D., and MOURA, F. (1993a). Mixed models for making small area estimates. In: Small Area Statistics and Survey Design, (G. Kalton, J. Kordos, and R. Platek, Eds.) 1, 221-231. Warsaw: Central Statistical Office.
HOLT, D., and MOURA, F. (1993b). Small area estimation using multilevel models. Proceedings of the Section on Survey Research Methods, American Statistical Association, 21-31.
KACKAR, R.N., and HARVILLE, D.A. (1984). Approximations for standard errors of estimators of fixed and random effects in mixed linear models. Journal of the American Statistical Association, 79, 853-862.
LONGFORD, N. (1987). A fast scoring algorithm for maximum likelihood estimation in unbalanced mixed models with nested effects. Biometrika, 79, 817-827.
MOURA, F.A.S. (1994). Small Area Estimation Using Multilevel Models. University of Southampton. Unpublished Ph. D. Thesis
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.

SINGH, A.C., STUKEL, D., and PFEFFERMANN, D. (1998). Bayesian versus frequentist measures of error in small area estimation. Journal of the Royal Statistical Society, Series B, 377-396.

# Composite Estimation of Drug Prevalences for Sub-State Areas 

# MANAS CHATTOPADHYAY, PARTHA LAHIRI, MICHAEL LARSEN and JOHN REIMNITZ ${ }^{1}$ 


#### Abstract

The Gallup Organization has been conducting household surveys to study state-wide prevalences of alcohol and drug (e.g., cocaine, marijuana, etc.) use. Traditional design-based survey estimates of use and dependence for counties and select demographic groups have unacceptably large standard errors because sample sizes in sub-state groups are too small. Synthetic estimation incorporates demographic information and social indicators in estimates of prevalence through an implicit regression model. Synthetic estimates tend to have smaller variances than design-based estimates, but can be very homogeneous across counties when auxiliary variables are homogeneous. Composite estimates for small areas are weighted averages of design-based survey estimates and synthetic estimates. A second problem generally not encountered at the state level but present for sub-state areas and groups concerns estimating standard errors of estimated prevalences that are close to zero. This difficulty affects not only telephone household survey estimates, but also composite estimates. A hierarchical model is proposed to address this problem. Empirical Bayes composite estimators, which incorporate survey weights, of prevalences and jackknife estimators of their mean squared errors are presented and illustrated.


KEY WORDS: Alcohol abuse; Drug abuse; Empirical Bayes; Jackknife; Mean squared error; Small area estimation; Synthetic estimation.

## 1. INTRODUCTION

The Gallup Organization has been conducting a series of household surveys for different states to study state-wide prevalences of the use of alcohol and drugs (e.g., cocaine, marijuana) among civilian, non-institutionalized adults and adolescents. The common goal of these surveys is to estimate the use and dependence prevalences for alcohol and drugs and, on that basis, to project the treatment needs of dependent users. For planning and resource allocation, states need precise estimates of prevalences for certain subgroups of the target population. For example, it is of interest to estimate prevalences for sub-state planning regions and counties in demographic subpopulations (e.g., older white males).

Traditional design-based procedures to estimate use and dependence for subpopulations have two drawbacks. First, if the traditional design-based survey estimate for a subgroup is positive, but sample size is small, then the corresponding standard error is unacceptably large. Second, since the problem is to estimate the proportion of a rare event, it is possible that the design-based procedure produces an estimate of zero and standard error estimation formulas for a particular subgroup, if applied, would give a false impression of the true underlying variability.

To improve on the traditional design-based estimators, one can use certain supplementary information usually available from administrative records in conjunction with the telephone survey data. This generally is done by using either implicit or explicit models that "borrow strength" or incorporate additional information that relates the various
groups, counties, and planning regions to one another. The method proposed here combines information across counties in order to deal with problem of zero estimates in some counties. It is derived from a model that bounds the proportions away from 1 , which is reasonable in an application with proportions expected to be very small, and estimates parameters using empirical Bayes methods. The procedure also incorporates the survey sampling weights in estimation.

For a detailed account of small-area estimation methods, see Ghosh and Rao (1994). Other recent references can be found in Farrell, MacGibbon and Tomberlin (1997) and Malec, Sedransk, Moriarity and Leclere (1997). Farrell et al., (1997) propose estimating small-area proportions with empirical Bayes procedures. They model the proportions via a logistic regression that relates expected proportions to respondent variables and includes random effects for the small areas. Malec et al., (1997) use hierarchical Bayes models. They use logistic regression models to relate individual characteristics to probabilities of an outcome and then use a linear regression model to relate coefficients across small areas. Most existing methods, including those of Farrell et al., (1997) and Malec et al., (1997) do not directly use survey sampling weights in estimation.

The survey design used by Gallup is described in section 2. In section 3, notations used in the paper are introduced. A direct design-based estimator and two synthetic estimators are presented in section 4. In section 5, several composite estimators of prevalences of alcohol and drug use and dependence are given. In this section, certain

[^10]empirical Bayes estimators and jackknife estimators of their mean squared errors (MSE) are proposed. In section 6, estimators presented in sections 4 and 5 are applied to a data set from a particular state. The focus of the analysis in this study is taken to be county level estimates. Sample size planning considerations originally were concerned with larger sub-state planning areas.

## 2. SURVEY

For sampling purposes, the state is divided into a few planning regions and samples are collected independently for each planning region using a truncated stratified random digit dialing (RDD) method of Casady and Lepkowski (1993). This design stratifies the Bellcore (BCR) frame into two strata: a high density stratum consisting of $100-$ banks with one or more listed residential numbers and a low density stratum consisting of all the remaining numbers in the BCR frame. About 52 percent of the numbers in the high-density stratum are estimated to be working residential numbers whereas in the low-density stratum, the corresponding percentage is only about 2 percent. The CasadyLepkowski procedure exploits the significant difference in the cost of sampling between the two strata by optimally determining the sample size in each stratum. In the truncated version of the procedure, sampling is done only from the high-density stratum.

Sample size in the original study was determined in order to estimate statewide prevalence with a desired degree of accuracy. Sample sizes were allocated to the planning regions using an optimal allocation scheme. Data on drug treatment admissions for the adult population in each county were used to compute the index prevalence (rate of admissions) percent in every planning region. These indices were then used to calculate the optimum sample size for each planning region. As a result of optimal allocation, relatively larger sample sizes were allocated to planning regions with higher index prevalences. The optimal allocation also minimizes the variance of the estimators. Gallup also oversampled the $18-45$ age group by planning region, because it is the age group with relatively higher rates of illicit drug use. Due to optimal allocation (which may be disproportional), the age oversampling and the complex design, weighting was needed to compute estimates from the sample data. The necessary weights, commonly known as sampling weights, were computed using current estimates of the population based on census data.

Due to budgetary constraints, it is not possible to increase sample size for all sub-state regions and groups in order to achieve the desired accuracy. To estimate alcohol and drug prevalences, we consider empirical Bayes procedures (see Efron and Morris 1973, Fay and Herriot 1979, Ghosh and Lahiri 1987, among others) to improve on usual design-based estimates of drug prevalences by taking advantage of demographic measurements and social indicator data.

Other variables that possibly are related to use and dependence prevalence by county and that are available from Census include the percent of population that is over 65 , under 30 , white, male, married, and renters. Local governments can provide data by county on social indicators, such as DUI (Driving Under the Influence) rate, mortality rate, per capita liquor licenses, and drug and alcohol treatment admission rates. The more closely auxiliary variables relate to use and dependence prevalence, the more likely it is that methods that "borrow strength" across areas and groups, such as the empirical Bayes methods presented here, can be employed to meet the desired accuracy levels for sub-state areas.

## 3. NOTATIONS

Let $n_{i}$ be the sample size allocated to the $i$-th planning region, $i=1, \ldots, l\left(n=\sum_{i=1}^{I} n_{i}\right)$. Samples are drawn independently in each planning region using RDD telephone surveys. After the sample is observed, suppose each region is post-stratified into $K$ demographic groups. These groups are formed by cross-classifying gender (Male, Female) and age ( $18-24,25-44,45-64,65+$ ), resulting in $K=2 \times 4=8$ groups. Suppose there are $J_{i}$ counties in the $i$-th planning region ( $i=1, \ldots, I$ ) and $n_{i j k}$ observations within the $k$-th demographic group in the $j$-th county belonging to the $i$-th planning region ( $i=1, \ldots, I ; j=1, \ldots, J_{i} ; k=1, \ldots, K$ ). Since typically $n_{i j k}$ is small, there is a good chance that some of the $k$ demographic groups are not represented in a particular county. Let $S_{i j}$ be the set of demographic groups in the $j$-th county within the $i$-th stratum ( $i=1, \ldots, I ; j=1$, ..., $J_{i}$ ) for which individuals have completed surveys.

Let $y_{i j k l}$ be the $l$-th observation ( 0 or 1) for the $k$-th demographic group in the $j$-th county belonging to the $i$-th planning area ( $i=1, \ldots, I ; j=1, \ldots, J_{i} ; k \in S_{i j} ; l=1, \ldots, n_{i j k}$ ). Let $w_{i j k l}$ be the corresponding sampling weight available from the survey. The goal is to estimate $\pi_{i j}$, the true prevalence of substance use or dependence for the $j$-th county within $i$-th planning area ( $i=1, \ldots, I ; j=1, \ldots, J_{i}$ ).

## 4. DIRECT SURVEY ESTIMATOR AND SYNTHETIC ESTIMATORS

The direct sample survey estimator of $\pi_{i j}$ is given by

$$
\pi_{\pi_{i j}} D=\frac{\sum_{k \in S_{i j}} \sum_{l=1}^{n_{i j k}} w_{i j k l} y_{i j k l}}{\sum_{k \in S_{i j}} \sum_{l=1}^{n_{i j k}} w_{i j k l}}
$$

The sample size available from a county could be very small (sometimes as small as 3 or 4). Thus, the estimator is highly unreliable. Other direct survey estimators are defined similarly. For example, the direct survey estimator
of $\pi_{i k}$, the true prevalence in the $k$-th demographic group in the $i$-th planning region is

$$
\widehat{\pi}_{i k}^{D}=\frac{\sum_{j: k \in S_{i j}} \sum_{l=1}^{n_{i k}} w_{i j k l} y_{i j k l}}{\sum_{j: k \in S_{i j}} \sum_{l=1}^{n_{i k}} w_{i j k l}}
$$

where the notation $j: k \in S_{i j}$ means that the summation is over counties $j$ in which demographic group $k$ is observed.

Additional problems arise when estimating the proportion of rare events. It is quite likely that all observations in a county may be zero, resulting in a zero estimate for a county. If usual estimates of standard error were applied, an estimated zero standard error of the estimate would give a false impression of the uncertainty of the estimate. Thus, it is very important to improve on the direct survey estimator.

Synthetic estimators borrow strength from related counties through implicit modeling of supplementary data from the U.S. Census Bureau along with the telephone survey data. A synthetic estimator, which has been used in the past to estimate alcohol prevalence at the county level, is given by

$$
\widehat{\pi}_{i j}^{S 1}=\sum_{k=1}^{K} a_{i j k}{\widehat{\pi_{k}}}^{D}
$$

where $\widehat{\pi}_{k}^{D}$ is the statewide direct survey estimator of prevalence of alcohol for the $k$-th demographic group and $a_{i j k}$ is the proportion of individuals belonging to the $k$-th demographic group in the $j$-th county within the $i$-th planning area ( $i=1, \ldots, I ; j=1, \ldots, J_{i} ; k=1, \ldots, K$ ). The value $a_{i j k}$ is available from current census estimates. For the household survey reported in this paper, the $a_{i j k}$ values were obtained from database vendors like Claritas Data Services of Ithaca, New York. Based on latest available census data, the $a_{i j k}$ values are typically estimated using projection models. In practice, therefore, the $a_{i j k}$ values are not true proportions but are current census estimates of reasonable precision. Outdated or inaccurate $a_{i j k}$ values cause the estimators using them to be biased. If population projections are used to calculate poststratification weighting adjustments in the survey, the direct survey estimator also suffers from this source of bias. It is beyond the scope of this paper to study the impact of alternate population projections. In proposing $\widehat{\pi}_{i j}{ }^{s 1}$, it is implicitly assumed that the prevalences for alcohol and drug use for the $k$-th group in all the counties are the same (or nearly the same).

A less restrictive synthetic estimator of prevalence of alcohol and drug use is given by

$$
\widehat{\pi_{i j}} s_{2}=\sum_{k=1}^{K} a_{i j k} \widehat{\pi}_{i k}^{D}
$$

where $\widehat{\pi} i k_{D}^{D}$ is a direct survey estimator of $\pi_{i k}$, the prevalence for alcohol or drug use for the $k$-th demographic group in the $i$-th planning region. It is implicitly assumed that the prevalences in the $k$-th group are the same (or nearly the same) for all the counties in a planning region. This assumption is more "regional" or less restrictive than the one made in proposing $\widetilde{\pi} i j^{s 1}$. A similar direct survey estimator $\widehat{\pi}_{i j k} D$ for the $k$-th demographic group within a specific county $j$ in region $i$ may be defined by restricting the sample to county $j$ only. As compared to $\widehat{\pi} i j k_{D}$, the estimator ${\widehat{\pi_{i k}}}^{D}$ will have relatively lower variance although it may have some bias since it does not distinguish the counties. $\widehat{\pi}_{i j k}{ }^{D}$, on the other hand, may be based on a very small sample size and hence may be significantly less reliable in terms of its variability.

The above synthetic estimators achieve reductions in variances at the cost of increasing bias. The synthetic estimators distinguish counties only through an indirect variable $a_{i j k}$ obtained from the census, whereas the direct estimator treats each county separately.

## 5. COMPOSITE ESTIMATORS OF $\pi_{i j}$ USING TELEPHONE SURVEY AND CENSUS DATA

A compromise between a direct survey estimator and a synthetic estimator is a composite estimator. A number of different composite estimators are proposed here based on the following identity:

$$
\pi_{i j}=\sum_{k \in S_{i j}} a_{i j k} \pi_{i j k}+\sum_{k \in S_{i j}} a_{i j k} \pi_{i j k}
$$

where $\pi_{i j k}$ is the prevalence for alcohol and drug use and $a_{i j k}$, as defined above, is the proportion of individuals belonging to the $k$-th demographic group in the $j$-th county within the $i$-th planning area ( $i=1, \ldots, j ; j=1, \ldots, J_{i}$; $k=1, \ldots, K$ ).

A simple composite estimator of $\pi_{i j}$ is obtained when,
 estimator of $\pi_{i j k}$, and for $k \notin S_{i j}, \pi_{i j k}$ is estimated by $\widetilde{\pi}_{i j k}{ }^{D}$. The estimator is then given by

$$
{\widehat{\pi_{i j}}}^{c}=\sum_{k \in S_{i j}} a_{i j k}{\widehat{\pi_{i j k}}}^{D}+\sum_{k \in S_{i j}} a_{i j k}{\widehat{\pi_{i k}}}^{D} .
$$

In the above formula, $\pi_{i j k}\left(k \in S_{i j}\right)$ is estimated using a small sample and thus there is the possibility for improving on ${\widehat{\pi_{i j k}}}^{D}$ (and, hence, on ${\widehat{\pi_{i j}}}^{c}$ ) by borrowing strength from relevant resources. To this end, an empirical Bayes estimate of $\pi_{i j}$ is proposed based on the following model.

## Model

1. Given the $\pi_{i j k}$ 's, the $y_{i j k l}$ 's are uncorrelated with one another with $E\left(y_{i j k l} \mid \pi_{i j k}\right)=\pi_{i j k}$ and $\operatorname{Var}\left(y_{i j k l} \mid \pi_{i j k}\right)=\pi_{i j k}$ ( $1-\pi_{i j k}$ ) for $i=1, \ldots, l ; j=1, \ldots, J_{i} ; k=1, \ldots, K ; l=1$, $\ldots, n_{i j k}$.
2. The $\pi_{i j k}$ 's are uncorrelated with $\mathrm{E}\left(\pi_{i j k}\right)=\mu_{i k}$; $\operatorname{Var}\left(\pi_{i j k}\right)=d \mu_{i k}^{2}\left(i=1, \ldots, I ; j=1, \ldots, J_{i} ; k=1, \ldots, K\right)$. If $\pi_{i j k} \sim \operatorname{Uniform}\left(0,2 \mu_{i k}\right)$, then in statement (2) $d=1 / 3$. Thus, unlike the implicit assumption made in the synthetic estimator $\widehat{\pi}_{i j}^{s 2}\left(i . e ., \pi_{i j k}=\mu_{i k}\right)$, some variability of proportions across counties within a region for a particular demographic group is allowed.

The first assumption of the model implies that given $\pi_{i j k}$, the $\widehat{\pi_{i j k}} D$ 's are uncorrelated with one another, $\mathrm{E}\left(\widehat{\pi i j k}^{D} \mid \pi_{i j k}\right)=\pi_{i j k}$, and $\operatorname{Var}\left(\widehat{\pi}_{i j k}^{D} \mid \pi_{i j k}\right)=c_{i j k} \pi_{i j k}\left(1-\pi_{i j k}\right)$, where $c_{i j k}=\sum_{l=1}^{n_{i j}} w_{i j k l}^{2} /\left(\sum_{l=1}^{n_{i j k}} w_{i j k l}\right)^{2}$ for $i=1, \ldots, I ; j=1, \ldots, J_{i}$; $k=1, \ldots, K$. The linear Bayes estimator of $\pi_{i j}$, under the model and squared error loss function, is given by

$$
{\widehat{\pi_{i j}}}^{B}=\sum_{k \in S_{i j}} a_{i j k}\left(B_{i j k} \widehat{i j k k}^{D}+\left(1-B_{i j k}\right) \mu_{i k}\right)+\sum_{k \in S_{i j}} a_{i j k} \mu_{i k},
$$

where $B_{i j k}=d \mu_{i k}^{2} /\left(d \mu_{i k}^{2}+c_{i j k}\left(\mu_{i k}-(d+1) \mu_{i k}^{2}\right)\right)$.
Since the Bayes estimator involves the unknown parameter $\mu_{i k}$, it cannot be used in practice. The following empirical Bayes estimator of $\pi_{i j}$ is obtained when $\mu_{i k}$ is replaced by an estimator, say $\widehat{\mu_{i k}}$, of $\mu_{i k}$ :

$$
\begin{aligned}
{\widehat{\pi_{i j}}}^{E B}=\sum_{k \in S_{i j}} a_{i j k}\left(\widehat{B_{i j k}} \widehat{\pi}_{i j k}^{D}\right. & \left.+\left(1-\widehat{B_{i j k}}\right) \widehat{\mu_{i k}}\right) \\
& +\sum_{k \in S_{i j}} a_{i j k} \widehat{\mu_{i k}}
\end{aligned}
$$

where $\widehat{B_{i j k}}=d{\widehat{\mu_{i k}}}^{2} /\left(d{\widehat{\mu_{i k}}}^{2}+c_{i j k}\left(\widehat{\mu_{i k}}-(d+1){\widehat{\mu_{i k}}}^{2}\right)\right)$. The weight or shrinkage factor $\mathcal{B}_{i j k}$ is a ratio of the variance of $\pi_{i j k}$ in the model to the (unconditional) variance of $\widehat{\pi_{i j k}}$. The estimator of $\mu_{i k}$ is taken to be $\widehat{\mu_{i k}}=\widehat{\pi_{i k}}{ }^{D}$.

## Mean Square Errors

The mean squared error (MSE) of the Bayes estimator ${\widehat{\pi_{i j}}}^{B}$ is defined as MSE $\left({\widetilde{\pi_{i j}}}^{B}\right)=\mathrm{E}\left(\widehat{\pi}_{i j}{ }^{B}-\pi_{i j}\right)^{2}$, where (unconditional) expectation is taken with respect to the model. It can be checked that

$$
\begin{aligned}
\operatorname{MSE}\left({\widehat{\pi_{i j}}}^{B}\right) & =\operatorname{Var}\left({\widehat{\pi_{i j}}}^{B}-\pi_{i j}\right) \\
& =\operatorname{Var}\left({\widehat{\pi_{i j}}}^{B}\right)+\operatorname{Var}\left(\pi_{i j}\right)-2 \operatorname{Cov}\left({\widehat{\pi_{i j}}}^{B}, \pi_{i j}\right) \\
& =\operatorname{Var}\left(\pi_{i j}\right)-\operatorname{Var}\left({\widehat{\pi_{i j}}}^{B}\right) \\
& =d\left(\sum_{k \in S_{i j}} a_{i j k}^{2}\left(1-B_{i j k}\right) \mu_{i k}^{2}+\sum_{k \in S_{i j}} a_{i j k}^{2} \mu_{i k}^{2}\right) .
\end{aligned}
$$

It is customary to take $\operatorname{MSE}\left({\widehat{\pi_{B}}}^{B}\right)$ as the MSE of the empirical Bayes estimator $\widehat{\pi}_{i j}{ }^{E B}$. However, $\operatorname{MSE}\left(\widehat{\pi}_{i j}{ }^{B}\right)$ will underestimate the MSE of $\widehat{\pi_{i j}}{ }^{E B}$ since it does not incorporate the uncertainty due to the estimation of the parameter $\mu_{i k^{k}}$. See Prasad and Rao (1990) and Lahiri and Rao (1995) in this context. Using a standard Bayesian argument, it can be shown that

$$
\operatorname{MSE}\left({\widetilde{\pi_{i j}}}^{E B}\right)=\operatorname{MSE}\left(\widehat{\pi}_{i j}^{B}\right)+\mathrm{E}\left({\widehat{\pi_{i j}}}^{E B}-{\widehat{\pi_{i j}}}^{B}\right)^{2} .
$$

It is necessary to estimate $\operatorname{MSE}\left(\widehat{\pi}_{i j}{ }^{E B}\right)$ since it contains the unknown parameter $\mu_{i k}$. The first term $\operatorname{MSE}\left({\widehat{\pi_{i j}}}^{B}\right)$ can be estimated by

$$
\begin{aligned}
& \operatorname{mse}_{f}\left({\widehat{\pi_{i j}}}^{B}\right)=\operatorname{mse}\left({\widehat{\pi_{i j}}}^{B}\right)- \\
& \frac{J_{i}-1}{J_{i}} \sum_{u=1}^{J_{i}}\left(\operatorname{mse}_{(-u)}\left({\widetilde{\pi_{i j}}}^{B}\right)-\operatorname{mse}\left({\widehat{\pi_{i j}}}^{B}\right)\right),
\end{aligned}
$$

where

$$
\operatorname{mse}\left(\widehat{\pi} i j_{B}^{B}\right)=d\left(\sum_{k \in S_{i j}} a_{i j k}^{2}\left(1-\widehat{B_{i j k}}\right) \widehat{\mu}_{i k}^{2}+\sum_{k \in S_{i j}} a_{i j k}^{2}{\widehat{\mu_{i k}}}^{2}\right)
$$

and

$$
\begin{aligned}
& \operatorname{mse}_{(-u)}({\widehat{(\overbrace{i j}}}^{B})= \\
& \quad d\left(\sum_{k \in S_{i j}} a_{i j k}^{2}\left(1-B_{i j k(-u)}\right) \widehat{\mu_{i k(-u)}^{2}}+\sum_{k \in S_{i j}} a_{i j k}^{2} \widehat{\mu}_{i k(-u)}^{2}\right),
\end{aligned}
$$

with

$$
\widehat{\mu_{i k(-u)}}=\sum_{j \neq u}^{J_{i}} \sum_{l=1}^{n_{i j k}} w_{i j k l} y_{i j k l} / \sum_{j \neq u}^{J_{i}} \sum_{l=1}^{n_{i j k}} w_{i j k l},
$$

and

$$
\begin{aligned}
B_{i j k(-u)}= & d{\widehat{\mu_{i k(-u)}}}^{2} / \\
& \left(d{\widehat{\mu_{i k(-u)}}}^{2}+c_{i j k}\left({\widehat{\mu_{i k(-u)}}}-(d+1){\widehat{\mu_{i k(-u)}} 2}_{2}\right) .\right.
\end{aligned}
$$

See Jiang, Lahiri, and Wan (1998) for comment on these estimators. The second term $\mathrm{E}\left(\widehat{\pi}_{i j}{ }^{E B}-{\widehat{\pi_{i j}}}^{B}\right)^{2}$ can be estimated with the following jackknife estimator:

$$
\mathrm{E}_{j}\left({\widehat{\pi_{i j}}}^{E B}-{\widehat{\pi_{i j}}}^{B}\right)^{2}=\frac{J_{i}-1}{J_{i}} \sum_{u=1}^{J_{i}}\left({\widehat{\pi_{i j-u)}}}^{E B}-{\widehat{\pi_{i j}}}^{E B}\right)^{2},
$$

where

$$
\begin{gathered}
\pi_{i j(-u)}^{E B}=\sum_{k \in S_{i j}} a_{i j k}\left(\widehat{B_{i j k(-u)}}{\widehat{\pi_{i j k}}}^{D}+\left(1-B_{i j k(-u)}\right) \widehat{\mu_{i k(-u)}}\right) \\
+\sum_{k \in S_{i j}} a_{i j k} \widehat{\mu_{i k(-u)}}
\end{gathered}
$$

Thus $\operatorname{MSE}\left(\widehat{\pi}_{i j}{ }^{E B}\right)$ is estimated by

$$
\operatorname{mse}\left(\widehat{\pi}_{i j}{ }^{E B}\right)=\operatorname{mse}_{j}\left(\widehat{\pi}_{i j}^{B}\right)+\mathrm{E}_{J}\left(\widehat{\pi}_{i j}^{E B}-{\widehat{\pi_{i j}}}^{B}\right)^{2}
$$

Jackknife methods are reviewed in the recent text by Shao and Tu (1995).

## 6. AN EXAMPLE

In this study, the primary objective is to provide information about treatment need. Anyone who meets the criteria for lifetime dependence or abuse as defined by the National Technical Center's DSM-III-R criteria, is considered a member of the group of respondents who may have needed treatment during the last year. Several indicator variables were created in the dataset to identify respondents with a diagnosis for substance dependence or abuse for alcohol or drugs. For the purpose of numerical calculations, these indicator variables with 0 and 1 as possible values were treated as response variables ( $y_{i j k l}$ ).

In order to save space, results are presented for the outcome variable Alcohol Dependence only. Results on
other response variables can be obtained from the authors. In order to preserve confidentiality, results for only 40 counties, identified as counties 1 through 40, are reported. Table 1 contains five different estimates of prevalence for alcohol dependence. In general, the direct estimates are highly variable and are often zero. The first synthetic estimator ( S 1 ) is the most stable, producing no zero estimates and estimates with little variability. The second synthetic estimator (S2) is similar to S1, but not as restrictive. The first synthetic estimates are very homogeneous, while the second synthetic estimates are homogeneous within the four planning areas. The estimates produced by the composite estimator are more variable than the other estimates. The empirical Bayes estimator produces estimates very similar to those of S2. In the model leading

Table 1
Five Estimators of Alcohol Dependence Prevalence Expressed as Percents for Forty Counties.
Estimated Standard Errors for Direct (Est.se) and Square Root of Estimated Mean Square Error for Empirical Bayes
( $\sqrt{\text { Est.mse }})$ Estimates in Parentheses Also as Percents

| Estimator |  |  |  |  |  |  |  |  | Number of Groups Observed in County |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
|  | Direct |  | Synthetic 1 | Synthetic 2 | Composite | Emp | Bayes |  |  |
| County | $\widehat{\pi_{i j}}$ | (Est.se) | $\widehat{\pi i j}^{\text {c }}$ | $\pi_{\pi_{i j}}{ }^{\text {S }}$ | $\widetilde{\pi}_{i j} c$ | $\widetilde{\pi}^{\text {ij }}{ }^{\text {cr }}$ | ( $\sqrt{\text { Est.mse }}$ ) | Sample Size |  |
| 1 | 1.7 | (2.4) | 3.4 | 1.6 | 0.9 | 1.6 | (0.33) | 30 | 8 |
| 2 | 4.4 | (2.0) | 3.8 | 1.8 | 7.2 | 2.1 | (0.35) | 111 | 8 |
| 3 | 0.0 | (0.0) | 3.6 | 3.3 | 0.0 | 3.0 | (0.85) | 36 | 8 |
| 4 | 0.0 | (0.0) | 3.3 | 5.6 | 1.6 | 5.3 | (1.79) | 6 | 5 |
| 5 | 9.4 | (4.8) | 3.3 | 5.6 | 14.1 | 6.9 | (1.78) | 37 | 8 |
| 6 | 1.6 | (1.1) | 3.4 | 3.0 | 1.7 | 2.7 | (0.67) | 136 | 8 |
| 7 | 9.3 | (5.8) | 3.4 | 3.1 | 9.9 | 3.1 | (0.81) | 25 | 6 |
| 8 | 0.0 | (0.0) | 3.6 | 3.2 | 0.4 | 3.1 | (0.84) | 20 | 7 |
| 9 | 0.0 | (0.0) | 3.4 | 5.8 | 5.6 | 5.8 | (1.93) | 3 | 3 |
| 10 | 1.5 | (1.3) | 3.4 | 2.1 | 0.7 | 1.9 | (0.54) | 81 | 8 |
| 11 | 0.0 | (0.0) | 3.3 | 1.6 | 0.0 | 1.5 | (0.33) | 58 | 8 |
| 12 | 7.0 | (6.8) | 3.5 | 1.7 | 5.0 | 1.8 | (0.35) | 14 | 6 |
| 13 | 5.7 | (3.8) | 3.3 | 5.5 | 12.9 | 6.4 | (1.75) | 37 | 8 |
| 14 | 0.0 | (0.0) | 3.5 | 1.7 | 0.8 | 1.6 | (0.33) | 12 | 4 |
| 15 | 2.4 | (1.4) | 3.3 | 5.6 | 2.0 | 4.4 | (1.56) | 120 | 8 |
| 16 | 4.1 | (3.5) | 3.3 | 3.0 | 2.5 | 3.0 | (0.77) | 32 | 7 |
| 17 | 2.8 | (2.4) | 3.8 | 1.8 | 1.3 | 1.8 | (0.37) | 48 | 8 |
| 18 | 3.9 | (1.1) | 3.4 | 3.0 | 3.2 | 3.2 | (0.60) | 316 | 8 |
| 19 | 0.0 | (0.0) | 3.4 | 5.7 | 3.7 | 5.7 | (1.95) | 19 | 5 |
| 20 | 3.1 | (3.9) | 3.6 | 3.2 | 14.9 | 3.2 | (0.82) | 20 | 6 |
| 21 | 2.7 | (1.6) | 3.3 | 5.6 | 4.1 | 5.8 | (1.50) | 102 | 8 |
| 22 | 4.2 | (1.8) | 3.3 | 2.1 | 1.8 | 2.2 | (0.42) | 124 | 8 |
| 23 | 9.7 | (2.7) | 4.3 | 8.0 | 11.8 | 8.8 | (2.11) | 121 | 8 |
| 24 | 0.0 | (0.0) | 3.3 | 2.0 | 0.2 | 1.9 | (0.54) | 22 | 6 |
| 25 | 7.8 | (4.7) | 3.3 | 1.6 | 2.8 | 1.8 | (0.33) | 32 | 6 |
| 26 | 0.0 | (0.0) | 3.5 | 1.7 | 0.0 | 1.6 | (0.37) | 28 | 7 |
| 27 | 2.2 | (1.8) | 3.2 | 5.6 | 1.6 | 4.9 | (1.74) | 63 | 8 |
| 28 | 10.5 | (13.7) | 3.4 | 1.6 | 14.2 | 1.7 | (0.35) | 5 | 5 |
| 29 | 0.0 | (0.0) | 3.5 | 3.1 | 1.8 | 3.0 | (0.81) | 12 | 5 |
| 30 | 0.0 | (0.0) | 3.2 | 1.5 | 0.0 | 1.5 | (0.33) | 11 | 6 |
| 31 | 4.6 | (3.2) | 3.5 | 5.9 | 17.0 | 5.8 | (1.87) | 44 | 8 |
| 32 | 8.4 | (3.8) | 3.7 | 3.4 | 8.4 | 4.1 | (0.84) | 52 | 8 |
| 33 | 2.5 | (1.3) | 3.4 | 2.2 | 2.5 | 2.1 | (0.50) | 144 | 8 |
| 34 | 2.9 | (2.4) | 3.6 | 1.7 | 1.3 | 1.7 | (0.35) | 49 | 7 |
| 35 | 0.0 | (0.0) | 3.3 | 3.0 | 0.0 | 2.8 | (0.77) | 22 | 8 |
| 36 | 0.0 | (0.0) | 3.4 | 3.1 | 0.3 | 2.9 | (0.82) | 17 | 6 |
| 37 | 4.2 | (4.0) | 3.0 | 2.0 | 3.4 | 2.1 | (0.54) | 26 | 6 |
| 38 | 0.0 | (0.0) | 3.4 | 5.8 | 3.7 | 5.7 | (1.97) | 16 | 6 |
| 39 | 0.0 | (0.0) | 3.5 | 3.1 | 0.6 | 3.0 | (0.81) | 10 | 6 |
| 40. | 5.3 | (1.9) | 3.4 | 3.1 | 2.9 | 3.5 | (0.69) | 144 | 8 |

Table 2
Summary of Five Estimators of Alcohol Dependence Prevalence for All Counties. Results Expressed as Percents.

| Estimator | minimum | $1^{\text {st }}$ quartile | median | $3^{\text {rd }}$ quartile | maximum | mean | standard deviation |
| :--- | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| Direct | 0.0 | 0.0 | 2.2 | 4.3 | 10.5 | 2.8 |  |
| Synthetic 1 | 3.0 | 3.3 | 3.4 | 3.5 | 4.3 | 3.5 | 0.2 |
| Synthetic 2 | 1.5 | 1.8 | 3.0 | 4.4 | 8.0 | 3.2 | 1.7 |
| Composite | 0.0 | 0.4 | 1.7 | 4.6 | 17.5 | 3.7 | 4.8 |
| Empirical Bayes | 1.5 | 1.8 | 2.8 | 4.2 | 8.8 | 3.2 |  |

to the empirical Bayes estimator, $d$ was chosen to be one third.

Table 1 also displays the estimated standard errors of the direct estimates and the square root of the estimated mean squared errors (see section 4) of the empirical Bayes estimates. The standard errors of the direct estimates, which are calculated as

$$
\sqrt{\widehat{\pi}_{i j}^{D}\left(1-\widehat{\pi}_{i j}^{D}\right) / n_{i j}}
$$

are often (incorrectly) estimated to be zero and are quite variable. The square roots of the estimated MSE of the empirical Bayes estimates are relatively stable and always below 025 .

Table 2 summarizes alcohol dependence estimates in the previous table for all counties in the state. The means of the synthetic and composite estimates are higher than the mean of the direct estimates, because there are fewer zero estimates and the means in the summary tables are unweighted.

## 7. CONCLUSION

We have proposed simple empirical Bayes estimators to estimate county level prevalences. Empirical Bayes estimators are found to be very effective when sample sizes for the counties are small and when prevalences are extremely small. We have introduced a measure of uncertainty of the proposed empirical Bayes estimator based on the jackknife method. The proposed measure incorporates additional sources of variability due to estimation of various model parameters. In our model, presented in this paper, we have implicitly assumed that the selection probabilities are unrelated to $y_{i j k k}$. In the household study reported in this paper, the selection probabilities were unequal and depended on several factors like number of telephone lines and number of adult household members in the household. None of these variables were related to $y_{i j k l}$. The sample allocation to different regions, however, was done based on the number of "treatment admissions" in each region. Hence, the selection probabilities might be indirectly related to $y_{i j k}$. In this paper, we have not addressed the issue of sample selection bias, which can be handled appropriately by following procedures discussed in Pfeffermann (1993).

In this paper, we have not considered the use of auxiliary variables in the model to relate small areas to one another and to facilitate improved estimation. The use of available auxiliary data from the U.S. Census and other
administrative records may be a sensible use of resources that can be used to improve planning for treatment of drug and alcohol abuse and dependence. We plan to do further work in this area with an actual example in a future paper.

## ACKNOWLEDGEMENTS

We wish to thank an anonymous referee who had many useful suggestions on improving our paper and the Gallup Organization for partial support. Additionally, Partha Lahiri wishes to acknowledge partial support from U.S. National Science Foundation Grant SBR-9705574.

## REFERENCES

CASADY, R.J., and LEPKOWSKI, J.M. (1993). Stratified telephone survey designs. Survey Methodology, 19, 103-113.
EFRON, B., and MORRIS, C. (1973). Stein's estimation rule and its competitors - an empirical Bayes approach. Journal of the American Statistical Association, 68, 117-130.
FARRELL, P.J., MacGIBBON, B., and TOMBERLIN, T.J. (1997). Empirical Bayes estimators of small area proportions in multistage designs. Statistica Sinica, 7, 1065-1083.
FAY, R., and HERRIOT, R. (1979). Estimates of income for small places: an application of James-Stein procedures to census data. Journal of the American Statistical Association, 74, 269-277.
GHOSH, M., and LAHIRI, P. (1987). Robust empirical Bayes estimation of means from stratified samples. Journal of the American Statistical Association, 82, 1153-1162.
GHOSH, M., and RAO, J.N.K. (1994). Small area estimation: an appraisal. Statistical Science, 9, 55-93.
JIANG, J., LAHIRI, P., and WAN, S. (1998). Jackknifing Mean Squared Error of Empirical Best Predictor. Unpublished manuscript.
LAHIRI, P., and RAO, J.N.K. (1995). Robust estimation of mean squared error of small area estimators. Joumal of the American Statistical Association, 90, 758-766.
MALEC, D., SEDRANSK, J., MORIARITY, C.L., and LECLERE, F.B. (1997). Small area inference for binary variables in the National Health Interview Survey. Journal of the American Statistical Association, 92, 815-826.
PFEFFERMANN, D. (1993). The role of sampling weights when modeling survey data. International Statistical Review, 61, 317-337.
PRASAD, N.G.N., and RAO, J.N.K. (1990). The estimation of mean squared errors of small area estimators. Journal of the American Statistical Association, 85, 163-171.
SHAO, J., and TU, D. (1995). The Jackknife and the Bootstrap. New York: Springer-Verlag.

# Some Issues in the Estimation of Income Dynamics 

SUSANA RUBIN BLEUER and MILORAD KOVAČEVIĆ́́


#### Abstract

Two design-based estimators of gross flows and transition rates are considered. One makes use of the cross-sectional samples for the estimation of the income class boundaries at each time period and the longitudinal sample for the estimation of counts of units in the longitudinal population (longitudinal counts); this is the mixed estimator. The other one is entirely based on the longitudinal sample, both for the estimation of the class boundaries and the longitudinal counts; this is the longitudinal estimator. We compare the two estimators in the presence of large attrition rates, by means of a simulation. We find that under a less than perfect model of compensation for attrition, the mixed estimator is usually more sensitive to model bias than the longitudinal estimator. Furthermore, we find that for the mixed estimator, the magnitude of this bias overshadows the small gain in precision when compared to the longitudinal estimator. The results are illustrated with data from the Survey of Labour and Income Dynamics and the Longitudinal Administrative Database of Statistics Canada.


KEY WORDS: Attrition; Gross flows; Transition rates; Longitudinal weighting; Cross-sectional weighting; Bootstrap variance estimator.

## 1. INTRODUCTION

Gross flows are counts of transitions from one time point to the other between a number of states for individuals in a population. Related parameters are longitudinal proportions and transition rates. Longitudinal proportions are relative gross flows, while transition rates are relative gross flows conditional on the initial transition state. Estimates of these parameters for transitions between different income classes are required in studies of income dynamics and can be obtained from longitudinal surveys. The boundaries of the income classes often have to be estimated from the survey as well. An example is the low income measure defined as half of the median income, where income is adjusted for family size. Thus, in this case, estimators of counts of transitions to and from "low income state" require the estimation of the income medians at the time period of interest.

The income class boundaries usually refer to the respective cross-sectional populations and have to be estimated from the cross-sectional samples to obtain unbiased estimators. If the change in population from one wave to the other (that is the number of "births" and "deaths") is negligible, a longitudinal sample may represent the respective populations at both time points, and we may estimate the income class boundaries from the longitudinal sample. Otherwise, estimation of income class boundaries from the longitudinal sample may yield biased estimates. By "deaths" we mean real deaths and/or emigration; similarly, "births" means real births and/or immigration.

Two design-based approaches are considered for estimation of longitudinal parameters involving two waves. One approach is based on the cross-sectional samples for the estimation of the class boundaries at each time period and on the longitudinal sample for the estimation of counts
of units in the longitudinal population (longitudinal counts). This results in an estimator that we term the mixed estimator. The other approach uses an estimator based on the longitudinal sample for both the class boundaries and the longitudinal counts, and we call it the longitudinal estimator. The main objective of this study is to compare the two approaches in terms of their performance under different attrition adjustment models.

In order to make the comparison we address two related issues: the impact of attrition on the considered estimators and the estimation of their variance. Attrition refers to the type of non-response that occurs from a certain wave on, until the end of the period of observation. The real issue with attrition is that non-respondents cumulate over time, and the longer the study lasts, the greater is the non-response. In some surveys like SIPP (Survey of Income Program Participation), attrition reached $20 \%$ by the time of the third wave (Rizzo, Kalton, and Brick 1996). Even if extra care is taken in the development of adjustments to compensate for the missing data, the resulting estimators may still be sensitive to a less-than-perfect model of compensation. We investigate empirically the sensitivity to attrition of the estimators considered.

Variance estimation is also an issue because the parameters of interest are non-linear functions of the observations and are dependent on the income class boundaries. The problem of variance estimation of low income proportions, and other measures of income inequality from complex cross-sectional samples was studied by Shao and Rao (1993), Binder and Kovačević (1995) and Kovačević and Yung (1997), among others. In the longitudinal situation, changes in the population over time imply the need to combine different samples and different systems of weights, which complicate variance estimation. The ultimate units

[^11]in a longitudinal sample may belong to different primary sampling units (PSU's) at different waves, some PSU's in the sample at one wave may not belong to the sample at another wave, etc. In this study, we develop an appropriate bootstrap variance estimator for estimators of income dynamics and for the complex design used in the example.

The data used for illustration come from Statistics Canada's Survey of Labour and Income Dynamics (SLID) and Longitudinal Administrative Database (LAD); both are sources with quite accurate longitudinal income data obtained from income tax returns. At the time of the study, SLID had response data from only two waves and the attrition rate was about $10 \%$.

Section 2 outlines the general assumptions for the population under study and for the design. In section 3 we deal with the issue of estimation of longitudinal low income proportion and the impact of attrition on it by means of a small simulation study performed on an artificial data set created assuming the log-normal distribution. Section 4 deals with the bootstrap variance estimation for longitudinal complex surveys. A more extensive simulation of various attrition models, using different adjustment methods and data from a complex design, is described in section 5 , and the results are presented and discussed in section 6.

## 2. LONGITUDINAL POPULATION, SAMPLE, AND WEIGHTING

Let $U_{0}$ represent the population at time 0 and $U_{1}$ represent the population at time 1 . In this study we only consider parameters involving two periods of time, and therefore, the longitudinal population is defined in terms of two waves by $U_{L}$, where $U_{L}=U_{0} \cap U_{1}$.
"Deaths" and "births" from one time period to the next cause a change in the population. If we denote by $U_{D}$ the set of individuals who belong to the population $U_{0}$ at time $t=0$ and do not belong to the population $U_{1}$ at $t=1$ due to "death", and by $U_{B}$ the set of "births" from time 0 to 1 , then the longitudinal population can be expressed as $U_{L}=$ $U_{0} \backslash U_{D}=U_{1} \backslash U_{B}$.
Similarly, we denote by $s_{0}$, a representative sample of $U_{0}$, by $s_{1}$, a representative sample of $U_{1}$ and by $s_{d}$ and $s_{b}$ the respective subsamples of individuals in $s_{0}$ who "died" between $t=0$ and $t=1$, and of individuals in $s_{1}$ "born" between $t=0$ and $t=1$. Hence, the longitudinal sample, representing $U_{L}$, is defined by $s_{L}=s_{0} \cap s_{1}=s_{0} \backslash s_{d}=s_{1} \backslash s_{b}$.

Non-respondents to the initial wave at $t=0$ exist but they are relatively few compared to non-respondents in later waves. For the sake of simplicity, assume that $s_{0}$ is the sample without the initial non-response and with the associated weights already adjusted for it. Attrition from wave 0 to wave 1 will be represented by a subset of individuals in $s_{0}$ denoted by $s_{a}$. Hence, the longitudinal sample affected by attrition can be expressed by $s_{L}^{A}=s_{0} \backslash\left(s_{d} \cup s_{a}\right)$.

Note that for some parameters of interest $s_{d}$ should remain in the longitudinal sample for weighting purposes
(Tambay, Schiopu-Kratina, Mayda, Stukel and Nadon 1997). The parameters considered in this paper refer to the longitudinal population $U_{L}$, and therefore units that "die" from one wave to another are out of scope.

Large scale surveys often employ stratified multistage designs with a large number of strata and relatively few clusters or primary sampling units (PSU) sampled within each stratum. The selected PSU's are subsampled in one or more stages until the ultimate units are obtained. Here we assume that the number of strata and clusters within strata does not change from one wave to the other.

We assume that the cross-sectional samples $s_{t}$ consist of $n_{h}$ sampled PSU's with replacement within stratum $h$ and $m_{h i}^{t}$ units sampled within the $i$-th PSU in stratum $h$, for $t=0,1, h=1, \ldots, H$ and $i=1, \ldots, n_{h}$. Let $\left\{w_{h i j}^{t}\right\}, j=1,2, \ldots, m_{h i}^{t}$ be the set of survey weights corresponding to the crosssectional sample $s_{t}$. We assume that the survey weights provide approximately unbiased estimators of population totals so that $E_{p}\left(\Sigma_{s_{t}} w_{h i j}^{t}\right) \approx N^{t}$, where $N^{t}$ is the size of $U_{t}$, for $t=0,1$. Here $E_{p}$ is the expectation with respect to the design $p(s)$. When the set $s_{a}$ of attritors is large, the original weights $w_{h i j}^{0}$ have to be adjusted to account for the missing units and the adjusted weights should add up, in average, to the size $N_{L}$ of the longitudinal population:

$$
E_{p} E_{m}\left(\sum_{s_{0} \backslash\left(s_{d} \cup s_{a}\right)} w_{h i j}^{L}\right) \approx N_{L} .
$$

Here the expectation $E_{m}$ is taken with respect to the model $m$ assumed for the probability of response.

## Examples:

1. In the Survey of Labour and Income Dynamics (SLDD), every wave has an added component that consists of "cohabitants", i.e., individuals who live in the households of the longitudinal individuals (Lavallée and Hunter 1992). SLID has a stratified two-stage design with approximately $H=400$ strata at each wave. The number of clusters within stratum $h$ may change if there is growth in it. The number of sampled clusters is usually 2 or 3 . When a new panel is selected or an old one is replaced from time $t=0$ to time $t=1$, then the number of sample clusters per stratum may vary.
2. The Longitudinal Administrative Database (LAD) of Statistics Canada is a longitudinal sample obtained from administrative data files and is a representative sample of the income-tax-filing population at any year. The LAD is a collection of many panels since a panel is "bom" at each wave (year). Here non-response is approximately $5 \%$ of the cross-sectional sample every year. Longitudinal administrative samples, like LAD, do not have attrition, but are subject to wave nonresponse caused usually by late filing (Rubin Bleuer 1996). The design for LAD is non-stratified and single stage. We use LAD as a base for a simulation of
attrition because its data are representative of the Canadian income population at every wave.

## 3. ESTIMATION UNDER ATTRITION

Without loss of generality (wlog), in the following we define and explain the two estimation methods in terms of longitudinal low income proportions. In section 6 results on the impact of attrition are given for other two-wave parameters like gross-flows and transition rates. We also assume a negligible amount of "births" and "deaths" in the finite population, relative to the attrition rate. In fact, we assume that $U_{0}=U_{1}$ and that though the units are the same at both time points, the incomes attached to the units can vary.

Let $y_{h i j}^{t}$ be the value of the characteristic of interest (family income adjusted for family size) for the $j$-th ultimate unit in the $i$-th PSU of stratum $h, j=1, \ldots, M_{h i}^{t}, i=1, \ldots, N_{h}^{\prime}$, $h=1, \ldots, H$, and $t=0,1$. Then the longitudinal proportion of individuals with income less than or equal to $x$ at $t=0$ and income less than or equal to $y$ at $t=1$ is given by

$$
\begin{equation*}
F(x, y)=\frac{1}{N_{L}} \sum_{h=1}^{H} \sum_{i=1}^{N_{n}^{0}} \sum_{j=1}^{M_{h i}^{0}} I\left(y_{h i j}^{0} \leq x\right) I\left(y_{h i j}^{1} \leq y\right) \tag{3.1}
\end{equation*}
$$

where, since the two populations coincide, $M_{h i}^{0}=M_{h i}^{1}$, and

$$
N_{L}=\sum_{h=1}^{H} \sum_{i=1}^{N_{h}^{0}} M_{h i}^{0}
$$

coincides with the size of the original population $U_{0}$. I is the indicator function of the incomes smaller than or equal to $x$ and $y$ respectively. $F(x, y)$ is the bivariate distribution function of incomes at times 0 and 1 . Let us now denote by $M_{0} / 2$, half the median income at time $t=0$, and by $M_{1} / 2$ half the median income at $t=1$. Then the longitudinal low income proportion is defined by

$$
\begin{equation*}
\theta=F\left(M_{0} / 2, M_{1} / 2\right) . \tag{3.2}
\end{equation*}
$$

Under complete response, and $U_{0}=U_{1}, \theta$ is the bivariate version of the cross-sectional low income proportion which was studied, among others, by Shao and Rao (1993). Under a framework for the development of asymptotic theory in the design space and under certain regularity conditions on the design and the income distributions, Shao and Rao proved that the estimator of the cross-sectional low income proportion $\hat{F}_{t}\left(\hat{M}_{t} / 2\right)$ is consistent (as the number of PSU's, $N_{\text {PSU }}$, approaches infinity) for general stratified multistage designs where the PSU's are selected with replacement. The framework assumes: (i) the existence of a sequence of finite populations with either increasing number $N_{\text {PSU }}$ of PSU's or increasing number of independent units if the population is not clustered, and (ii) the existence of a corresponding sequence of probability designs with the first stage sample size $n_{\text {PSU }}$ increasing to infinity as $N_{\text {PSU }}{ }^{-\infty}$.

This result is easily extended to the longitudinal situation for the estimator

$$
\begin{equation*}
\hat{\theta}=\hat{F}\left(\hat{M}_{0} / 2, \hat{M}_{1} / 2\right) \tag{3.3}
\end{equation*}
$$

under the assumptions of no change in the population from $t=0$ to $t=1$ and of no attrition.

Let $\hat{M}_{t}$ denote the estimator of the median income at time $t$ based on the cross-sectional sample $s_{t}$ and corresponding cross-sectional weights $\left\{w_{h i j}^{t}\right\}$, for $t=0,1$

$$
\hat{M}_{t}=\inf \left\{y_{h i j}^{t} \in s_{t} \mid \hat{F}_{t}\left(y_{h i j}^{t}\right) \geq 1 / 2\right\},
$$

where

$$
\hat{F}_{t}(y)=\sum_{s_{t}} w_{h i j}^{t} I\left(y_{h i j}^{t} \leq y\right) / \sum_{s_{t}} w_{h i j}^{t},
$$

and let $\tilde{M}_{t}$ denote the estimator of the median income at time $t$ based on the longitudinal sample $s_{L}$ and the longitudinal weights $\left\{w_{h i j}^{L}\right\}$ :

$$
\tilde{M}_{t}=\inf \left\{y_{h i j}^{t} \in s_{L} \mid \tilde{F}_{t}\left(y_{h i j}^{t}\right) \geq 1 / 2\right\},
$$

where

$$
\tilde{F}_{t}(y)=\sum_{s_{L}} w_{h i j}^{L} I\left(y_{h i j}^{\prime} \leq y\right) / \sum_{s_{L}} w_{h i j}^{L} .
$$

Then, there are two possible ways to estimate the longitudinal parameter (3.2):

$$
\begin{align*}
& \hat{\theta}_{\text {mixed }}=\tilde{F}\left(\hat{M}_{0} / 2, \tilde{M}_{1} / 2\right)= \\
& \sum_{s_{L}} w_{h i j}^{L} I\left(y_{h i j}^{0} \leq \hat{M}_{0} / 2\right) I\left(y_{h i j}^{1} \leq \tilde{M}_{1} / 2\right) / \sum_{s_{L}} w_{h i j}^{L}, \tag{3.4}
\end{align*}
$$

and

$$
\begin{align*}
& \hat{\theta}_{\text {long }}=\tilde{F}\left(\tilde{M}_{0} / 2, \tilde{M}_{1} / 2\right)= \\
& \sum_{s_{L}} w_{h i j}^{L} I\left(y_{h i j}^{0} \leq \tilde{M}_{0} / 2\right) I\left(y_{h i j}^{1} \leq \tilde{M}_{1} / 2\right) / \sum_{s_{L}} w_{h i j}^{L} \tag{3.5}
\end{align*}
$$

The first estimator is termed "mixed" because it combines longitudinal and cross-sectional samples. The second is only based on the longitudinal sample. Note that when there are no "births" or "deaths" from one wave to the next, the median at $t=1$ can only be estimated from the longitudinal sample and thus we use $\tilde{M}_{1}$ in the definition of the mixed estimator.

Under attrition, most of the missing data may correspond to individuals who are different from the rest of the population, and failure to account for this may result in biased estimates. Hence, weights are adjusted to compensate for the missing information according to a model. The estimates will become more sensitive to model misspecification as attrition increases. Thus, estimators that are robust to the choice of the model for non-response adjustments are desirable.

In order to compare estimators (3.4) and (3.5) regarding their robustness against incorrect non-response adjustments we made a simple simulation study to empirically estimate the expected (with respect to the design and the attrition model) values of $\hat{\theta}_{\text {mixed }}$ and $\hat{\theta}_{\text {long }}$ when the adjustment model was both the correct one and an incorrect one.

As we already pointed out, if there is no change in the population from one wave to the next, and thus no new sample is selected in the second wave to represent the change, the estimation of the median in the second wave can only be based on the longitudinal sample. Thus the only difference between the two estimators lies in the estimation of the low income measure in the first wave. Hence, wlog, we consider for our simulation the parameter $\theta=F\left(M_{0} / 2, \infty\right)$. In that case, $\theta$ coincides with the crosssectional low income proportion, and the estimator of $\theta$ under complete response, $\hat{\theta}_{\text {cross }}=\hat{F}_{0}\left(\hat{M}_{0} / 2\right)$, is consistent (and thus asymptotically unbiased) as $N_{\text {PSU }}$ tends to infinity.

The simulation study, described in detail in Appendix A, consisted in simulating 1,000 samples of size 1,000 from a log-normal income population similar to the Canadian income population. We first selected a simple random sample without replacement (SRSWOR) from a large finite population of incomes and then we simulated attrition in that sample. Here we call a model missing at random (MAR) if the probability of non-response in the second wave is constant within response classes; and we call a model of attrition missing completely at random (MCAR) if the probability of non-response in the second wave is constant in the whole population. The attrition was simulated following a missing at random model where the nonresponse was induced in a low income class. The boundary of the low income class was the first quintile of the finite population, known a priori. For every sample, we calculated $\hat{\theta}_{\text {mixed }}$, $\hat{\theta}_{\text {long }}$ and $\hat{\theta}_{\text {cross }}$ with adjustments under both the correct (MÅR) and a MCAR attrition model. The arithmetic mean of the estimates approximates the double expectation (with respect to the model and the design) of the first two estimators and approximates the expectation of $\hat{\theta}_{\text {cross }}$ with respect to the design. This last expectation approximates, in turn, the parameter $\theta$, since $\hat{\theta}_{\text {cross }}$ is asymptotically unbiased as $n_{\text {PSU }} \rightarrow \infty$. When the weight adjustments are calculated under the MCAR (incorrect) attrition model, the following relationship is empirically found:

$$
\left|E_{p} E_{m}\left(\hat{\theta}_{\text {mixed }}\right)-\theta\right|>\left|E_{p} E_{m}\left(\hat{\theta}_{\text {long }}\right)-\theta\right|,
$$

where $m$ refers to the simulated attrition model and $p$ refers to the design under SRSWOR. Note that attrition from low income individuals will always bias upwards the estimator of the median and thus we will always obtain $\hat{\theta}_{\text {mixed }} \leq \hat{\theta}_{\text {long: }}$ :

The somewhat surprising result is that the estimator which utilises less information is, in average, nearer the true parameter, meaning that more information, if it is not used well, does not improve the estimator. Similarly, when $\theta_{+}$is the proportion of incomes higher than the income category
boundary estimated from the sample, and attrition is heavier in the lower income categories, we will always have the inequality

$$
\hat{\theta}_{\text {long }}, \leqslant \hat{\theta}_{\text {mixed. }},
$$

and as with the low income proportion $\theta$, the estimator of $\theta_{+}$using less information is, in average, nearer the truth. The description of the simulation and the numerical results are in Appendix A.

The question now is if the bias caused by model misspecification is larger than the increase in variance caused by the attrition. In sections 5 and 6 we tackle this issue by simulating attrition on data from SLID and LAD, calculating $\hat{\theta}$, given by (3.3), $\hat{\theta}_{\text {mixed }}$ and $\hat{\theta}_{\text {long, }}$, and calculating the design variance of the estimators as well.

## 4. BOOTSTRAP VARIANCE ESTIMATION FOR LONGITUDINAL SAMPLES

In order to compare the two approaches to estimation we need to study them in terms of variance and bias under different attrition situations. The estimators $\hat{\theta}_{\text {mixed }}$ and $\hat{\theta}_{\text {ong }}$ defined in section 3 are nonlinear functions of the observations; in addition, the income data come from complex surveys. The variances of these estimators cannot be expressed in simple terms, and we have to rely on approximate variance estimation techniques. We seek a method that is easy to apply to many different complex parameters, and under different designs. We would like to evaluate the two estimation approaches for any parameter, using the same criteria and a consistent method of variance estimation. We concentrate on developing a bootstrap variance estimator that can be applied to a stratified multistage longitudinal design. It is important to emphasize that only the primary sampling units are resampled, not the units within them.

Kovačević and Yung (1997) compared several resampling methods and the Taylor linearization method for variance estimation of cross-sectional estimators of income inequality under a complex survey design. They found, by means of a simulation study, that the best method (in terms of relative bias, coverage properties, stability, robustness against assumptions, etc.) is the Taylor linearization method via the estimating equation approach, and that the next best is the bootstrap method.

In the calculation of the number of individuals in one income class at time 0 and another income class at time 1 , the units in the longitudinal sample $s_{L}$ are involved, and the bootstrap sampling scheme must ensure the selection of units in $s_{L}$. However, if we are confined only to the resampling of units in $s_{L}$, we would not allow enough variability for the consistent estimation of the variance of the cross-sectional quantile estimators $\hat{M}_{0}$ and $\hat{M}_{1}$. Therefore, the bootstrap sample should contain as well elements from $s_{0} \backslash s_{1}, s_{L}$ and $s_{1} \backslash s_{0}$ at each iteration.

We assume a stratified two-stage design, and we assume that the primary sampling units (PSU's) are exactly the same at both $t=0$ and $t=1$, that is there are no "deaths" or "births" of PSU's from one wave to the next.

This is the case of the first and second waves of SLID. The "births" and cohabitants that appear in the second wave live in dwellings with individuals who were selected in the first wave. Every unit $u$ in $s_{0}$ or $s_{1}$ is assigned to a PSU that was selected in the first wave:
if $u \in s_{0} \cap s_{1}$, then we assign $u$ to the PSU corresponding to its original dwelling at time $t=0$;
if $u \in s_{0} \backslash s_{1}$, then we assign $u$ to the PSU it belonged to at $t=0$;
if $u \in s_{1} \backslash s_{0}$ and $u$ lives with $v \in s_{0} \cap s_{1}$, then we assign $u$ to the PSU of $v$.

In this way we reduce the problem to a cross-sectional situation. Then we perform the following steps:

Suppose that the original weights of an individual are $w_{h j,}^{0}, w_{h i j}^{1}, w_{h i j}^{L}$.

1. We select a simple random sample with replacement. (SRSWR) of PSU's of size $n_{h}-1$, independently in each stratum. A union of such samples is denoted by $s_{\text {boor }}^{*}$ It contains a subsample $s_{0 \text { boot }}{ }^{\text {b }}$ of units from $s_{0}$ that are not in $s_{1}$, a subsample $s_{1 \text { boot }}^{*}$ of units from $s_{1}$ that are not in $s_{0}$ and a subsample $s_{\text {Lboot }}$ of units that are in both, $s_{0}$ and $s_{1}$.
2. Let $m_{h i}^{*}$ be the number of times the $h i$-th PSU is selected; the bootstrap modifications of the weights are

$$
\begin{aligned}
& w_{h i j}^{*(0)}=\frac{n_{h}}{n_{h}-1} m_{h i}^{*} w_{h i j}^{0} ; \\
& w_{h i j}^{*(1)}=\frac{n_{h}}{n_{h}-1} m_{h i}^{*} w_{h i j}^{1} ; \\
& w_{h i j}^{*(L)}=\frac{n_{h}}{n_{h}-1} m_{h i}^{*} w_{h i j}^{L}
\end{aligned}
$$

The weights $w_{h i j}^{(1)}$ and $w_{h i j}^{(L)}$ are obtained from the original weights by multiplying by an attrition adjustment factor. The adjustment factors are the inverses of the response probabilities (assumed different for each response class). These probabilities are estimated from the original data set. The process of estimation of the response probabilities and subsequent adjustment is imitated in the bootstrap resampling: for each bootstrap sample $s_{\text {boot }}^{*}$ the adjustments are recalculated, to produce new $w_{h i j}^{(1)}$ and $w_{h i j}^{(L)}$.

Then the estimate $\hat{\theta}_{\text {mixed }}$ computed from a bootstrap sample $s_{\text {boot }}^{*}$ is

$$
\hat{\theta}_{m b}^{*}=\sum_{s_{i \text { bood }}} w_{h i j}^{*(L)} I\left\{y_{h i j}^{0} \leq \hat{M}_{0}^{*} / 2\right\} I\left\{y_{h i j}^{1} \leq \hat{M}_{1}^{*} / 2\right\} / \sum_{s_{i \text { lood }}} w_{h i j}^{*(L)},
$$

where

$$
\hat{M}_{t}^{*}:=\inf \left\{y_{h i j}^{t} \in s_{t \text { boot }}^{*} \cup s_{L \text { boot }}^{*} \mid \hat{F}_{t}^{*}\left(y_{h i j}^{t}\right) \geq 1 / 2\right\}
$$

and

$$
\hat{F}_{t}^{*}(y):=\sum_{s_{i \text { box }}^{*} \cup s_{L \text { boot }}^{*}} w_{h i j}^{*(t)} I\left(y_{h i j}^{\prime} \leq y\right) / \sum w_{h i j}^{*(t)}, t=0,1
$$

The estimate $\hat{\theta}_{\text {long }}$ computed from $s_{\text {Lboot }}^{*}$ is

$$
\hat{\theta}_{l b}^{*}=\sum_{s_{i b o a t}} w_{h i j}^{*(L)}\left\{y_{h i j}^{0} \leq \tilde{M}_{0}^{*} / 2\right\} I\left\{y_{h i j}^{1} \leq \tilde{M}_{1}^{*} / 2\right\} / \sum_{s_{L \text { bood }}} w_{h i j}^{*(L)},
$$

where the medians $\bar{M}_{\text {* }}{ }^{\text {t }}$ are estimated from the longitudinal sample $s_{\text {Lboot }}^{*}$
3. Repeat steps 1 and 2 a large number of times, say $B$. $A$ Monte Carlo estimate of the variance is obtained as

$$
v_{B}\left(\hat{\theta}_{\text {mixed }}\right)=\frac{1}{B} \sum_{b}\left(\hat{\theta}_{m b}^{*}-\overline{\hat{\theta}}_{\text {mixed }}\right)^{2}
$$

and

$$
v_{B}\left(\hat{\theta}_{\text {long }}\right)=\frac{1}{B} \sum_{b}\left(\hat{\theta}_{l b}^{*}-\overline{\hat{\theta}}_{\text {long }}\right)^{2},
$$

where

$$
\overline{\hat{\theta}}_{\text {mixed }}=\sum_{b} \hat{\theta}_{\mathrm{mb}}^{*} / B \text { and } \overline{\hat{\theta}}_{\text {long }}=\sum_{b} \hat{\theta}_{l b}^{*} / B .
$$

By resampling the original PSU's, we reduce the problem of variance estimation in a longitudinal survey to that of a cross-sectional framework. This is an extension of the bootstrap variance estimator developed by Rao and Wu (1988) and later Kovačević and Yung (1997) for variance estimation of cross-sectional income inequality measures from a stratified multistage sample survey.

In order to accommodate attrition, we look at the original data set as a set of longitudinal records. Then, attrition can be viewed as item non-response and accordingly the weight adjustment for attrition can be considered as ratio imputation (Hajek-ratio).

Indeed, recalling that $s_{0}=s_{L} \cup s_{a}$, let us denote by $y_{h i j}^{(1) *}$ ( $h i j \in s_{a}$ ) the ratio-imputed value of wave one information, based on the observed data in the longitudinal sample; and let us denote by $y_{h i j}^{(0) *}\left(h i j \in s_{a}\right)$ the ratio-imputed value of wave zero information based on the longitudinal sample $\left(s_{L}\right)$. Note that the values $y_{h i j}^{(0)}\left(h i j \in s_{a}\right)$ are not missing, but we need $y_{h i j}^{(0) *}$ to represent the weight adjustment in the longitudinal sample.

The estimation of $\theta$ with weight adjustments for attrition is equivalent to estimation on a "complete" imputed data set weighted by $w_{h i j}^{(0)},\left(h i j \in s_{0}\right)$. The set

$$
D_{\text {mixed }}=\left\{\left(y_{h i j}^{(0)}, y_{h i j}^{(1)}\right), h i j \in s_{L},\left(y_{h i j}^{(0)}, y_{h i j}^{(1) *}\right), h i j \in s_{a}\right\},
$$

is used to calculate $\hat{\theta}_{\text {mixed }}$ and the set

$$
\left.D_{\text {long }}=\left\{\left(y_{h i j}^{(0)}, y_{h i j}^{(1)}\right), h i j \in s_{L}, \partial_{h i j}^{(0) *}, y_{h i j}^{(1) *}\right), h i j \in s_{a}\right\},
$$

is used to calculate $\hat{\theta}_{\text {long }}$.
We noted above that adjustment due to attrition is equivalent to ratio imputation for item non response. Hence the variance estimator proposed here has the same properties as the cross-sectional variance estimators for imputed survey data: consistency now follows from the consistency of the bootstrap variance for imputed survey data (Shao and Sitter 1996) and good coverage properties and small relative bias, as documented by Kovačević and Yung (1997). In the case of small number of PSU's per stratum (SLID has two or three PSU's per stratum) Kovar, Rao and Wu (1988) showed empirically that the bootstrap variance estimate overestimates the true variance by no more than $10 \%$.

## 5. EMPIRICAL STUDY

In order to compare the two approaches to estimation, we now consider two real longitudinal surveys, SLID and LAD, and simulate different attrition situations. The study begins with a sample of complete respondents in two waves. The response pattern of the individuals in this sample can be presented as

| RESPONDENTS | $t=0$ | $t=1$ |
| :---: | :---: | :---: |
| Responded in both waves | X | X |

Here X means that the response is available for the individual at the corresponding wave, and 0 means the opposite. We omitted "births" and "deaths" from both SLID and LAD.

The respondents in the first wave are divided into two response classes, low income class and else. The boundary is given by $\hat{M}_{0} / 2$, where $\hat{M}_{0}$ is the estimate of the median income at time $t=0$, based on all respondents in the first wave. The size of the SLID sample (Ontario) is approximately 10,000 , from which 2,000 were low income individuals.

We simulate three different attrition situations. In all of them we select a subsample of the complete respondents with pattern XX and convert them into the following pattern

| RESPONDENTS | $t=0$ | $t=1$ |
| :--- | :---: | :---: |
| Responded in the first wave 1, <br> not in the second | X | 0 |

1) $10 \%$ attrition. We select a subsample at random of individuals with low income $\left\{y_{h i j}^{0} \leq \hat{M}_{0} / 2\right\}$ at time $t=0$, and make them non-respondents. In order to have $10 \%$ of the overall sample missing at time $t=1$, we convert $50 \%$ of low income individuals into nonrespondents in the second wave.
2) $\mathbf{2 0 \%}$ attrition. We select $70 \%$ of individuals from $\left\{y_{h i j}^{0} \leqslant \hat{M}_{0} / 2\right\}$ and $7.5 \%$ from $\left\{y_{h i j}^{0}>\hat{M}_{0} / 2\right\}$ at random and convert them into non-respondents at the second wave. This results in $20 \%$ overall attrition for SLID.
3) $30 \%$ attrition. We select $80 \%$ of individuals from $\left\{y_{h i j}^{0} \leq \hat{M}_{0} / 2\right\}$ and $17.5 \%$ from $\left\{y_{h i j}^{0}>\hat{M}_{0} / 2\right\}$ at random and convert them into non-respondents at the second wave. This results into $30 \%$ attrition for SLD.

We then consider two different adjustment models for each situation:

Model 1: The non-respondents are missing completely at random (MCAR). This model is the worst possible model that we might use given that attrition is usually experienced by the group of low income individuals.

Model 2: The non-respondents are missing at random from the low-income class. We allow for a small increase of the upper boundary for the low-income class, so that the response classes are defined as $\left\{y_{h i j}^{0} \leq \hat{M}_{0} / 2+\hat{M}_{0} / 10\right\}$ and $\left\{y_{h i j}^{0}>\hat{M}_{0} / 2+\hat{M}_{0} / 10\right\}$. We assume that this model as one of the best possible models under our setup, since it recognizes the response classes as separated by low income boundaries.

We chose these two models of adjustment because they represent the two extremes. In practice, we may only be able to choose a model between these two.

Let $l_{t}$ denote the low income measure defined as half the median: $l_{t}=M_{t} / 2, t=0,1$. Several longitudinal parameters were studied. We define some of them in Table 5.1. The values of the estimates are presented in Tables B1 to B3. The standard errors were obtained using the bootstrap method described in section 4 assuming that the corresponding adjusted weights are known a priori and do not change for each bootstrap sample.

Table 5.1
Some Longitudinal Parameters Evaluated in the Empirical Study

| Proportions |  |  |
| :--- | :--- | :---: |
| The proportion of individuals with income |  |  |
| (adjusted family income) below $l_{t}$ at $t=0,1$. | $p\left(y_{0} \leq l_{0}, y_{1} \leq l_{1}\right)$ |  |
| The proportion of individuals with income |  |  |
| below $l_{0}$ at $t=0$ and above $l_{1}^{*}=(11 / 10) l_{1}$ at |  |  |
| $t=1 ;$ the factor $11 / 10$ is used to be able to | $p\left(y_{0} \leq l_{0}, y_{1} \geq l_{1}^{*}\right)$ |  |
| detect true transitions from one state to the |  |  |
| other. |  |  |
| The proportion of individuals with income |  |  |
| above $l_{0}^{*}=(11 / 10) l_{0}$ at $t=0$ and below $l_{1}$ at | $p\left(y_{0} \geq l_{0}^{*}, y_{1} \leq l_{1}\right)$ |  |
| $t=1$. |  |  |
| The proportion of individuals with income |  |  |
| above $l_{0}^{*}$ and $l_{1}^{*}$ at times 0 and 1 | $p\left(y_{0} \geq l_{0}^{*}, y_{1} \geq l_{1}^{*}\right)$ |  |
| respectively. |  |  |

Conditional Rates
The probability of an individual having low income at the second wave (second year), given that he or she had low income in the $p\left(y_{1} \leq l_{1} \mid y_{0} \leq l_{0}\right)$ first wave.
The probability of not having low income at the second wave, given that the individual
$p\left(y_{1}>l_{1}^{*} \mid y_{0} \leq l_{0}\right)$ had low income at the first wave.

## 6. RESULTS AND DISCUSSION

The empirical study shows that attrition does affect estimates adversely, but different outcomes result depending on whether the parameter of interest is cross-sectional or longitudinal. In the estimation of later-wave cross-sectional parameters, estimators based on the actual longitudinal sample are more biased than estimators based on the crosssectional sample whether the model of adjustment is sound or not, (see for example the estimates of the median at $t=0$ in Tables B.1-B.3.) However, in the estimation of longitudinal parameters, longitudinal estimators (based entirely on the longitudinal sample) are less biased than mixed estimators (based on the three samples.)

Tables 6.1 a and 6.1 b present gross flows estimated from SLD and LAD data, respectively. The estimates are calculated with the complete data set and after $20 \%$ nonresponse is simulated in the second wave. For the complete data set the longitudinal and the mixed estimators coincide (this is the no-attrition situation). As explained in the previous section, $20 \%$ of non-response was simulated by eliminating $70 \%$ of the responses from individuals who were low-income and $7.5 \%$ of individuals with income higher than $l_{0}$ in the first wave. The adjustment for non-response was done assuming that the individuals were missing completely at random.

The applied adjustment model means that the original survey weights were adjusted with a factor of 1.25 (representing 20\% attrition) across the sample, whereas the correct adjustment should have been with a factor of 3.33
(representing $70 \%$ attrition) in the domain of individuals who were low income in the first wave (1993), and with a factor of 1.08 (representing $7.5 \%$ attrition) in the domain of individuals who had an income higher than the LIM in 1993. Thus, by adjusting with an incorrect model, we incur in a much greater error in the estimation of one domain ( $y_{o} \leq M_{0} / 2$ ) than in the estimation of the other ( $y_{o}>M_{0} / 2$ ).

We see from these tables that both the mixed and longitudinal estimates seriously underestimate the parameter of interest in the first column and overestimate it in the second column, assuming that estimation based on the complete data set results in reasonable and acceptably good estimates. It is obvious that the mixed estimates are more affected by the wrong adjustment, verifying the inequality stated in section 3.

Table 6.1
Gross Flows Estimated From SLID and LAD 20\% Attrition
(70\% of the Low Income Missing)
a. SLID, Ontario

| a. SLID, Ontario |  |  |  |
| :---: | ---: | ---: | ---: |
| 1994 | 1993 |  |  |
|  | $1,602,000$ | 113,000 | No attrition |
|  | 425,000 | 152,000 | Mixed |
|  | 710,000 | 125,700 | Longitudinal |
| $y_{1}>1.1 M_{1} / 2$ | 70,000 | $8,080,000$ | No attrition |
|  | 15,000 | $8,975,000$ | Mixed |
|  | 30,000 | $8,870,000$ | Longitudinal |

b. LAD, Sub-Area from Toronto

| 1992 | $y_{o} \leq M_{0} / 2$ | $y_{o}>1.1 M_{0} / 2$ |  |
| :---: | :---: | :---: | ---: |
|  | 2,700 | 640 | No attrition |
|  | 1,100 | 800 | Mixed |
|  | 1,500 | 740 | Longitudinal |
| $y_{1}>1.1 M_{1} / 2$ | 580 | 10,420 | No attrition |
|  | 190 | 12,150 | Mixed |
|  | 380 | 11,650 | Longitudinal |

Tables B. 1 to B.3, given in Appendix B, show the results for SLID at three different attrition levels: $10 \%, 20 \%$ and $30 \%$. For each parameter the estimates and their corresponding bootstrap standard errors were calculated using both the longitudinal and mixed estimators. First, we calculated them for the ideal longitudinal "no attrition" sample, and then for the reduced sample adjusted under the two non-response models described in section 5 . We provide the estimate of the model bias as the difference $\hat{\Delta}$ between the estimate obtained under the model and the "no attrition" estimate.

The numbers in Tables B. 1 to B. 3 repeat the same pattern that was shown for gross flows in Tables 6.1a and 6.1 b , i.e., the estimates obtained using the longitudinal
estimators are "less sensitive" to a choice of the non-response adjustment model. At the same time, there is almost no difference between the corresponding standard errors (given in parentheses). Overall, we found that the estimates of the standard errors of $\hat{\theta}_{\text {mixed }}$ are slightly smaller than for $\hat{\theta}_{\text {long }}$, and that this negligible difference in favour of $\hat{\theta}_{\text {mixed }}$ is not enough to compensate for the larger bias affecting $\hat{\theta}_{\text {mixed }}$, induced by the wrong adjustment for non-response.

There is no difference between $\hat{\theta}_{\text {mixed }}$ and $\hat{\theta}_{\text {long }}$ when the second (best) adjustment for non-response is used. This is true for most parameters except for the conditional rates: for example, in Table B. 1 ( $10 \%$ attrition) the empirical bias of the mixed estimate of the conditional rate of remaining low income in 1994 was found statistically significant, whereas the empirical biases of the longitudinal estimates of the two conditional rates were found non-significant (see Appendix C). Of course the "perfect" model for adjustment (not shown here) yields exactly the same numbers from the two estimation methods, and $\hat{\theta}_{\text {mixed }}$ is approximately equal to $\hat{\theta}_{\text {long }}$ for any parameter considered, but their variances differ.

We introduce a single "sensitivity measure" that combines information on sampling standard error and model bias caused by the applied attrition adjustment model:

$$
\begin{equation*}
s\left(\hat{\theta}_{\text {mixed }}^{A}\right)=\left|\cdot \sqrt{\frac{\left(\hat{\theta}_{\text {mixed }}^{A}-\hat{\theta}_{0}\right)^{2}+\text { s.e. } .^{2}\left(\hat{\theta}_{\text {mixed }}^{A}\right)}{\text { s.e. }^{2}\left(\hat{\theta}_{0}\right)}}-1\right| . \tag{6.1}
\end{equation*}
$$

Here, $\hat{\theta}_{0}$ and $\hat{\theta}_{\text {mixed }}^{A}$ (A = MCAR or MAR) denote estimates obtained under "no attrition" and under an attrition adjustment model respectively, and s.e.(.) stands for the standard error due to sampling. Similarly, we define $s\left(\hat{\theta}_{\text {long }}^{A}\right)$. If an attrition adjustment does not change by much the value of an estimator and its standard error (compared with the estimate obtained using another attrition adjustment), we say that the estimator is relatively insensitive to the applied attrition model. The ratios of the two sensitivity measures of the two applied adjustment models are defined by

$$
\begin{align*}
\text { ratio }_{\text {mixed }} & =s\left(\hat{\theta}_{\text {mixed }}^{\mathrm{MCAR}}\right) / s\left(\hat{\theta}_{\text {mixed }}^{\mathrm{MAR}}\right) \text { and } \\
\text { ratio }_{\text {long }} & =s\left(\hat{\theta}_{\text {long }}^{\mathrm{MCAR}}\right) / s\left(\hat{\theta}_{\text {long }}^{\mathrm{MAR}}\right) . \tag{6.2}
\end{align*}
$$

Values of the ratio for different attrition scenarios are presented in Charts B.1-B. 3 in Appendix B.

From Charts B.1-B. 3 it is evident that the ratio of the sensitivity measures (6.2) is systematically lower for the longitudinal estimator. This further means that the sensitivity measure of longitudinal estimator under applied adjustment models are more alike than those of the mixed estimator. The longitudinal estimator seems to be more insensitive on the applied adjustment models. We refer to this as "robustness". Regarding the simulated attrition rates, the charts show that as the attrition rate increases the sensitivity measures ratio approaches 1 . This means that both models of adjustment perform alike for higher attrition rates, and then the choice
of the estimators becomes more important than the model used in the adjustments.

We summarize our findings as following

1) In the estimation of later wave cross-sectional parameters, estimators based on the actual longitudinal sample are more biased than estimators based on the cross-sectional sample.
2) In the estimation of longitudinal parameters, both the mixed and longitudinal estimators are considerably biased if the wrong model of attrition adjustment is used.
3) The longitudinal estimator is more robust against the inappropriate adjustment for attrition than the mixed estimator. Under the perfect adjustment model, these two estimators perform alike.
4) In general, the sampling variance of the mixed estimator is smaller than the variance of the longitudinal estimator. This relationship remains steady over different attrition rates and different adjustment models.
5) For the mixed estimators, the magnitude of the bias coming from inappropriate non-response adjustments overshadows the small gain in precision when compared to the longitudinal estimates.
6) Different models of adjustment perform alike for higher attrition rates. In this case the choice of the estimator is more important than the efforts at model improvement.

## ACKNOWLEDGEMENTS

The authors would like to thank the associate editor and the referee for their helpful comments and suggestions.

## APPENDIX A

## Description of Simulation for Section 3.

The simulation consisted of the following steps.

1. Let $X$ be a log-normal random variable, $X \sim \exp$ ( N $(\mu=10.3, \sigma 2=0.64)\}$. These parameters correspond to a median similar to that of the SLID estimate for 1992, and a spread similar to that of the Canadian population. The low income boundary was set to the first quintile of the income population. The first quintile $q_{1}$ was estimated from a simulated sample of size 50,000 . The value obtained was $q_{1}=14,901$.
2. From this infinite population, 1,000 independent random samples of size 1,000 were selected.
3. To simulate attrition, from each sample, $50 \%$ of the units with income below $q_{1}$ were selected at random and dropped from the sample for the calculations
pertinent to the second wave. Thus $10 \%$ attrition from the low income class was simulated (MAR model).
4. For each sample $i$, the low income proportion estimators $\hat{\theta}_{\text {mixed }}(i)$ and $\hat{\theta}_{\text {long }}(i)$ were estimated with weights adjusted under both the correct attrition model (MAR) and under the incorrect assumption of units missing completely at random (MCAR).
5. Also, for each sample $i$, we calculated the cross-sectional estimator under complete response, $\hat{\theta}_{\text {cross }}(i)$, which is entirely based on the sample before attrition.
6. For each type of adjustment, the expectation with respect to design and the attrition model of $\hat{\theta}_{\text {mixed }}, \hat{\theta}_{\text {long }}$ and $\hat{\theta}_{\text {cross }}$ was estimated by their respective arithmetic means over 1,000 samples with the incorrect adjustment, and over 539 samples for the correct adjustment.

## Result of the Simulation

The next two tables show that under the incorrect specification, the longitudinal estimator has less bias than the mixed estimator, and under the correct specification for the adjustment, both estimators are unbiased (with respect to the model and the design).

The arithmetic mean of $\hat{\theta}_{\text {cross }}$ over the 1,000 different values was 0.193 , and the standard deviation of the 1,000
values was 0.012 . This implies that for a SRSWOR sample of size 1,000 , the estimate $\hat{\theta}_{\text {cross }}$ is quite stable and its expectation can be used as surrogate for $\theta$. The expected values of $\hat{\theta}_{\text {mixed }}$ and $\hat{\theta}_{\text {long }}$ are estimated by 0.109 and 0.145 , with standard deviations of 0.011 and 0.013 respectively.

Expected Values Under Mis-Specification of the Adjustment Model

| Estimator | Expected <br> value | Standard <br> deviation | Number of <br> samples |
| :---: | :---: | :---: | :---: |
| $\hat{\theta}_{\text {mined }}$ | 0.109 | 0.011 | 1,000 |
| $\hat{\theta}_{\text {long }}$ | 0.145 | 0.013 | 1,000 |
| $\hat{\theta}_{\text {cross }}$ | 0.193 | 0.012 | 1,000 |

Expected Values Under the Correct Adjustment of the Attrition Model

| Estimator | Expected <br> value | Standard <br> deviation | Number of <br> samples |
| :---: | :---: | :---: | :---: |
| $\hat{\theta}_{\text {mixed }}$ | 0.193 | 0.016 | 539 |
| $\hat{\theta}_{\text {long }}$ | 0.193 | 0.014 | 539 |
| $\hat{\theta}_{\text {cross }}$ | 0.194 | 0.012 | 539 |

We see from the table above that both estimators approximate the true value of the parameter if the adjustment model is correct.

## APPENDIX B

## Table B. 1

Estimates of Different Population Characteristics and Their Standard Errors Obtained From the Complete Data Set and Under $10 \%$ Attrition in the Second Wave

| Parameter | Type of Estimator | No Attrition |  | Attrition Adjustment Model |  |  |  |  |  |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
|  |  |  |  | MCAR |  |  | MAR |  |  |
|  |  | $\hat{\theta}$ | s.e. | $\hat{\theta}$ | s.e. | $\hat{\Delta}$ | $\hat{\theta}$ | s.e. | $\hat{\Delta}$ |
| Quantiles |  |  |  |  |  |  |  |  |  |
| $M_{0}$ (median, $t=0$ ) | M | 29,300 | $(1,000)$ | 29,300 | $(1,000)$ | 0 | 29,300 | $(1,100)$ | 0 |
|  | L | 29,300 | $(1,000)$ | 30,900 | $(1,000)$ | -1,600 | 29,300 | $(1,200)$ | 0 |
| $M_{1}($ median, $t=1)$ | M | 28,600 | $(1,100)$ | 30,400 | (900) | -1,800 | 28,600 | $(1,100)$ | 0 |
|  | L | 28,600 | $(1,000)$ | 30,400 | $(1,000)$ | -1,800 | 28,600 | $(1,200)$ | 0 |
| Proportions |  |  |  |  |  |  |  |  |  |
| $p\left(y_{0} \leq l_{0}, y_{1} \leq l_{1}\right)$ | M | 0.156 | (0.010) | 0.092 | (0.008) | 0.064 | 0.124 | (0.009) | 0.032 |
|  | L | 0.156 | (0.010) | 0.111 | (0.008) | 0.045 | 0.124 | (0.009) | 0.032 |
| $p\left(y_{0} \leq l_{0}, y_{1} \geq l_{1}^{*}\right)$ | M | 0.007 | (0.002) | 0.003 | (0.001) | 0.004 | 0.005 | (0.002) | 0.002 |
|  | L | 0.007 | (0.002) | 0.004 | (0.001) | 0.003 | 0.005 | (0.002) | 0.002 |
| $p\left(y_{0} \geq l_{0}^{*}, y_{1} \leq l_{1}\right)$ | M | 0.011 | (0.002) | 0.013 | (0.003) | -0.002 | 0.012 | (0.002) | 0.001 |
|  | L | 0.011 | (0.002) | 0.011 | (0.003) | 0.000 | 0.012 | (0.002) | -0.001 |
| $p\left(y_{0} \geq l_{0}^{*}, y_{1} \geq l_{1}^{*}\right)$ | M | 0.790 | (0.040) | 0.840 | (0.040) | -0.050 | 0.804 | (0.042) | -0.005 |
|  | L | 0.790 | (0.040) | 0.831 | (0.042) | -0.041 | 0.804 | (0.043) | -0.005 |
| Conditional Rates |  |  |  |  |  |  |  |  |  |
| $p\left(y_{1} \leq l_{1} \mid y_{0} \leq l_{0}\right)$ | M | 0.923 | (0.023) | 0.546 | (0.025) | 0.377 | 0.734 | (0.033) | 0.189 |
|  | L | 0.923 | (0.023) | 0.926 | (0.030) | -0.003 | 0.921 | (0.031) | 0.002 |
| $p\left(y_{1}>l_{1}^{*} \mid y_{0} \leq l_{0}\right)$ | M | 0.040 | (0.010) | 0.018 | (0.006) | 0.022 | 0.030 | (0.009) | 0.010 |
|  | L | 0.040 | (0.010) | 0.036 | (0.010) | 0.004 | 0.038 | (0.012) | 0.002 |

$l_{i}^{*}=(11 / 10) l_{i} ; l_{i}^{*}$ is used to identify true transitions from one wave to the next.
$M$ denotes Mixed ( $\hat{\theta}$ ) and L denotes Longitudinal ( $\bar{\theta}$ ) estimates.
$\hat{\theta}$ is the estimate, s.e. denotes the standard error of the estimate, and $\hat{\Delta}$ is the difference between the corresponding estimates obtained using the attrition adjustment model and assuming no attrition.

Table B. 2
Estimates of Different Population Characteristics and Their Standard Errors Obtained From the Complete Data Set and Under 20\% Attrition in the Second Wave

| Parameter | Type of Estimator | No Attrition |  | Attrition Adjustment Model |  |  |  |  |  |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
|  |  |  |  | MCAR |  |  | MAR |  |  |
|  |  | $\hat{\theta}$ | s.e. | $\hat{\theta}$ | s.e. | $\hat{\Delta}$ | $\hat{\theta}$ | s.e. | $\hat{\Delta}$ |
| Quantiles |  |  |  |  |  |  |  |  |  |
| $M_{0}($ median, $t=0)$ | M | 29,300 | $(1,000)$ | 29,300 | $(1,000)$ | 0 | 29,300 | $(1,000)$ | 0 |
|  | L | 29,300 | $(1,000)$ | 31,800 | $(1,100)$ | -2500 | 29,300 | $(1,000)$ | 0 |
| $M_{1}($ median, $t=1)$ | M | 28,600 | $(1,100)$ | 31,100 | (800) | -2500 | 28,600 | $(1,000)$ | 0 |
|  | L | 28,600 | $(1,000)$ | 31,100 | (900) | -2500 | 28,600 | $(1,000)$ | 0 |
| Proportions |  |  |  |  |  |  |  |  |  |
| $p\left(y_{0} \leq l_{0}, y_{1} \leq l_{1}\right)$ | M | 0.156 | (0.010) | 0.055 | (0.008) | 0.101 | 0.096 | (0.012) | 0.060 |
|  | L | 0.156 | (0.010) | 0.080 | (0.006) | 0.076 | 0.096 | (0.013) | 0.060 |
| $p\left(y_{0} \leq l_{0}, y_{1} \leq l_{1}^{*}\right)$ | M | 0.007 | (0.002) | 0.001 | (0.001) | 0.006 | 0.004 | (0.002) | 0.003 |
|  | L | 0.007 | (0.002) | 0.003 | (0.001) | 0.004 | 0.004 | (0.002) | 0.003 |
| $p\left(y_{0} \geq l_{0}^{*}, y_{1} \leq l_{1}\right)$ | M | 0.011 | (0.002) | 0.014 | (0.003) | -0.003 | 0.012 | (0.002) | -0.001 |
|  | L | 0.011 | (0.002) | 0.012 | (0.002) | -0.002 | 0.012 | (0.002) | -0.001 |
| $p\left(y_{0} \geq l_{0}^{*}, y_{1} \geq l_{1}^{*}\right)$ | M | 0.790 | (0.040) | 0.864 | (0.041) | -0.074 | 0.820 | (0.049) | -0.030 |
|  | L | 0.790 | (0.040) | 0.855 | (0.042) | -0.065 | 0.820 | (0.048) | -0.030 |
| Conditional Rates |  |  |  |  |  |  |  |  |  |
| $p\left(y_{1} \leq l_{1} \mid y_{0} \leq l_{0}\right)$ | M | 0.923 | (0.023) | 0.323 | (0.051) | 0.600 | 0.570 | (0.057) | 0.353 |
|  | L | 0.923 | (0.023) | 0.914 | (0.040) | 0.009 | 0.928 | (0.058) | -0.005 |
| $\underline{p\left(y_{1}>l_{1}^{*} \mid y_{0} \leq l_{0}\right)}$ | M | 0.040 | (0.010) | 0.007 | (0.005) | 0.033 | 0.026 | $(0.010)$ | $0.014$ |
|  | L | 0.040 | (0.010) | 0.030 | (0.012) | 0.010 | 0.042 | (0.014) | -0.002 |

Table B. 3
Estimates of Different Population Characteristics and Their Standard Errors Obtained From the Complete Data Set and Under 30\% Attrition in the Second Wave

| Parameter | Type of Estimator | No Attrition |  | Attrition Adjustment |  |  |  |  |  |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
|  |  |  |  | MCAR |  |  | MAR |  |  |
|  |  | $\hat{\theta}$ | s.e. | $\hat{\theta}$ | s.e. | 人 | $\hat{\theta}$ | s.e. | $\hat{\Delta}$ |
| Quantiles |  |  |  |  |  |  |  |  |  |
| $M_{0}($ median, $t=0)$ | M | 29,300 | $(1,000)$ | 29,300 | $(1,000)$ | 0 | 29,300 | $(1,000)$ | 0 |
|  | L | 29,300 | $(1,000)$ | 32,000 | (900) | -2,700 | 29,300 | $(1,000)$ | 0 |
| $M_{1}($ median, $t=1)$ | M | 28,600 | $(1,100)$ | 31,200 | (800) | -2,600 | 28,600 | $(1,000)$ | 0 |
|  | L | 28,600 | $(1,000)$ | 31,300 | (900) | -2,700 | 28,600 | $(1,100)$ | 0 |
| Proportions |  |  |  |  |  |  |  |  |  |
| $p\left(y_{0} \leq l_{0}, y_{1} \leq l_{1}\right)$ | M | 0.156 | (0.01) | 0.04 | (0.011) | 0.116 | 0.080 | (0.014) | 0.076 |
|  | L | 0.156 | (0.01) | 0.07 | (0.006) | 0.086 | 0.080 | (0.016) | 0.076 |
| $p\left(y_{0} \leq l_{0}, y_{1} \geq l_{1}^{*}\right)$ | M | 0.007 | (0.002) | 0.001 | (0.001) | 0.006 | 0.004 | (0.002) | 0.003 |
|  | L | 0.007 | (0.002) | 0.003 | (0.001) | 0.004 | 0.004 | (0.002) | 0.003 |
| $p\left(y_{0} \geq l_{0}^{*}, y_{1} \leq l_{1}\right)$ | M | 0.011 | (0.002) | 0.015 | (0.003) | -0.005 | 0.013 | (0.002) | -0.002 |
|  | L | 0.011 | (0.002) | 0.012 | (0.003) | -0.001 | 0.013 | (0.002) | -0.002 |
| $p\left(y_{0} \geq l_{0}^{*}, y_{1} \geq l_{1}^{*}\right)$ | M | 0.790 | (0.04) | 0.874 | (0.037) | -0.084 | 0.829 | (0.049) | -0.039 |
|  | L | 0.790 | (0.04) | 0.864 | (0.038) | -0.074 | 0.828 | (0.051) | -0.038 |
| Conditional Rates |  |  |  |  |  |  |  |  |  |
| $p\left(y_{1} \leq l_{1} \mid y_{0} \leq l_{0}\right)$ | M | 0.923 | (0.023) | 0.245 | (0.045) | 0.678 | 0.465 | (0.071) | 0.458 |
|  | L | 0.923 | (0.023) | 0.885 | (0.045) | 0.038 | 0.930 | (0.099) | -0.007 |
| $p\left(y_{1}>l_{1}^{*} \mid y_{0} \leq l_{0}\right)$ | M | 0.040 | (0.010) | 0.008 | (0.006) | 0.032 | 0.022 | (0.012) | 0.018 |
|  | L | 0.040 | (0.010) | 0.037 | (0.016) | 0.003 | 0.044 | (0.017) | -0.004 |

Charts B
Sensitivity Measures Ratio of the Mixed and Longitudinal Estimators

1. $10 \%$ attrition rate in the second wave

2. $20 \%$ attrition rate in the second wave;

3. $30 \%$ attrition rate in the second wave



#### Abstract

APPENDIX C


We found the empirical bias of the mixed estimates (under a MAR adjustment model) of the conditional rate of remaining low income in 1994 given that the individual was low income in 1993, statistically significant when performing a conservative test of the form

$$
\left(\hat{\theta}_{0}-\hat{\theta}_{\text {mixed }}\right) / \sqrt{\operatorname{var}\left(\hat{\theta}_{\text {mixed }}\right)+\operatorname{var}\left(\hat{\theta}_{0}\right)} .
$$

We found the empirical bias of the mixed estimates of the conditional rate of having an income higher than the LIM in 1994 given that the individual was low income in 1993, (under a MAR adjustment model) non-significant when performing a "radical" test of the form

$$
\left(\hat{\theta}_{0}-\hat{\theta}_{\text {mixed }}\right) / \text { s.e. }\left(\hat{\theta}_{\text {mixed }}\right) .
$$

Similarly, we found the empirical bias of the longitudinal estimates (under a MAR adjustment model) of both conditional rates, non-significant when performing the same type of "radical" test as above, i.e., when assuming that the estimate under no attrition is non-stochastic.

These results hold for $10 \%, 20 \%$ and $30 \%$ attrition rates.

## REFERENCES

BINDER, D., and KOVAČEVIĆ, M.S (1995). Estimating some measures of income inequality from survey data: An application of the estimating equation approach. Survey Methodology, 21, 137-145.

KOVAČEVIĆ, M.S., and YUNG, W. (1997). Variance estimation for measures of income inequality and polarization - An empirical study. Survey Methodology, 23, 41-52.
KOVAR, J.G., RAO, J.N.K., and WU, C.F.J. (1988). Bootstrap and other methods to measure errors in survey estimates. The Canadian Journal of Statistics, 16, 25-45.
LAVALLÉE, P., and HUNTER, L. (1992). Weighting for the survey of labour and income dynamics. In Proceedings: Symposium 92, Design and Analysis of Longitudinal Surveys, Statistics Canada.
LEPKOWSKI, J.M. (1989). Non-response adjustments for wave nonresponse. In Panel Surveys, (Eds. D. Kasprzyk, et al.). New York: John Wiley and Sons.

RAO, J.N.K., and WU, C.F.J. (1988). Resampling inference with complex survey data. Journal of American Statistical Association, 83, 231-241.
RIZZO, L., KALTON, G., and BRICK, J.M. (1996). A comparison of some weighting adjustment methods for panel nonresponse. Survey Methodology, 22, 43-53.
RUBIN BLEUER, S. (1996). Gross flows estimation from longitudinal administrative data with missing waves. Proceedings of the Section on Survey Research Methods, American Statistical Association, II, 681-686.
SHAO, J., and RAO, J.N.K. (1993). Standard errors for low income proportions estimated from stratified multi-stage samples. Sankhyā, B, 55, 393-414.
SHAO, J., and SITTER, R. (1996). Bootstrap for imputed data, Journal of the American Statistical Association, 91, 435, 12781288.

TAMBAY, J-L., SCHIOPU-KRATINA, I., MAYDA, J., STUKEL, D., and NADON, S. (1998). Treatment of nonresponse in cycle two of the National Population Health Survey. Survey Methodology, 24, 147-156.

# Utilising Longitudinally Linked Data from the British Labour Force Survey 

PAM F.TATE ${ }^{1}$


#### Abstract

The British Labour Force Survey (LFS) uses a rotating sample design, with each sample household retained for five consecutive quarters. Linking together the information on the same persons across quarters produces a potentially very rich source of longitudinal data. There are however serious risks of distortion in the resuits from such longitudinal linking, mainly arising from sample attrition, and from response errors, which can produce spurious flows between economic activity states. This paper describes the initial results of investigations by the Office for National Statistics (ONS) into the nature and extent of the problems.


KEY WORDS: Longitudinal data; Labour Force Survey; Economic activity; Attrition bias; Response error.

## 1. INTRODUCTION

The British Labour Force Survey (LFS) is a household survey, gathering information on a wide range of labour force characteristics and related topics. Since 1992 it has been conducted on a quarterly basis, with each sample household retained for five consecutive quarters, and a fifth of the sample replaced each quarter.

The survey is designed to produce cross-sectional data, but in recent years it has been recognised that linking together data on each individual across quarters could produce a rich source of longitudinal data, the uses of which include estimation of labour force gross flows.

The process of linking information on the same individual from different quarters in the LFS is relatively straightforward. However, there are methodological problems which pose serious risks of distortion in the results from this new, hitherto untested use of LFS data. Similar problems have been identified in other countries' labour force surveys, but there are as yet no generally accepted methods of dealing with them. The Office for National Statistics (ONS) has therefore undertaken a programme of work to address this issue.

This paper describes the results so far of investigations into the nature and extent of the problems, and the proposed methods of dealing with them. The issues fall into two main groups: biases arising from sample attrition and related factors; and biases arising from response errors, particularly their effects in producing spurious flows between economic activity states. These are considered in turn.

## 2. SAMPLE ATTRITION AND ITS BIASING EFFECTS

Some sample members are lost at the initial stage, because of nonresponse in the first interview, either because it has not been possible for them to be contacted during the
narrow time window available, or because they have refused to be interviewed. After that, further sample members are lost from each successive quarterly interview round, either because they have moved house (the basic sampling unit for this survey being the dwelling), or because it proves impossible to contact them or they refuse to continue. All these groups of people are, in different ways, atypical of the population as a whole, so their loss from the sample can introduce biases.

Some of these biases are compensated for in the course of applying the normal LFS weighting procedure, which produces population level estimates which are consistent with census-based control totals by sex, age group and region. This process will compensate for biases arising, at all stages of the survey, from differential attrition by sex, age and region. However, biases in other characteristics which are not themselves used in the weighting procedure will not be compensated for (and may even be increased) in that process, except when they are related to age, sex or region, in such a way that the bias is caused entirely by the under- or overrepresentation of particular age, sex or region categories.

Work on this subject therefore looked first at what characteristics are more or less represented in the LFS sample than the whole population, and in different waves of the LFS sample. (Each quarter, the sample is made up of five waves, the people in the first wave having their first interview, those in the second wave their second interview, and so on.) It then examined whether and to what extent these characteristics are related to each other, and whether it is possible to define a set of variables which characterise those people who are likely to be under-represented.

## 3. CHARACTERISTICS OF NON-RESPONDENTS

Analysis of the proportions which could not be linked to the next quarter, by wave, for key demographic and

[^12]economic variables (Table 1 gives an illustration for broad age/sex groups), showed that, consistently across all waves, there is a greater propensity to be under-represented for young people aged 18 to 29 (and especially 18 to 24), single people, those living in London, people in rented accommodation (especially privately rented), the unemployed, and those in temporary employment. Most of these characteristics have also been found (by Foster (1994) in a study which linked data from the 1991 Census with non-responding LFS sample households) to be associated with high non response at the first interview, particularly young adults, single people, one person households, and those living in London.

Table 1
Percentage of Unlinked Cases by Sex and Age Group by Wave

|  <br> category | Unlinked percentage <br> Wave 1 | Wave 4 |
| :---: | :---: | :---: |
| All persons | 8.6 | 4.8 |
| AGE \& SEX |  |  |
| Male | 8.6 | 4.9 |
| $15-17$ | 6.8 | 4.9 |
| $18-29$ | 13.8 | 10.2 |
| $30-44$ | 6.8 | 4.1 |
| $45-64$ | 7.2 | 2.5 |
| Female | 8.7 | 4.7 |
| $15-17$ | 5.9 | 3.0 |
| $18-29$ | 13.9 | 10.9 |
| $30-44$ | 6.3 | 2.8 |
| $45-59$ | 7.5 | 2.2 |

Note: More detailed analyses are available from the author.
Several of the characteristics of those who are lost to the sample appear likely to be related, and this was investigated in the first instance using logistic regression. The variables identified as being independently associated with whether the cases were lost from the sample, were found to be largely consistent for the four waves. In each case they included age group, marital status, tenure, (i.e. whether the accommodation was owned, rented from a private landlord, or rented from a local authority or housing association), qualification level, and a combined economic activity variable incorporating broad economic activity (employed, unemployed or inactive), and, for the employed, employment status, part-time/full-time and temporary/permanent. Region was found to be independently associated in only two of the four waves, and sex in none.

For the five variables consistently appearing for all waves, there was a good degree of consistency conceming which categories were associated with sample attrition. Table 2 gives the multiplying factors for the odds ratio for all categories with a consistent association with increasing attrition. Being in the younger age groups, between 18 and

29 (and especially 18 to 24 ), has a particularly strong effect, as does being in privately rented accommodation. Being single (i.e., never married and not cohabiting) has a moderate association. There are no consistent associations with particular categories of economic activity or qualification level, except for a slight one with full-time, temporary employees. The effect of region is not consistent even for the two waves in which it appears.

Table 2
Multiplying Factors for Odds Ratios Categories Associated with High Attrition

|  | Multiplying factor for odds ratio |  |  |  |  |
| :--- | :---: | :---: | :---: | :---: | :---: |
| Variable | Category | Wave 1 | Wave 2 | Wave 3 | Wave 4 |
| Age Group | $18-19$ | 1.89 | 2.56 | 2.86 | 1.92 |
|  | $20-24$ | 1.79 | 2.08 | 2.10 | 2.83 |
|  | $25-29$ | 1.17 | 1.30 | 1.44 | 1.55 |
| Tenure | Privately <br> rented | 2.12 | 1.52 | 1.86 | 2.29 |
| Marital | Single | 1.25 | $(1.12)$ | 1.27 | 1.49 |
| Status | Economic <br> Ectivity/ <br> status | Employee, <br> full time, <br> temp | $(1.12)$ | $(1.36)$ | $(1.13)$ |

Note: () indicates coefficient is not significant at $5 \%$ level.
The logistic regression analysis performed did not allow for interactions between the variables, and to investigate this possibility a further analysis was performed, using the CHAID module of SPSS to produce a segmentation of the data set into groups which have as great a variation as possible with respect to the proportion of unlinked cases. The results of this were however very similar to those of the logistic regression analysis. Overall, the main characteristics independently associated with a high proportion of sample loss were the younger adult age groups (18 to 29, especially 18 to 24 ) and living in privately rented accommodation, with some relatively minor additional effects of being in temporary employment for the youngest age groups. Separate analyses of the characteristics of those sample members who had been lost through moving away, and those lost through non-contact (or, more rarely, refusal) produced similar results.

## 4. COMPENSATING FOR ATTRITION BIAS

The analysis so far has been directed at the biasing effect of sample attrition on the cross-sectional characteristics of the longitudinal sample, and has identified the characteristics independently associated with greater nonresponse. A possible approach to compensating for the bias arising from this is to incorporate tenure as well as age into the
weighting procedure for the longitudinal data. This is being explored, using a calibration approach with CALMAR software, and including prior weights derived from the work described above to compensate for differential nonresponse by tenure.

However, there may be a problem which would limit the effectiveness of this approach. The propensity to not respond may be directly dependent on the unobserved labour force status of the individual, and possibly independently of their observed characteristics. Nonresponse of this kind is an example of non-ignorable nonresponse (Rubin 1976), and its presence would imply that estimates of the important measures of labour force gross flows would be biased even after the application of a weighting process of the type being explored.

There are two indirect approaches which give some indication of whether non-ignorable nonresponse might be a problem for longitudinal LFS data. One is to investigate whether the proportion of the gross flows in the sample which are transitions between different economic activity states systematically decrease (or increase) from wave 1-2 to wave $4-5$; if so, this would suggest that people changing from one state to another are more (or less) likely to be nonrespondents than those in a stable state. However, Table 3 shows that there is no consistent systematic pattern across waves - though this does not exclude the possibility of other pattems of differential nonresponse by labour force flows category.

Table 3
Percentage of Transitions Between Different Economic Activity States by Wave for Pairs of Adjacent Quarters

| Data Set | Percentages |  |  |  |  |
| :--- | :---: | :---: | :---: | :---: | :---: |
|  | Wave <br> $1-2$ | Wave <br> $2-3$ | Wave <br> $3-4$ | Wave <br> $4-5$ |  |
| Summer/autumn 1995 | 8.0 | 7.3 | 7.3 | 7.1 |  |
| Autumn 1995/ <br> winter 1995-1996 | 7.2 | 6.7 | 6.5 | 6.5 |  |
| Summer/ autumn 1996 | 7.6 | 7.0 | 7.3 | 7.5 |  |
| Autumn 1996/ <br> winter 1996-1997 | 6.8 | 6.5 | 6.2 | 6.5 |  |

Another possibility is that people moving addresses (and thereby lost to the LFS sample) may have a different pattern of labour force flows than the rest of the population. We do not have any information on the people who have moved away, but we do know something about the people who have moved into the sample addresses from elsewhere. These movers-in can reasonably be taken to represent the movers-out, since they are equally samples from the same population of movers (ignoring the possible effects of the small proportion of international moves). Table 4 shows the distribution of the linked sample (all adults whose records were able to be matched) and of the identifiable
movers-in for a pair of adjacent quarters in 1995. (It should however be noted that the flows categories are not strictly comparable, since the previous economic activity state for the movers-in is obtained by retrospective reporting.) It is clear that the sample of movers does differ, with a lower proportion in stable inactivity, and a higher proportion in all the other flows categories, and in particular a greater proportion of people changing their economic activity state; but that the movers make up such a small proportion overall that the effect on the whole sample is negligible.

Table 4
Gross Flows for Movers - in Compared with Linked Sample

| Activity states | Linked sample <br> $(\%)$ | Movers-in <br> $(\%)$ | Linked + movers <br> $(\%)$ |
| :--- | :---: | :---: | :---: |
| EE | 55.1 | 56.9 | 55.1 |
| EU | 0.8 | 1.5 | 0.8 |
| EN | 1.1 | 1.6 | 1.1 |
| UE | 1.0 | 1.7 | 1.0 |
| UU | 2.9 | 6.5 | 3.0 |
| UN | 0.7 | 1.1 | 0.7 |
| NE | 1.2 | 2.5 | 1.3 |
| NU | 1.0 | 2.3 | 1.1 |
| NN | 36.2 | 26.0 | 35.9 |
| All transitions | 5.9 | 10.6 | 6.0 |
| TOTAL (no.) | 80,664 | 1,790 | 82,454 |

Note: E represents in employment
$\quad \mathrm{U}$ represents ILO unemployed
$\quad \mathrm{N}$ represents economically inactive
hence EE represents in employment at both quarters
EU represents in employment then ILO unemployed etc.
These indirect approaches do not indicate any very strong effect of non-ignorable nonresponse, but they do not rule it out. This possibility is therefore being investigated by work involving the modelling of nonresponse in the LFS.

## 5. RESPONSE ERROR AND ITS BIASING EFFECTS

All surveys in general, and household surveys in particular, are subject to response error, when the information given by the respondent is not an accurate reflection of the actuality. This may occur for a variety of reasons - the respondent may misunderstand the question; the interviewer may misunderstand or misrecord the response; the respondent may not know or remember the correct answer; or the respondent may knowingly give an incorrect answer for reasons of embarrassment, prestige, fear of breach of confidentiality or a wish to give the "expected" answer.

In the field of labour force surveys it has generally been found (for an overview see Lemaître 1994) that, for cross-sectional data, there is no particular tendency for the errors to be systematic, so that on average they tend to cancel out. However, for longitudinal data produced by linking together data collected on the same person at different points in time, this cancellation may not occur.

In particular, this is likely to be the case for data on gross flows between economic activity states. The numbers of people who move from one state (in employment, unemployed, economically inactive) to another during the relatively short period usually considered (a month, a quarter, or perhaps a year) are small compared with the numbers of people who remain in the same state. A response error at one point of time is much more likely to lead to an apparent change of state when the true situation is one of stability, than the reverse. Thus response errors are likely to have a very disproportionate effect in upwardly biasing flows between reported states. In the LFS, they may arise from the use of proxy respondents, where one person answers questions on behalf of someone else in the same household; and from respondent errors. We will consider these in turn.

## 6. THE EFFECT OF PROXY RESPONDENTS

To investigate the effect of proxy respondents, we need to look at the distribution of activity states at the two quarters according to whether the first quarter's interview was in person or by proxy, and whether the second quarter's interview was in person or by proxy. Very young adults under 20 are both exceptionally likely to be represented by proxies and also likely to be particularly volatile in terms of their economic activity category, and so may distort any relationship between these two factors. Table 5 therefore shows the distribution of activity states at the two quarters, for men aged 20 to 64 and women aged 20 to 59 . There is a higher proportion of transitions for personal followed by proxy interviews than for personal at both quarters, but proxy followed by personal interviews show only a very slightly higher proportion than personal at both quarters. Thus switching between proxy and personal interviews does not show a consistently greater proportion of transitions. Cases with both interviews by proxy have the lowest proportion of transitions of all, and the inclusion of these brings the overall proportion to a level consistent with that for personal interviews at both quarters. Thus there do appear to be differences between the various combinations of interview types, which merit further investigation, but in the LFS the use of proxy respondents does not of itself produce an exaggerated estimate of gross flows.

Table 5
Percentage of Transitions by Interview Type

| Interview <br> type | Men (20-64) |  | Women (20-59) <br> Sample no. |  |
| :--- | :---: | :---: | :---: | :---: |
| \% trans. | Sample no. | \% trans. |  |  |
| Personal/ <br> personal | 14,527 | 5.2 | 19,582 | 7.3 |
| Personal/ <br> proxy | 2,044 | 7.0 | 1,597 | 8.3 |
| Proxy/ <br> personal | 2,214 | 5.4 | 1,632 | 7.6 |
| Proxy/ <br> proxy | 8,602 | 4.7 | 4,206 | 5.4 |
| All | 27,387 | 5.2 | 27,017 | 7.1 |

## 7. RESPONDENT ERRORS

By their nature, respondent errors are impossible to identify directly, (except perhaps by re-interview, and even then there may be doubt about what is the correct answer). It is however sometimes possible to identify internal inconsistencies in the survey data, which may indicate response error. In the LFS, respondents who are in employment, and respondents who are unemployed, are asked how long they have been in that state. If the period is greater than three months, but they stated in the previous quarter that they were in a different state, there is an inconsistency which may indicate a false transition between economic activity states.

Table 6 shows the percentage of inconsistencies for various kinds of transitions - these are high throughout. Transitions from economic inactivity produce the highest percentages, especially when the transition is into unemployment. (There are no large or consistent differences between the different subcategories of the inactive.) Separating those in employment into part-time and full-time shows that there is a consistent pattern of a greater proportion of inconsistencies for part time employment, and similar but less pronounced results were found for the self-employed.

Table 6
Percentage of Inconsistencies by Transition Type

| Transition Type | Percentage of inconsistencies |  |  |
| :--- | :---: | :---: | :---: |
|  | All <br> (\%) | Full time <br> $(\%)$ | Part time <br> $(\%)$ |
|  | 8.7 | 7.8 | 12.2 |
| Inactive to Employment | 26.2 | 18.1 | 30.4 |
| Employment to Unemployment | 18.7 | 14.7 | 23.3 |
| Inactive to Unemployment | 49.5 |  |  |
| All | 23.9 |  |  |

It is possible that the inconsistencies may have arisen through errors in the reported length of time in the economic activity state at the second quarter, rather than in the initial state at the first quarter. The distribution of the length of time does not however show heaping at around four to five months (as would be expected in the case of errors in the duration data). Also, the duration data reported in consecutive quarters for people in a stable state were found to be very consistent. These findings tend to suggest, though the evidence is indirect and by no means conclusive, that the errors are more likely to be in the reporting of economic activity at one or other of the interviews. This is not the only possibility - for example, it may be that some respondents have correct transition data, but incorrect duration data through using an interpretation of their past economic activity which is not consistent with the standard definitions applied to the reporting of their current state but the findings so far suggest that it is likely to be the most common.

Some light on which of the inconsistent categories is correct may emerge by looking at the pattern of responses over three interviews. Table 7 shows the proportions of each group of inconsistent transitions from one quarter to the next which are followed by each economic activity category in the third quarter. (All relevant waves are combined in order to obtain reasonable sample sizes.) It is clear that of the transitions into employment in the second quarter, the great majority remain in that category in the third quarter. The transitions into unemployment show a much more mixed pattern, with a little over half remaining in unemployment, but a substantial group of about 30 to 40 per cent reverting to the state reported in the first quarter. It is noteworthy that scarcely any of the transitions from the second to the third quarter for this group were found to have a repeated inconsistency between the transition and the reported duration data. The results so far suggest that, in the case of an inconsistent transition into employment, that is likely to be the correct state, but more investigation is needed to achieve further clarification.

Table 7
Percentages of Inconsistent Transitions by Economic Activity at Following Quarter

| Transition <br> type | Total <br> inconsis- <br> tent | Activity state in next quarter <br> Employed <br> (\%) | Unempl. <br> (\%) | Inactive <br> $(\%)$ |
| :--- | :---: | :---: | :---: | :---: |
| Unempl. to <br> Employment | 60 | 90 | 7 | 3 |
| Inactive to | 159 | 79 | 4 | 17 |
| Employment |  | 39 | 53 | 8 |
| Employment <br> to Unempl. | 87 | 39 | 55 | 28 |
| Inactive to | 229 | 17 | 55 |  |

## 8. ADJUSTING FOR RESPONSE ERROR BIAS

It is clear from the above that there is likely to be a substantial level of response error affecting the raw data on gross flows. Work on adjusting for such errors has so far been largely confined to the USA and Canada. A review of three methods proposed for USA data is given by Flaim and Hogue (1985), and a later proposal for Canadian data is given in Singh and Rao (1995), but to date, to the author's knowledge no official adjusted gross flows data are being published, though several countries are publishing unadjusted data while drawing attention to their limitations. The adjustment methods so far proposed all rely on assumptions about the nature of the errors which seem unlikely to be met in practice - either full independence of the classification errors or very limited departures from that assumption. (See Lemaitre 1994 for a review of problems with these adjustment methods.)

It seems worthwhile to explore different routes to the development of methods of adjustment or compensation for response error bias. As a first stage, work is continuing on the investigation of the characteristics and circumstances of cases of inconsistency, and of other possible ways of identifying false transitions. It is also proposed to investigate the circumstances of people giving inconsistent responses of the kind analysed above, by means of more detailed follow-up interviews. This should provide better indications of the extent to which the inconsistencies do represent response error, and may provide results useful for both reducing, and adjusting for, response error. Both these strands will provide inputs to a third element of the forward programme, in which it is proposed to develop models of classification error in reporting economic activity.

## ACKNOWLEDGEMENTS

The author wishes to thank the editor and two referees for their helpful comments.

## REFERENCES

FLAIM, P. O., and HOGUE, C. R. (1985). Measuring labor force flows: a special conference examines the problems. Monthly Labor Review, July 1985, U.S. Bureau of Labor Statistics.
FOSTER, K. (1994). The Labour Force Survey - Report of the 1991 Census-linked Study of Survey Nonrespondents. Office of Population Censuses and Surveys.
LEMAITTRE, G. (1994). Data on Labour Force Dynamics from Labour Force Surveys. Organisation for Economic Co-operation and Development.
RUBIN, D.B. (1976). Inference and missing data. Biometrika, 63, 581-592.
SINGH, A.C., and RAO, J.N.K. (1995). On the adjustment of gross flow estimates for classification error with application to data from the Canadian Labour Force Survey. Journal of the American Statistical Association, 90, 478-488.

# A Model Based Justification of Kish's Formula for Design Effects for Weighting and Clustering 

SIEGFRIED GABLER, SABINE HAEDER and PARTHA LAHIRI ${ }^{1}$


#### Abstract

In this short note, we demonstrate that the well-known formula for the design effect intuitively proposed by Kish has a model-based justification. The formula can be interpreted as a conservative value for the actual design effect.


KEY WORDS: Cluster size; Intraclass correlation coefficient; Selection probabilities.

## 1. INTRODUCTION

We consider multistage, clustered, sample designs where each observation belongs to a weighting class. For example, the clusters are blocks which are selected proportional to the number of its households. Within each block the same number of households is selected with equal probabilities. A randomly chosen person of the household has to be interviewed. Then, the household sizes determine the weighting classes. Kish (1987) proposed the following formula for determining the design effect in order to incorporate the effects due to both weighting needed to counter unequal selection probabilities, and clustered selection:

$$
\operatorname{deff}_{\mathrm{Kish}}=m \frac{\sum_{i=1}^{I} w_{i}^{2} m_{i}}{\left(\sum_{i=1}^{l} w_{i} m_{i}\right)^{2}}[1+(\bar{b}-1) \rho]
$$

where $m_{i}$ and $w_{i}$ denote the number of observations and the weight attached to the $i$-th weighting class ( $i=1, \ldots, I$ ), $m=\sum_{i=1}^{J} m_{i}$, the total sample size, $\bar{b}$ is the average cluster size and $\rho$ is the intraclass correlation coefficient. Kish's formula is very intuitive and novel, but he said that his "treatment may be incomplete and imperfect."

Kish's formula is now used by many survey samplers. In fact, the above formula will be used in the sample size determination in the European Social Surveys to be conducted by its member countries. The purpose of this note is to provide a model-based justification for using Kish's formula.

## 2. A MODEL BASED JUSTIFICATION OF KISH'S FORMULA

Let $m_{i c}$ be the number of observations in the $c$-th sampled cluster belonging to the $I$-th weighting class
$(i=1, \ldots, I ; c=1, \ldots, C)$. Then $m_{i}=\sum_{c=1}^{C} m_{i c}$, the number of observations in the $I$-th weighting class. Let $b_{c}=$ $\sum_{i=1}^{i} m_{i c}$, the number of observations in the $c$-th cluster $(i=1, \ldots, I ; c=1, \ldots, C)$ so that $\bar{b}=C^{-1} \sum_{c=1}^{C} b_{c}$. Let $y_{c j}$ and $w_{c j}$ be the observation and the weight for the $j$-th sampling unit in the $c$-th cluster ( $c=1, \ldots, C ; j=1, \ldots, b_{c}$ ). The usual design-based estimator for the population mean is defined as

$$
\bar{y}_{w}=\frac{\sum_{c=1}^{c} \sum_{j=1}^{b_{c}} w_{c j} y_{c j}}{\sum_{c=1}^{c} \sum_{j=1}^{b_{c}} w_{c j}}
$$

To justify Kish's formula, we assume the following model:

$$
\begin{align*}
& \operatorname{Var}\left(y_{c j}\right)=\sigma^{2} \text { for } c=1, \ldots, C ; j=1, \ldots, b_{c} \\
& \operatorname{Cov}\left(y_{c j^{\prime}}, y_{c^{\prime} j^{\prime}}\right)= \begin{cases}\rho \sigma^{2} & \text { if } c=c^{\prime} ; j \neq j^{\prime} \\
0 & \text { otherwise } .\end{cases} \tag{1}
\end{align*}
$$

The above model is appropriate to account for the cluster effect and was used earlier by others (see, e.g., Skinner, Holt and Smith 1989). We shall then define design effect as deff $=\operatorname{Var}_{1}\left(\bar{y}_{w}\right) / \operatorname{Var}_{2}(\bar{y})$, where $\operatorname{Var}_{1}\left(\bar{y}_{w}\right)$ is the variance of $\bar{y}_{w}$ under model (1) and $\operatorname{Var}_{2}(\bar{y})$ is the variance of the overall sample mean $\bar{y}$, defined as $\sum_{c=1}^{C} \sum_{j=1}^{b_{c}} y_{c j} / m$, computed under the following model:

$$
\begin{align*}
& \operatorname{Var}\left(y_{c j}\right)=\sigma^{2} \text { for } c=1, \ldots, C ; j=1, \ldots, b_{c} \\
& \operatorname{Cov}\left(y_{c j}, y_{c j^{\prime}}\right)=0 \text { for all }(c, j) \neq\left(c^{\prime}, j^{\prime}\right) \tag{2}
\end{align*}
$$

Note that model (2) is appropriate under simple random sampling and provides the usual formula $\sigma^{2} / m$ for $\operatorname{Var}_{2}(\bar{y})$.

[^13]Now, turning our attention to $\operatorname{Var}_{1}\left(\bar{y}_{w}\right)$, first note that

$$
\begin{align*}
& \operatorname{Var}_{1}\left(\sum_{c=1}^{c} \sum_{j=1}^{b_{c}} w_{c j^{\prime}} y_{c j}\right) \\
= & \sum_{c=1}^{c} \operatorname{Var}\left(\sum_{j=1}^{b_{c}} w_{c j} y_{c j}\right) \\
= & \sum_{c=1}^{c}\left\{\sum_{j=1}^{b_{c}} w_{c j}^{2} \operatorname{Var}\left(y_{c j}\right)+\sum_{j \neq j}^{b_{c}} w_{c j^{\prime}} w_{c j^{\prime}} \operatorname{Cov}\left(y_{c j^{\prime}}, y_{c j^{\prime}}\right)\right\} \\
= & \sigma^{2} \sum_{c=1}^{c}\left\{\sum_{j=1}^{b_{c}} w_{c j}^{2}+\rho \sum_{j \neq j^{\prime}}^{b_{c}} w_{c j^{\prime}} w_{c j^{\prime}}\right\} \\
= & \sigma^{2}\left\{\sum_{i=1}^{I} w_{i}^{2} m_{i}+\rho \sum_{c=1}^{c}\left(\sum_{i=1}^{1} w_{i} m_{i c}\right)^{2}-\rho \sum_{i=1}^{l} w_{i}^{2} m_{i}\right\} \tag{3}
\end{align*}
$$

$\operatorname{since} \sum_{c=1}^{C} \sum_{j=1}^{b_{c}} w_{c j}^{2}=\sum_{i=1}^{I} w_{i}^{2} m_{i}$ and

$$
\begin{aligned}
\sum_{c=1}^{c} \sum_{j \neq j^{\prime}}^{b_{c}} w_{c j} w_{c j^{\prime}} & =\sum_{c=1}^{c}\left\{\left(\sum_{j=1}^{b_{c}} w_{c j}\right)^{2}-\sum_{j=1}^{b_{c}} w_{c j}^{2}\right\} \\
& =\sum_{c=1}^{c}\left\{\left(\sum_{i=1}^{1} w_{i} m_{i c}\right)^{2}-\sum_{i=1}^{1} w_{i}^{2} m_{i c}\right\} \\
& =\sum_{c=1}^{c}\left(\sum_{i=1}^{1} w_{i} m_{i c}\right)^{2}-\sum_{i=1}^{1} w_{i}^{2} m_{i}
\end{aligned}
$$

Noting that $\sum_{c=1}^{C} \sum_{j=1}^{b_{c}} w_{c j}=\sum_{i=1}^{l} w_{i} m_{i}$, we have

$$
\begin{aligned}
& \operatorname{Var}_{1}\left(\bar{y}_{w}\right) \\
& =\frac{\operatorname{Var}_{1}\left(\sum_{c=1}^{c} \sum_{j=1}^{b_{c}} w_{c j} y_{c j}\right)}{\left(\sum_{i=1}^{l} m_{i} w_{i}\right)^{2}} \\
& =\frac{\sigma^{2}\left\{\sum_{i=1}^{l} w_{i}^{2} m_{i}+\rho \sum_{c=1}^{c}\left(\sum_{i=1}^{l} w_{i} m_{i c}\right)^{2}-\rho \sum_{i=1}^{l} w_{i}^{2} m_{i}\right\}}{\left(\sum_{i=1}^{l} w_{i} m_{i}\right)^{2}}
\end{aligned}
$$

so that

$$
\begin{equation*}
\operatorname{deff}=m \frac{\sum_{i=1}^{1} w_{i}^{2} m_{i}}{\left(\sum_{i=1}^{1} w_{i} m_{i}\right)^{2}}\left[1+\left(b^{*}-1\right) \rho\right], \tag{4}
\end{equation*}
$$

where $b^{*}=\sum_{c=1}^{C}\left(\sum_{i=1}^{l} w_{i} m_{i c}\right)^{2} / \sum_{i=1}^{I} w_{i}^{2} m_{i}$.
Using the Cauchy-Schwarz inequality, we get

$$
\begin{gathered}
\left(\sum_{i=1}^{1} w_{i} m_{i c}\right)^{2}=b_{c}^{2}\left(\sum_{i=1}^{l} w_{i} \frac{m_{i c}}{b_{c}}\right)^{2} \leq b_{c}^{2} \sum_{i=1}^{l} w_{i}^{2} \frac{m_{i c}}{b_{c}} \\
=b_{c} \sum_{i=1}^{l} w_{i}^{2} m_{i c}
\end{gathered}
$$

so that

$$
\begin{equation*}
b^{*} \leq \frac{\sum_{c=1}^{c} b_{c} \sum_{i=1}^{1} w_{i}^{2} m_{i c}}{\sum_{c=1}^{c} \sum_{i=1}^{1} w_{i}^{2} m_{i c}}=\bar{b}_{w} \text {, say. } \tag{5}
\end{equation*}
$$

Thus (4) and (5) yield

$$
\begin{equation*}
\operatorname{deff} \leq m \frac{\sum_{i=1}^{I} w_{i}^{2} m_{i}}{\left(\sum_{i=1}^{l} w_{i} m_{i}\right)^{2}}\left[1+\left(\bar{b}_{w}-1\right) p\right] \tag{6}
\end{equation*}
$$

Note that $\bar{b}_{w}$ can be interpreted as an average (weighted) cluster size. If $\bar{b}_{w}$ is equal to $\bar{b}$, e.g., if all $b_{c}$ are equal, the upper bound of deff is simply Kish's formula. Thus Kish's formula serves as a conservative value for the actual design effect.

## ACKNOWLEDGEMENTS

The authors are thankful to the editor and the referees for their remarks which led to an improvement of the paper. The work was completed while the last author was a Guest Professor at ZUMA, the Center for Survey Research and Methodology, Mannheim, Germany.

## REFERENCES

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

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

## JOURNAL OF OFFICIAL STATISTICS

## An International Review Published by Statistics Sweden

JOS is a scholarly quarterly that specializes in statistical methodology and applications. Survey methodology and other issues pertinent to the production of statistics at national offices and other statistical organizations are emphasized. All manuscripts are rigorously reviewed by independent referees and members of the Editorial Board
Contents
Volume 14, Number 4, 1998
Introduction to the Special Issue: Disclosure Limitation Methods for Protecting the Confidentiality of Statistical Data Steven E. Fienberg and Leon C.R.J. Willenborg ..... 337
A Database System Prototype for Remote Access to Information Based on Confidential Data Sallie Keller-McNulty and Eiizabeth A. Unger ..... 347
Estimating the Re-identification Risk Per Record in Microdata C.J. Skinner and D.J. Holmes ..... 361
A Bayesian Species-Sampling-Inspired Approach to the Uniques Problem in Microdata Disclosure Risk Assessment Stephen M. Samuels ..... 373
Confidentiality, Uniqueness, and Disclosure Limitation for Categorical Data Stephen E. Fienberg and Udi E. Makov ..... 385
Synthetic and Combined Estimators in Statistical Disclosure Control Jeroen Pannekoek and Ton de Waal ..... 399
Balancing Disclosure Risk Against the Loss of Nonpublication Alan M. Zaslavsky and Nicolas J. Horton ..... 411
Optimal Local Suppression in Microdata
A.G.de Waal and L.C.R.J. Willenborg ..... 421
Models and Methods for the Microdata Protection Problem C.A.J. Hurkens and S.R. Tiourine ..... 437
Masking Microdata Using Micro - Aggregation D. Defays and M.N. Anwar ..... 449
Post Randomisation for Statistical Disclosure Control: Theory and Implementation J.M. Gouweleeuw, P. Kooiman, L.C.R.J. Willenborg and P.-P. de Wolf ..... 463
Comment G. Sande ..... 479
Disclosure Limitation Using Perturbation and Related Methods for Categorical Data Stephen E. Fienberg, Udi E. Makov, and Russell J. Steel ..... 485
Comment Peter Kooiman ..... 503
Rejoinder
Stephen E. Fienberg, Udi E. Makov, and Russell J. Steel ..... 509
Comparision of Systems Implementing Automated Cell Suppression for Economic Statistics N: Kirkendall and G. Sande ..... 513
Using Noise for Disclosure Limitation Establishment Tabular Data Timothy Evans, Laura Zayatz, and John Slanta ..... 537
Experiments with Controlled Rounding for Statistical Disclosure Control in Tabular Data with Linear Constraints Matteo Fischetti and Juan-José Salazar-González ..... 553
Editorial Collaborators ..... 567
Index to Volume 14, 1998 ..... 571

## Volume 27, No. 1, March/mars 1999

## Christian GENEST

Editor's report/Rapport du rédacteur en Chef
Mousumi BANERJEE, Michelle CAPOZZOLI, Laura MCSWEENEY and Debajyoti SINHA
Beyond kappa: A review of interrater agreement measures
L.MANCHESTER, C. A. FIELD and A. MCDOUGALL

Regression for overdetermined systems: A fisheries example

## E.MONGA and S. TARDIF

Asymptotic optimality of a class of rank tests for replicated Latin square designs
Guohua PAN and Winson TAAM
Distribution-free subset selection for incompletely ranked data
Gary SNEDDON
Smoothing in an underdetermined linear model with random explanatory variables

## John R.COLLINS

Robust $M$-estimators of scale: Minimax bias versus maximal variance

## Dongchu SUN and Keying YE

Reference priors for a product of normal means when variances are unknown

## Sonia PETRONE

Bayesian density estimation using Bernstein polynomials

## M. PENSKY and R. S. SINGH

Empirical Bayes estimation of reliability characteristics for an exponential family
M. C. FINOCCHIARO and D. SACCHETTI

The variance function of the natural exponential family generated by a measure on $n$ integers

## John E. KOLASSA

Confidence intervals for parameters lying in a random polygon

## Nabendu PAL and Wooi K. LIM

Second order properties of intraclass correlation estimators for a symmetric normal distribution

## Andrew HEARD and Tim SWARTZ

Extended voting measures

## Michael BARON

Convergence rates of change-point estimators and tail probabilities of the first-passage-time process

## Célestin C. KOKONENDJI

Le problème d'Anscombe pour les lois binomiales négatives généralisées
Helge BLAKER
A class of shrinkage estimators in linear regression
Acknowledgement of referees' services / Remerciements aux membres des jurys
Forthcoming papers/Articles à paraître

## Volume 27, No. 2, June/juin 1999

Andrey FEUERVERGER, John ROBINSON and Augustine WONG I
On the relative accuracy of certain bootstrap procedures

## Biao ZHANG

Bootstrapping with auxiliary information
Steven N. MACEACHERN, Merlise CLYDE and Jun S. LIU
Sequential importance sampling for nonparametric Bayes models: The next generation

## George D. PAPANDONATOS and Seymour GEISSER

Bayesian interim analysis of lifetime data
Paul DAMIEN and Stephen WALKER
A full Bayesian analysis of circular data using the von Mises distribution
Craig A. COOLEY and Steven N. MACEACHERN
Prior elicitation in the classification problem
Schultz CHAN and Malay GHOSH
A geometric optimality of Cox's partial likelihood

## Y. LEE and John A. NELDER

The robustness of the quasilikelihood estimator
Hui CHEN and Joseph P. ROMANO
An invariance principle for triangular arrays of dependent variables with application to autocovariance Estimation

## Kanchan MUKHERJEE

The asymptotic behavior of a class of L-estimators under long-range dependence

## Ross H. TAPLIN

Robust F-tests for linear models

## Ping ZHANG

The optimal prediction of cross-sectional proportions in categorical panel-data analysis

## Dan NETTLETON

Order-restricted hypothesis testing in a variation of the normal mixture model

## K. KRISHNAMOORTHY and Maruthy K. PANNALA <br> Confidence estimation of a normal mean vector with incomplete data

## Marten H. WEGKAMP

Quasi-universal bandwidth selection for kernel density estimators

## Christian GENEST

Probability and statistics: A tale of two worlds?
Forthcoming Papers / Articles à paraître


## GUIDELINES FOR MANUSCRIPTS

Before having a manuscript typed for submission, please examine a recent issue of Survey Methodology (Vol. 19, No. 1 and onward) as a guide and note particularly the points below. Accepted articles must be submitted in machine-readable form, preferably in WordPerfect. Other word processors are acceptable, but these also require paper copies for formulas and figures.

## 1. Layout

1.1. Manuscripts should be typed on white bond paper of standard size ( $81 / 2 \times 11 \mathrm{inch}$ ), one side only, entirely double spaced with margins of at least $11 / 2$ 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 " $\log (\cdot)$ "; 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, $\omega ; 0,0,0 ; 1,1$ ).
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 $\mathrm{a}, \mathrm{b}, \mathrm{c}$ to the year of publication. Journal titles should not be abbreviated. Follow the same format used in recent issues.


[^0]:    1 Hannu Kröger and Ismo Teikari, P.O. Box 3A, Fin-00022, Statistics Finland, Finland; Carl-Erik Särndal, 11 th floor, R.H. Coats Bldg., Statistics Canada, Ottawa, Ontario, K1A 0T6, Canada.

[^1]:    1 William R. Bell and Matthew Kramer, Statistical Research Division, Room 3000-4, U.S. Bureau of the Census, Washington, D.C. 20233-9100, U.S.A.

[^2]:    Figure 2. U.S. 5+ Units Housing Starts: Variance Decomposition after X-11 Seasonal Adjustment. The top panel gives the variance of the seasonally adjusted series as the total of the three components. The second panel give the contribution of sampling error (e), which is the largest component and constant across the series. The third panel gives the contribution of back/forecast error ( $\mathbf{0} \mathbf{0}$ ), which is zero in the middle of the series, where no back/forecasts are needed, but increases towards either end of the series as more back/forecasts are used. The bottom panel is the sum of the two covariance terms $(\operatorname{cov}(\mathbf{e},(\mathbf{b} \mathbf{0} \mathbf{f}))+\operatorname{cov}((\mathbf{b} \mathbf{0}), \mathbf{e}))$, which tend to offset the contribution from back/forecast error.

[^3]:    $\bar{J}$ Jan de Haan, Eddy Opperdoes and Cecile M. Schut, Division Socio-economic Statistics, Statistics Netherlands, P.O. Box 4000, 2270 JM, Voorburg, The Netherlands, e-mail: jhhn@cbs.nl

[^4]:    ${ }^{1}$ Pierre Duchesne, Département de Mathématiques et de Statistique, Université de Montréal, C.P. 6128, succursale centre-ville, Montréal, Québec, H3C 3J7.

[^5]:    ' The limits for the CARVs are $[0.20,32]$ for $n=10$ and $[0.20,16]$ for $n=15$.

[^6]:    ${ }^{1}$ The limits for the CARVs are $[0.20,16]$ for $n=30$ and $[0.20,7]$ for $n=60$.

[^7]:    1 Yves Tille, Crest - Ensai, Ecole Nationale de la Statistique et de l'Analyse de l'Information, rue Blaise Pascal, Campus de Ker Lann, 35170 Bruz, France, e-mail: tille@ensai.fr.

[^8]:    $\overline{1}$ N.G.N. Prasad, Department of Mathematical Sciences, University of Alberta, Edmonton, Alberta, T6G 2Gl; J.N.K. Rao, Department of Mathematics and Statistics, Carleton University, Ottawa, Ontario, K1S SB6.

[^9]:    ${ }^{1}$ Fernando A.S. Moura, Instituto de Matemática, UFRJ, Rio de Janeiro, Brazil, CP: 68530, CEP: 21941-590, e-mail: fmoura@dme.ufrj.br; David Holt, Office for National Statistics, 1 Drummond Gate, London, SWIP 2QQ, e-mail: tholt@ons.go.uk.

[^10]:    ${ }^{1}$ Manas Chattopadhyay, The Gallup Organization; Partha Lahiri, University of Nebraska/Lincoln; Michael Larsen, Harvard University, Department of Statistics, Science Center, One Oxford Street, Cambridge, MA 02138, U.S.A.; John Reimnitz, The Gallup Organization.

[^11]:    ${ }^{1}$ Susana Rubin Bleuer, Business Survey Methods Division, Statistics Canada, Ottawa, Ontario, K1A 0T6, e-mail: rubisus@statcan.ca; Milorad S. KovaCevic, Social Survey Methods Division, Statistics Canada, Ottawa, Ontario, K1A 0T6, e-mail: kovamil@statcan.ca.

[^12]:    1 Pam F. Tate, Office for National Statistics, Room RG/11, 1 Drummond Gate, London SW1V 2QQ, United Kingdom.

[^13]:    ${ }^{1}$ Siegfried Gabler and Sabine Haeder, ZUMA, B 2,1, D-68159 Mannheim; Partha Lahiri, University of Nebraska-Lincoln, NE 68588-0323.

