



SURVEY METHODOLOGY

STATISTICS
CANADA

STATISTIQUE
CANADA

MAR 12 1998

LIBRARY
BIBLIOTHÈQUE

Catalogue No. 12-001-XPB

A JOURNAL
PUBLISHED BY
STATISTICS CANADA

DECEMBER 1997

•

VOLUME 23

•

NUMBER 2



Statistics
Canada

Statistique
Canada

Canada



SURVEY METHODOLOGY

A JOURNAL
PUBLISHED BY
STATISTICS CANADA

DECEMBER 1997 • VOLUME 23 • NUMBER 2

Published by authority of the Minister
responsible for Statistics Canada

© Minister of Industry, 1998

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

February 1998

Catalogue no. 12-001-XPB

Frequency: Semi-annual

ISSN 0714-0045

Ottawa



Statistics
Canada

Statistique
Canada

Canada

SURVEY METHODOLOGY

A Journal Published by Statistics Canada

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

MANAGEMENT BOARD

Chairman G.J. Brackstone

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

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*
W.A. Fuller, *Iowa State University*
R.M. Groves, *University of Maryland*
M.A. Hidirolou, *Statistics Canada*
D. Holt, *Central Statistical Office, U.K.*
G. Kalton, *Westat, Inc.*
R. Lachapelle, *Statistics Canada*
S. Linacre, *Australian Bureau of Statistics*
G. Nathan, *Central Bureau of Statistics, Israel*
D. Pfeffermann, *Hebrew University*

J.N.K. Rao, *Carleton University*
L.-P. Rivest, *Université Laval*
I. Sande, *Bell Communications Research, U.S.A.*
F.J. Scheuren, *George Washington University*
J. Sedransk, *Case Western Reserve University*
R. Sitter, *Simon Fraser University*
C.J. Skinner, *University of Southampton*
R. Valliant, *U.S. Bureau of Labor Statistics*
V.K. Verma, *University of Essex*
P.J. Waite, *U.S. Bureau of the Census*
J. Waksberg, *Westat, Inc.*
K.M. Wolter, *National Opinion Research Center*
A. Zaslavsky, *Harvard University*

Assistant Editors J. Denis, P. Dick, H. Mantel 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 1 800 700-1033, by fax (613) 951-1584 or 1 800 889-9734 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 23, Number 2, December 1997

CONTENTS

In This Issue	79
P.S. KOTT and D.M. STUKEL Can the Jackknife Be Used With a Two-Phase Sample?	81
G. DECAUDIN and J.-C. LABAT A Synthetic, Robust and Efficient Method of Making Small Area Population Estimates in France	91
P. RAVALET An Adaptive Procedure for the Robust Estimation of the Rate of Change of Investment	99
F. COTTON and C. HESSE Sampling and Maintenance of a Stratified Panel of Fixed Size	109
P.J. FARRELL Empirical Bayes Estimation of Small Area Proportions Based on Ordinal Outcome Variables	119
A. GELMAN and T.C. LITTLE Poststratification Into Many Categories Using Hierarchical Logistic Regression	127
K.K. SINGH, A.O. TSUI, C.M. SUCHINDRAN and G. NARAYANA Estimating the Population and Characteristics of Health Facilities and Client Populations Using a Linked Multi-Stage Sample Survey Design	137
J. DUFOUR, R. KAUSHAL and S. MICHAUD Computer-assisted Interviewing in a Decentralised Environment: The Case of Household Surveys at Statistics Canada	147
F. SCHEUREN and W.E. WINKLER Regression Analysis of Data Files That Are Computer Matched - Part II	157
Acknowledgements	167

In This Issue

This issue of *Survey Methodology* contains articles on a variety of topics. Kott and Stukel consider jackknife variance estimation for a specific, but widely used two-phase design. At the first phase, clusters within strata are selected using SRS with replacement, and all units within the selected clusters are sampled. At the second phase, the sampled units are restratified and then second phase units are selected using SRS without replacement. Two point estimators are considered: the "reweighted expansion estimator" and the more commonly known "double expansion estimator". Under this design, it is shown that the jackknife variance estimator behaves remarkably better for the former point estimator than it does for the latter. A Monte Carlo study supports these findings.

Decaudin and Labat describe a "multi-source" population estimation system designed to produce local population estimates during intercensal periods in France. The system is robust and flexible in that it works with a variable number of sources. It is based on a robust combination of estimates from different sources, blending demographic reasoning with statistical methods.

Ravalet applies GM-estimators to INSEE's industrial investment survey with an adaptive procedure to produce a robust estimator. Tukey's biweight function and the Cauchy function are examined. Each function relies on a tuning constant based on the width of the tail of the distribution and the concentration of the residuals. Tuning constants that minimize the estimator's variance are determined for eight distributions representing various scenarios relating to the width of the tail and the concentration of the residuals, which are assumed to be symmetrical.

Cotton and Hesse study the characteristics of various methods of selecting a stratified panel of fixed size, along with their impact on initial selection, rotation, resampling and sample overlap. The authors propose a kind of algorithm based on transformations of permanent random numbers used for sampling purposes; the algorithm extends the pre-resampling rotation into the post-resampling period. The transformations can be performed on random numbers that have been made equidistant and on random numbers derived from a uniform distribution.

In his paper Farrell studies empirical Bayes estimation of small area proportions. Using data from the United States Census he compares empirical Bayes small area estimates of proportions of individuals in different income categories based on multinomial and ordinal logistic models with random effects. Inferences based on the ordinal model were slightly better than those based on the multinomial model. He also compares naive and bootstrap adjusted variance estimates and coverage probabilities of their associated confidence intervals. The bootstrap adjustment improves coverage significantly.

Gelman and Little describe a novel extension of analyzing poststratified survey data, using Bayesian hierarchical logistic regression modelling. The technique allows for many more stratification categories than are typically feasible using standard poststratification and weighting strategies, and thus much more population level information can be included in the model. The proposed method as well as some of the more standard methods are applied to pre-election opinion polling data in the U.S., and the various models are evaluated graphically by comparing them to actual election outcomes.

Singh, Tsui, Suchindran and Narayana describe the survey design and estimation techniques used for PERFORM (Project Evaluation Review for Organizational Resource Management), a large scale survey conducted in the state of Uttar Pradesh in India. The survey was designed to estimate the characteristics of health facilities and their target populations, in order to provide benchmark indicators for a large family planning project. PERFORM uses a stratified multi-stage design, where the ultimate sampling units are households and eligible females residing within. However, estimates of health facilities, which are not explicitly part of the sampling scheme, are also obtained by adjusting for multiplicity of the selected secondary sampling units served by those health facilities.

Dufour, Kaushal and Michaud review the tests and studies that preceeded the implementation of computer-assisted interviewing for most household surveys at Statistics Canada. The interviewing is conducted, in person at the respondent's home or by telephone from the interviewer's home, using laptop computers. They also discuss the challenges that were faced with the implementation of the new technology into ongoing surveys and the new opportunities for monitoring survey collection offered by it.

Scheuren and Winkler propose a method for using noncommon but correlated quantitative variables to improve record linkage. The basic idea is to use the linkages which are almost certainly correct to estimate a regression relationship between the noncommon variables and then to use the predicted values of these variables in a subsequent record linkage step. The procedure can then be iterated until convergence. The regression step uses a procedure which adjusts the regression for possible errors in the linkage, described in an article by the same authors in the June 1993 issue of *Survey Methodology*. The method is illustrated empirically and it is shown that it can lead to good results in situations that were hitherto hopeless.

The Editor

Dear *Survey Methodology* Reader,

I would like to take a moment to thank you for your interest and support of *Survey Methodology*. Since its inception, the journal remains committed to publishing articles relevant to statistical agencies and researchers with emphasis on the development and evaluation of specific methodologies as applied to data collection or to the data themselves.

Survey Methodology is approaching its 25th anniversary. From its beginning as an in-house review of developments in survey methodology in Statistics Canada, it has evolved into a widely read statistical journal with an editorial board of internationally recognized survey statisticians. Though many improvements to content and presentation have occurred during this period, there is always room for improvement. I would appreciate any suggestions, comments and recommendations you may have to assist us in our task of maintaining *Survey Methodology* as a viable platform for statistical development into the next millennium.

Should you wish to have complimentary copies of *Survey Methodology* sent to a colleague, please do not hesitate to contact us.

I thank you again for your interest and continued support of *Survey Methodology*.

Sincerely,

M.P. Singh
singhmp@statcan.ca

Can the Jackknife Be Used With a Two-Phase Sample?

PHILLIP S. KOTT and DIANA M. STUKEL¹

ABSTRACT

The jackknife variance estimator has been shown to have desirable properties when used with smooth estimators based on stratified multi-stage samples. This paper focuses on the use of the jackknife given a particular two-phase sampling design: a stratified with-replacement probability cluster sample is drawn, elements from sampled clusters are then restratified, and simple random subsamples are selected within each second-phase stratum. It turns out that the jackknife can behave reasonably well as an estimator for the variance for one common "expansion" estimator but not for another. Extensions to more complex estimation strategies are then discussed. A Monte Carlo study supports our principal findings.

KEY WORDS: Stratified; Reweighted expansion estimator; Double expansion estimator; Asymptotic.

1. INTRODUCTION

Krewski and Rao (1981) and Rao and Wu (1985) explore the design-based properties of the jackknife variance estimator given a stratified multi-stage sample incorporating with-replacement sampling in the first stage. Their results, although fairly general, cannot be directly applied to many multi-phase sampling designs. See also Wolter (1985; Chapter 4.5).

In this paper, we consider a simple example of two-phase sampling. A stratified with-replacement probability cluster sample is selected in a first phase of sampling. The elements in sampled clusters are then restratified, perhaps using information gathered from the first-phase sample, and a stratified simple random subsample is drawn without replacement.

One can estimate a total without auxiliary information in one of two ways. In the *double expansion estimator* – called "the π^* estimator" in Särndal, Swensson, and Wretman (1992, p. 347) – the value of each subsampled element is simply multiplied by the product of its expansion factor at each phase (*i.e.*, the inverses of its first-phase and second-phase selection probabilities) and then summed.

Although the double expansion estimator is more easily located in text books, the *reweighted expansion estimator* may be more common in practice, especially when element nonresponse is treated as a second phase of sampling, as in the weighting class estimator of Oh and Scheuren (1983, p. 150). An estimator for the population size of each second-phase stratum is computed by summing the first-phase expansion factors of all the elements in the second-phase stratum before subsampling. This value is then multiplied by the estimated second-phase stratum mean based on the subsample to yield an estimated stratum total. The second-phase estimated stratum totals are finally added together to produce the reweighted expansion estimator for the population total.

We are more concerned here with real two-phase sampling, rather than the artifice of treating nonresponse as

an additional sampling phase. The National Agricultural Statistics Service (NASS) presently uses the double expansion estimator in its Quarterly Agricultural Surveys (QAS). A stratified area cluster sample is enumerated in June. Farms identified in the June survey are restratified based on their June responses and then subsampled for enumeration in September, December, and March.

NASS uses a two-phase design and the reweighted expansion estimator for its on-farm chemical use surveys. The first phase of sampling identifies farms with specific crops, and the second phase measures pesticide use on those crops.

This paper shows that although the jackknife may be used to estimate the variance of the reweighted expansion estimator under certain conditions, it is not generally effective as a variance estimator for the double expansion estimator. Section 2 introduces the reweighted expansion estimator and discusses its mean squared error. Section 3 shows that the jackknife variance estimator can be nearly unbiased for the reweighted variance estimator, while Section 4 addresses the jackknife's failings as a variance estimator for the double expansion estimator. Section 5 describes a simulation study that appears to confirm the main assertions of the previous sections. Section 6 discusses extensions of the reweighted expansion estimator, and Section 7 offers some concluding remarks. An appendix provides an outline of our assumed asymptotic framework and some proofs.

2. THE REWEIGHTED EXPANSION ESTIMATOR

2.1 The Estimator

Let $h (= 1, \dots, H)$ denote the first-phase strata of a stratified with-replacement probability cluster sample, n_h the number of sampled clusters in stratum h , and F_h the set of those clusters. Let $g (= 1, \dots, G)$ be the second-phase

¹ Phillip S. Kott, National Agricultural Statistics Service, 3251 Old Lee Highway, Room 305, Fairfax, VA 22030; Diana M. Stukel, Household Survey Methods Division, Statistics Canada, Ottawa, Canada K1A 0T6.

strata from which a stratified simple random subsample is drawn without replacement. An element in a cluster sampled p times in the first phase is treated as p distinct elements for the subsample. Let M_g be the number of elements in g before subsampling and m_g the number of subsampled elements in g . In practice, the G second-phase strata are often not defined until after the first-phase sample has been drawn.

Let S_g be the set of elements in g before subsampling, s_g the set of subsampled elements in g , s the entire set of subsampled elements, and $m = \sum_g m_g$ the subsample size. Finally, let y_i be the value of interest for element i , and w_i the first-phase expansion factor for i (i.e., the inverse of the selection probability for the cluster containing i).

The estimator for the population total, T , one would use if all the elements in the first-phase sample were enumerated can be written as

$$t_1 = \sum_{g=1}^G \sum_{i \in S_g} w_i y_i. \quad (1)$$

Let the *reweighted expansion estimator* for T be:

$$\begin{aligned} t_2 &= \sum_{g=1}^G \left\{ \frac{\sum_{i \in S_g} (M_g/m_g) w_i y_i}{\sum_{i \in S_g} w_i} \right\} \\ &= \sum_{g=1}^G \left\{ \frac{\sum_{i \in S_g} w_i y_i}{\sum_{i \in S_g} w_i} \right\}. \end{aligned} \quad (2)$$

An alternative expression for t_2 is

$$t_2 = \sum_{g=1}^G \sum_{i \in s_g} a_i y_i = \sum_{i \in s} a_i y_i, \quad (3)$$

where

$$a_i = \left[\sum_{k \in S_g} w_k / \sum_{k \in s_g} w_k \right] w_i \text{ for } i \in s_g$$

is the *adjusted weight* for element i . Equation (3) is what gives the reweighted expansion estimator its name.

2.2 Its Mean Squared Error (Some Theory)

Now t_2 is not, in general, an unbiased estimator of T . Nevertheless, under certain mild conditions specified in the appendix, it is a design consistent estimator for T ; that is, $\text{plim}_{m \rightarrow \infty} (t_2 - T)/T = 0$ (Isaki and Fuller 1982). For the exposition in the text, it suffices to say that the m_g are assumed to be large.

Observe that

$$\begin{aligned} E[(t_2 - T)^2] &= E[(\{t_1 - T\} + \{t_2 - t_1\})^2] \\ &\approx \text{Var}_1(t_1) + E_1\{E_2[(t_2 - t_1)^2]\}, \end{aligned}$$

where the subscripts on Var and E denote the phase of sampling. Since the m_g are assumed to be large, $E_2[t_1(t_2 - t_1)] = t_1 E_2(t_2 - t_1) \approx 0$. Also, $E(t_2 - T) = E_1[E_2(t_2 - T)] \approx 0$, and the mean squared error of t_2 is effectively its (asymptotic) variance.

Since first phase of sampling was conducted with replacement, $\text{Var}_1(t_1)$ can, in principle, be estimated by

$$\begin{aligned} v_{L1} &= \sum_{h=1}^H (n_h/[n_h - 1]) \\ &\quad * \left(\sum_{j \in F_h} \left[\sum_{i \in U_{hj}} w_i y_i \right]^2 - \left[\sum_{j \in F_h} \sum_{i \in U_{hj}} w_i y_i \right]^2 / n_h \right), \end{aligned} \quad (4)$$

where U_{hj} is the set the elements in sampled cluster j of first-phase stratum h . The subscript L denotes "linearization" for historical reasons although there is nothing to linearize in this context. Note that when there is a second phase of sampling, it will generally not be possible to compute v_{L1} in practice.

Now

$$\begin{aligned} t_2 - t_1 &= \sum_{g=1}^G \sum_{i \in S_g} w_i \left\{ \frac{\sum_{i \in s_g} w_i y_i}{\sum_{i \in S_g} w_i} - \frac{\sum_{i \in S_g} w_i y_i}{\sum_{i \in S_g} w_i} \right\} \\ &= \sum_{g=1}^G \sum_{i \in S_g} w_i \frac{\sum_{i \in s_g} w_i r_i}{\sum_{i \in S_g} w_i}, \end{aligned}$$

where

$$r_i = y_i - \sum_{k \in S_g} w_k y_k / \sum_{k \in S_g} w_k \text{ for } i \in S_g.$$

It is crucial for the arguments below to realize that r_i has been defined so that $\sum_{i \in S_g} w_i r_i = 0$ for all g .

Continuing,

$$t_2 - t_1 \approx \sum_{g=1}^G \sum_{i \in s_g} (M_g/m_g) w_i r_i, \quad (5)$$

since $\sum_{i \in S_g} w_i \approx \sum_{i \in s_g} (M_g/m_g) w_i$ (see equation (A1) of the appendix). This implies

$$\begin{aligned} E_2[(t_2 - t_1)^2] &\approx \text{Var}_2 \left\{ \sum_{g=1}^G \sum_{i \in s_g} (M_g/m_g) w_i r_i \right\} \\ &= \sum_{g=1}^G (M_g^2 / [(M_g - 1) m_g]) (1 - m_g/M_g) \\ &\quad * \left\{ \sum_{i \in S_g} (w_i r_i)^2 - \left(\sum_{i \in S_g} w_i r_i \right)^2 / M_g \right\} \\ &\approx \sum_{g=1}^G ([M_g/m_g] - 1) \left\{ \sum_{i \in S_g} (w_i r_i)^2 \right\}. \end{aligned} \quad (6)$$

Observe that equation (6) does *not* ignore the finite population corrections from the second phase of sampling.

3. THE JACKKNIFE VARIANCE ESTIMATOR

3.1 The Variance Estimator

We are now ready to discuss the jackknife. For $j \in F_h$, define the jackknife replicate $t_{(hj)2}$ as

$$t_{(hj)2} = \sum_{g=1}^G \left\{ \frac{\sum_{i \in S_g} w_{hji} y_i}{\sum_{i \in S_g} w_{hji}} \right\}, \quad (7)$$

where

$$w_{hji} = \begin{cases} w_i n_h / (n_h - 1) & \text{when } i \in U_{hj'} \text{ and } j' \neq j \\ 0 & \text{when } i \in U_{hj} \\ w_i & \text{when } i \in U_{h'j'} \text{ and } h' \neq h. \end{cases}$$

Similarly, we define

$$t_{(hj)1} = \sum_{g=1}^G \sum_{i \in S_g} w_{hji} y_i.$$

Following Rust (1985), the *jackknife variance estimator*, v_{jf} ($f = 1$ or 2), is defined here simply as

$$v_{jf} = \sum_{h=1}^H (n_h - 1) / n_h \sum_{j \in F_h} (t_{(hj)f} - t_f)^2. \quad (8)$$

This form is labeled $v_j^{(2)}$ in Krewski and Rao (1981, equation (2.4)). It is easy to show that $v_{j1} = v_{L1}$.

3.2 Why it Works (More Theory)

We will soon see that v_{j2} provides a nearly unbiased estimator for the variance of the reweighted expansion estimator in equation (2). Rao and Shao (1992) indirectly make the same claim (our equation (2) is the expectation of their estimator in Section 3.3, pp. 818-819). Their work, however, treats nonresponse as an additional phase of sample selection in which Poisson sampling (Särndal *et al.* 1992, p. 85) is used in place of stratified simple random sampling. Each first-phase sample element in the Rao and Shao (1992) setup is effectively a second-phase stratum. Consequently, the near unbiasedness of v_{j2} reduces to a special case of a result in Krewski and Rao (Rao and Shao 1992, p. 821).

What we have called the second-phase strata are reweighting classes in the Rao and Shao (1992) setup. Elements in the same class are assumed to have the same unknown probability of selection/response. *Conditional on*

the realized subsample sizes within reweighting classes, Poisson sampling is equivalent to stratified simple random sampling. Rao and Shao's (1992) treatment, however, is *unconditional*.

Returning to the problem at hand, observe that

$$\begin{aligned} t_{(hj)2} - t_{(hj)1} &= \sum_{g=1}^G \sum_{i \in S_g} w_{hji} \left\{ \frac{\sum_{i \in S_g} w_{hji} y_i}{\sum_{i \in S_g} w_{hji}} - \frac{\sum_{i \in S_g} w_{hji} y_i}{\sum_{i \in S_g} w_{hji}} \right\} \\ &= \sum_{g=1}^G \left\{ \sum_{i \in S_g} w_{hji} \frac{\sum_{i \in S_g} w_{hji} r_{hji}}{\sum_{i \in S_g} w_{hji}} \right\}, \end{aligned}$$

where

$$r_{hji} = y_i - \sum_{k \in S_g} w_{hjk} y_k / \sum_{k \in S_g} w_{hjk} \quad \text{for } i \in S_g.$$

Under mild conditions (see equations (A2) and (A3) in the appendix), we have the following analogue to equation (5):

$$\begin{aligned} t_{(hj)2} &\approx t_{(hj)1} + \sum_{g=1}^G (M_g / m_g) \sum_{i \in S_g} w_{hji} r_{hji} \\ &= \sum_{g=1}^G \sum_{i \in S_g} w_{hji} (y_i + [M_g / m_g] c_i r_{hji}), \end{aligned} \quad (9)$$

where c_i is an indicator variable equal to 1 when i is in the subsample and zero otherwise.

Continuing,

$$\begin{aligned} t_{(hj)2} &\approx \sum_{g=1}^G \sum_{i \in S_g} w_{hji} (y_i + \{[M_g / m_g] c_i - 1\} r_{hji}) \\ &= \sum_{g=1}^G \sum_{i \in S_g} w_{hji} z_{hji}, \end{aligned} \quad (10)$$

where $z_{hji} = y_i + \{[M_g / m_g] c_i - 1\} r_{hji}$. Again, since every m_g is large, it is not unreasonable to assume $r_{hji} \approx r_i$ (see equation (A4) in the appendix). Thus,

$$t_{(hj)2} \approx \sum_{g=1}^G \sum_{i \in S_g} w_{hji} z_i,$$

where $z_i = y_i + \{[M_g / m_g] c_i - 1\} r_i$. Using similar arguments, $t_2 \approx \sum_{g=1}^G \sum_{i \in S_g} w_i z_i$. Since t_2 is linear in the z_i ,

$$\begin{aligned} v_{j2} &\approx v_{L1} \left(\sum_{h=1}^H \sum_{j \in F_h} w_j z_j \right) = \sum_{h=1}^H (n_h / [n_h - 1]) \\ &\quad * \left(\sum_{j \in F_h} \left[\sum_{i \in U_{hj}} w_i z_i \right]^2 - \left[\sum_{j \in F_h} \sum_{i \in U_{hj}} w_i z_i \right]^2 / n_h \right). \end{aligned} \quad (11)$$

Let $e_i = M_g/m_g$ be the second-phase expansion factor for $i \in S_g$. Observe that c_i is a random variable with $E(c_i) = m_g/M_g$ and $E(c_i c_k) = (m_g/M_g)(m_g - 1)/(M_g - 1)$ for $i, k \in S_g, i \neq k$.

Now

$$E_2 \left[\left(\sum_{i \in U_{hj}} w_i z_i \right)^2 \right] \approx \left(\sum_{i \in U_{hj}} w_i y_i \right)^2 + \sum_{i \in U_{hj}} (e_i - 1) (w_i r_i)^2 - \sum_{g=1}^G \sum_{i, k \in S_g \cap U_{hj} \atop i \neq k} [(1 - m_g/M_g)/m_g] w_i r_i w_k r_k. \quad (12)$$

Similarly, letting F_h^* be the set of elements from selected clusters in the first-phase stratum h before subsampling, we have

$$E_2 \left[\left(\sum_{j \in F_h} \sum_{i \in U_{hj}} w_i z_i \right)^2 \right] = E_2 \left[\left(\sum_{i \in F_h^*} w_i z_i \right)^2 \right] \approx \left(\sum_{i \in F_h^*} w_i y_i \right)^2 + \sum_{i \in F_h^*} (e_i - 1) (w_i r_i)^2 - \sum_{g=1}^G \sum_{i, k \in S_g \cap F_h^* \atop i \neq k} [(1 - m_g/M_g)/m_g] w_i r_i w_k r_k. \quad (13)$$

In the appendix, it is argued that under mild conditions that the last term in both equations (12) and (13) is negligible. As a result,

$$\begin{aligned} E_2(v_{J2}) &\approx v_{J1} + \sum_{h=1}^H \sum_{i \in F_h^*} (e_i - 1) (w_i r_i)^2 \\ &= v_{J1} + \sum_{g=1}^G \sum_{i \in S_g} [(M_g/m_g) - 1] (w_i r_i)^2 \\ &\approx v_{L1} + E_2[(t_2 - t_1)^2], \end{aligned} \quad (14)$$

which in turn implies that v_{J2} is a nearly unbiased estimator for $E[(t_2 - T)^2]$.

4. THE DOUBLE EXPANSION ESTIMATOR

An alternative to t_2 , the *double expansion* estimator, has the form:

$$t_3 = \sum_{g=1}^G \sum_{i \in S_g} (M_g/m_g) w_i y_i. \quad (15)$$

The definition of a jackknife replicate for t_3 is unclear. One simple possibility is

$$t_{(hj)3} = \sum_{g=1}^G \sum_{i \in S_g} w_{hji} (M_g/m_g) y_i. \quad (16)$$

Another, perhaps more in the spirit of "replication", is

$$t_{(hj)3}^* = \sum_{g=1}^G \sum_{i \in S_g} w_{hji} (M_{ghj}/m_{ghj}) y_i, \quad (17)$$

where M_{ghj} is the number of elements in the first-phase sample (*i.e.*, in a cluster in the first-phase sample) that are in S_g but *not* U_{hj} . Similarly, m_{ghj} is the number of elements in the second-phase sample that are in s_g but *not* U_{hj} . Through counter-examples given in the appendix, we show that neither version of the replicate produces a jackknife variance estimator (v_{J3} from equation (8)) that is asymptotically unbiased in general.

5. A MONTE CARLO SIMULATION STUDY

5.1 Design of the Study

The results given so far in the text are asymptotic. In order to assess the accuracy of the jackknife as a variance estimator for the reweighted expansion estimator in a finite world, we undertook a Monte Carlo simulation study. At the same time, we assessed the accuracy of the two jackknife estimators suggested for the double expansion estimator in Section 4.

We used December 1990 Canadian Labour Force Survey (LFS) sample data for the province of Newfoundland to simulate a finite population, from which repeated samples were drawn. The LFS is the largest ongoing household sample survey conducted by Statistics Canada. Monthly data relating to the labour market is collected using a complex multi-stage sampling design with several levels of stratification. The details of the design of the survey prior to the 1991 redesign can be found in Singh, Drew, Gambino and Mayda (1990) and Stukel and Boyer (1992). In general, provinces are stratified into "economic regions", which are large areas of similar economic structure; Newfoundland has four such economic regions. The economic regions are further substratified into lower level substrata. The lowest level of stratification in Newfoundland yielded 45 strata, each of which contained less than 6 clusters or *primary sampling units* (PSU's), which was an insufficient number from which to sample for the purposes of the simulation. Thus, the 45 strata were collapsed down to 18, each containing between 6 and 18 PSU's. In collapsing the strata, economic regions were kept intact, as were the Census Metropolitan Areas of St. John's and Cornerbrook.

For the Monte Carlo study, $R = 4,000$ samples were drawn from the Newfoundland "population" (which was 9,152 individuals), according to the following two-phase design: within each first-phase stratum, two PSU's were selected at the first phase using simple random sampling (SRS) *with* replacement. This yielded a total of 36 PSU's. All households within selected first-phase PSU's (as well as individuals within those households) were selected, resulting in a single-stage take-all cluster sample. At the second phase, all selected first-phase elements (individuals, treating each person in a PSU selected twice as two separate individuals) were restratified according to five age categories (≤ 14 , 15-24, 25-44, 45-64, ≥ 65), and second-phase sample elements (*i.e.*, individuals) were drawn using SRS *without* replacement sampling within each of the five second-phase strata.

We varied the second-phase stratum sample size to take on values $m_g = 5, 10, 20$, and 50 yielding overall second-phase sample sizes of $m = 25, 50, 100$, and 250. When the number of first-phase-sampled individuals in a second-phase stratum was less than our target m_g value, we planned to set $m_g = M_g$, but that event never occurred.

A popular rule of thumb for a "separate ratio estimator" such as the reweighted expansion estimator in equation (2) is that there should be at least 20 individuals within each second-phase stratum (see, for example, Särndal, Swensson and Wretman 1992, p. 270). By allowing m_g to be as small as 5 and 10, we are checking whether this rule is really necessary.

We considered two parameters of interest: T_y , the total number of employed, and T_y/T_z the employment rate. Here $T_y = \sum_{i \in U} y_i$, where $y_i = 1$ when individual i is employed; 0 otherwise. Similarly, $T_z = \sum_{i \in U} z_i$, where $z_i = 1$ when individual i is in the labour force (i.e., either employed or unemployed); 0 otherwise. For each of the $R = 4,000$ samples, we calculated the reweighted expansion estimator (REE), t_2 , given by equation (2), the double expansion estimator (DEE), t_3 , given by equation (15), and the full first-phase expansion estimator (FFPE), t_1 given by equation (1). Although these estimators are defined for totals (applicable for total number of employed), it is a simple matter to extend them to ratios of totals (applicable for employment rate).

For each of the $R = 4,000$ second-phase samples, we calculated the jackknife variance corresponding to the reweighted expansion estimator and the double expansion estimator, given by equation (8) with $f=2$ and $f=3$ respectively. In the case of the double expansion estimator, we attempted both the replicates defined in equations (16) and (17), which we will refer to as variant 1 and 2, respectively.

For each of the $R = 4,000$ first-phase samples, we also calculated the jackknife variance corresponding to the full first-phase estimator for comparison purposes. This is given by equation (8) with $f=1$.

For all of the above estimators and their corresponding jackknife variances, a number of frequentist properties were investigated. These are given below. For simplicity, they are expressed only in terms of estimates of the total number of employed.

The percent relative bias of the estimated number of employed with respect to the population value is estimated by

$$\text{PRB}(t^*) = \{[E_M(t^*)/T_y] - 1\} \times 100, \quad (18)$$

where

$$E_M(t^*) = (1/4,000) \sum_{r=1}^{4,000} t_r^*$$

is the Monte Carlo expectation of the point estimator t^* taken over the 4,000 samples. Here t^* can be either t_1, t_2 , or t_3 , and t_r^* is the value of t^* for sample r .

The percent relative bias of the jackknife variance estimator with respect to the true mean squared error is

estimated by

$$\text{PRB}[v_{Jf}(t^*)] = ((E_M[v_{Jf}(t^*)] - \text{MSE}_{\text{true}}) / \text{MSE}_{\text{true}}) \times 100, \quad (19)$$

where

$$E_M[v_{Jf}(t^*)] = (1/4,000) \sum_{r=1}^{4,000} v_{Jf}(t_r^*),$$

$$\text{MSE}_{\text{true}} = (1/4,000) \sum_{r=1}^{4,000} (t_r^* - T_y)^2,$$

and $v_{Jf}(t^*)$ is the value of $v_{Jf}(t^*)$ for sample r .

The (percent) coefficient of variation of the jackknife variance with respect to the true MSE is estimated by:

$$\text{CV}[v_{Jf}(t^*)] = ((1/4,000) \sum [v_{Jf}(t_r^*) - \text{MSE}_{\text{true}}]^2)^{1/2} / \text{MSE}_{\text{true}} \times 100; \quad (20)$$

that is, the estimated root mean squared error of the variance estimator divided by the estimated true MSE, expressed as a percentage.

5.2 Results of the Study

Table 1A gives the estimated percent relative biases of the three point estimates for the total number of employed using equation (18), and Table 1B gives the same for the employment rate. All biases are less than 1% in absolute value.

Table 1A
Percent Relative Bias of the Point Estimates
for Total Number of Employed

Estimator	$m_g = M_g$	$m_g = 50$	$m_g = 20$	$m_g = 10$	$m_g = 5$
REE	—	0.14	-0.3	-0.29	-0.56
DEE	—	0.16	-0.01	0.03	0.115
FFPE	0.04	—	—	—	—

Table 1B
Percent Relative Bias of the Point Estimates
for Employment Rate

Estimator	$m_g = M_g$	$m_g = 50$	$m_g = 20$	$m_g = 10$	$m_g = 5$
REE	—	-0.09	-0.31	-0.19	-0.26
DEE	—	-0.08	-0.27	-0.12	-0.13
FFPE	-0.09	—	—	—	—

REE - Reweighted Expansion Estimator (t_2)
DEE - Double Expansion Estimator (t_3)
FFPE - Full First Phase Estimator (t_1)

Not displayed are the Monte Carlo estimates of the mean squared errors (*i.e.*, the values of MSE_{true}) and the corresponding coefficients of variation from using either the reweighted or double expansion estimator. This is because the focus in this article is on mean squared error estimation. The mean squared errors (and coefficients of variation) from using the two estimators are comparable for each sample size (a relative difference in the coefficient of variation is roughly half of the corresponding relative difference in mean squared error). The reweighted expansion estimator is slightly more efficient when estimating the total number of employed individuals (*e.g.*, when $m_g = 5$, the double expansion estimator has 17% more mean squared error). There is less than a 1% difference in the mean squared errors from using the two approaches when estimating the employment rate. Not surprisingly, the mean squared errors for all estimators increase as the second-phase sample size decreases.

Table 2A gives the estimated percent relative biases of the jackknife variances for the total number of employed using equation (19), and Table 2B gives the same for the employment rate. Focusing first on Table 2A, the full first-phase estimator's variance is almost perfectly unbiased, at 0.94%. The jackknife for the reweighted expansion estimator works well, having small negative biases in the variances always less than -6%. The biases tend to become more negative (although not uniformly) as the second-phase sample sizes diminish.

Table 2A
Percent Relative Bias of Jackknife Variances
for Total Number of Employed

Estimator	$m_g = M_g$	$m_g = 50$	$m_g = 20$	$m_g = 10$	$m_g = 5$
REE	-	-0.99	-2.51	-5.81	-5.13
DEE (Variant 1)	-	46.35	68.24	78.18	86.22
DEE (Variant 2)	-	101.59	278.44	654.99	1997.51
FFPE	0.94	-	-	-	-

Table 2B
Percent Relative Bias of Jackknife Variances
for Employment Rate

Estimator	$m_g = M_g$	$m_g = 50$	$m_g = 20$	$m_g = 10$	$m_g = 5$
REE	-	-3.53	-3.45	-7.09	-6.55
DEE (Variant 1)	-	-2.46	-1.53	-5.21	-7.41
DEE (Variant 2)	-	-0.36	4.91	9.09	30.46
FFPE	2.08	-	-	-	-

REE - Reweighted Expansion Estimator (t_2)

DEE - Double Expansion Estimator (t_3)

FFPE - Full First Phase Estimator (t_1)

Variant 1 uses the jackknife replicates in equation (16)

Variant 2 uses the jackknife replicates in equation (17)

In contrast, both jackknife variants for the double expansion estimator fail miserably, with very large positive biases in the variances ranging from 46.35% to 1997.51%! The second variant is worse than the first, but both are well beyond the realm of acceptable behavior.

Table 2B repeats the analysis for the ratio estimate of employment rate. The results here are surprising since all variance estimators behave reasonably well, with the exception of variant 2 of the double expansion estimator when $m_g = 5$. Other than this case where the bias in the variance is 30.46%, all other biases are less than 10% in absolute value.

Overall, Table 2A and 2B provide strong support for using the jackknife variance estimator with a reweighted expansion estimator even when second-phase sample sizes are surprisingly small. By contrast, the jackknife can fail miserably for the double expansion estimator when estimating totals. Sometimes, however, variant 1 can also work reasonably well depending on the estimator and the data.

Although most studies focus on the *bias* of the variance estimators, it is also of secondary interest to look at the *coefficient of variation* of the variance estimators to see how stable the variance estimates themselves are. In Tables 3A and 3B, we investigate the estimated (percent) coefficients of variation corresponding to the total number of employed and the employment rate, respectively. In equation (20), the expression under the square root in the numerator gives the MSE of the variance, whose component parts are the square of the bias of the variance and the variance of the variance. For those entries in Tables 2A and 2B where the bias of the variance has been determined to be exceedingly large (say larger than 20%), the corresponding entries in Tables 3A and 3B are not reported (indicated by a *), since it is clear that those entries will be excessively large. In Table 3A, the estimated coefficients of variation corresponding to the reweighted expansion estimator range between 46.86% and 53.42%. Coefficients of variation of the magnitude exhibited here are typical for variance estimators, and have been encountered in other simulation studies relating to variances. See, for example, Kovačević and Yung (1997). To that end, note that even the estimated coefficients of variation corresponding to the full first-phase estimators are in the same range, and in fact, somewhat higher than those of the second-phase estimators in all cases.

Table 3B, which gives the coefficients of variation for the variances of the estimated employment rates, are entry by entry higher than their counterparts in Table 3A. In addition, all estimators exhibit the pattern that their corresponding coefficients of variation increase, quite substantially in fact, as the second-phase sample sizes diminish. This effect is more pronounced for the ratio estimators than it is for the estimators of the total. The very high coefficients of variation in the column $m_g = 5$ for both tables is not surprising, since the overall second-phase sample size (25) is actually smaller than the number of PSU's drawn in the first phase of sampling (36). In fact, a

Table 3A
Coefficient of Variation of Jackknife Variances
for Total Number of Employed

Estimator	$m_g = M_g$	$m_g = 50$	$m_g = 20$	$m_g = 10$	$m_g = 5$
REE	—	51.33	49.3	46.86	53.42
DEE (Variant 1)	—	*	*	*	*
DEE (Variant 2)	—	*	*	*	*
FFPE	56.71	—	—	—	—

Table 3B
Coefficient of Variation of Jackknife Variances
for Employment Rate

Estimator	$m_g = M_g$	$m_g = 50$	$m_g = 20$	$m_g = 10$	$m_g = 5$
REE	—	59.28	65.66	74.26	103.06
DEE (Variant 1)	—	59.24	66.16	72.89	99.1
DEE (Variant 2)	—	60.94	73.2	92.71	*
FFPE	78.42	—	—	—	—

REE - Reweighted Expansion Estimator (t_2)

DEE - Double Expansion Estimator (t_3)

FFPE - Full First Phase Estimator (t_1)

Variant 1 uses the jackknife replicates in equation (16)

Variant 2 uses the jackknife replicates in equation (17)

more relevant realized sample count for the ratio estimator is the number of sampled individuals in the labour force (i.e., in the denominator). This value varies from sample to sample and is often considerably less than 25.

6. EXTENDING THE REWEIGHTED EXPANSION ESTIMATOR

6.1 The Reweighted Expansion Estimator

It is not that difficult to develop a linearization variance estimator for the reweighted expansion estimator in equation (2). Suppose, however, one had a sample design with more than two phases or was interested in estimating the ratio of two totals. Linearization, although still possible, becomes increasingly cumbersome. The jackknife, on the other hand, does not.

It is a simple matter to generalize the results in Section 3 to p -phase sampling by induction. The h still refer the first-phase strata, but the g now denote the p -th-phase strata; S_g is the set of elements in the $(p-1)$ -th-phase sample from stratum g while s_g is the p -th-phase subsample from g . The w_i in equation (2) are replaced with the a_i from (3)

for the $(p-1)$ -th-phase estimator. Similarly, the $t_{(hj)2}$ in the jackknife are computed using a_{hji} from the $(p-1)$ -th phase in place of the w_{hji} .

It is also a simple matter (left to the reader) to replace the stratified cluster sample in the first phase of selection with a stratified multi-stage sample. The results in Section 3 follow as long as the first stage of the multi-stage sample is drawn with replacement.

Finally, it is not difficult to extend the results of Section 3 to more complicated estimators. Let U_2 be a vector of estimators each in the form of t_2 from equation (2). The mean squared error of any estimator $\Theta = g(U_2)$, where g is a smooth function, can be estimated with a jackknife in a nearly unbiased manner whenever the members of U_2 can be. This follows the proofs in the literature. Rao and Wu (1985), for example, address the asymptotic framework where the n_h are all bounded, while Wolter (1985; Chapter 4.5) treats the case where the n_h grow arbitrarily large.

6.2 Regression in the Second Phase

The estimator t_2 can be generalized into the regression estimator:

$$t_{2reg} = \sum_{i \in S} w_i x_i \left(\sum_{i \in S} w_i e_i d_i x_i' x_i \right)^{-1} \left(\sum_{i \in S} w_i e_i d_i x_i' y_i \right), \quad (21)$$

where S denotes the original sample, x_i is a row vector, d_i is a scalar, and there exists a row vector γ such that $d_i \gamma x_i' = 1$ for all i . In practice, d_i is usually 1 for all i . A popular exception occurs when $x_i = x_i$ and $d_i = 1/x_i$. In equation (2), $d_i = 1$ for all i , and x_i is a G -vector with a value of 1 in the g -th position and 0's elsewhere for $i \in S_g$.

Let

$$r_i = y_i - x_i \left(\sum_{i \in S} w_i d_i x_i' x_i \right)^{-1} \left(\sum_{i \in S} w_i d_i x_i' y_i \right).$$

The replicate $t_{2reg(hj)}$ has the same form as t_{2reg} except that w_{hji} replaces w_i everywhere. Similarly, r_{hji} has the same form as r_i except that w_{hji} replaces w_i . Note that the e_i are unchanged from t_{2reg} to $t_{2reg(hj)}$.

Since the sampling design hasn't changed, most of equation (6) stays as is except that now $(\sum_{i \in S_g} w_i r_i)^2$ is nonnegative rather than strictly zero. The interested reader can verify that equations (10) through (13) remain in their present form. It turns out that the jackknife has, if anything, an (approximate) upward bias in equation (14). That is to say, the jackknife is a conservative estimator of variance. Again, see the appendix (equations (A6) through (A9)) for a formal statement of the asymptotic assumptions.

The bias in the jackknife disappears when $\sum_{i \in S_g} w_i r_i = 0$ for all g . Formally, this will happen when there exists G row vectors $\gamma_1, \dots, \gamma_G$ such that $d_i \gamma_g x_i' = 1$ when $i \in S_g$ and 0 otherwise (since $\sum_{i \in S_g} w_i r_i = \sum_{i \in S_g} d_i \gamma_g x_i' w_i r_i = \gamma_g \sum_{i \in S_g} w_i d_i x_i' r_i = \gamma_g \{ \sum_{i \in S_g} w_i d_i x_i' (y_i - x_i [\sum_{i \in S_g} w_i d_i x_i' x_i]^{-1} \sum_{i \in S_g} w_i d_i x_i' y_i) \} = 0$). When all $d_i = 1$, the existence of γ_g

means that either one member of x_i is an indicator variable equal to 1 when $i \in S_g$ and 0 otherwise, or one member of a linear transform of x_i is such an indicator variable.

7. CONCLUDING REMARKS

The main purpose of this paper was to show that a simple jackknife variance estimator can be nearly unbiased for an estimation strategy involving two-phase sampling as long as that strategy employs a reweighted expansion estimator and not a double expansion estimator. Since the theoretical results for the reweighted expansion estimator rely on asymptotic arguments, their practical application will depend on the context. Nevertheless, a Monte Carlo simulation study performed here suggests that the jackknife can be an effective estimator for the variance of a reweighted expansion estimator even with surprisingly small second-phase stratum sample sizes, that is, sizes of 5 and 10.

APPENDIX

The Design Consistency of the Reweighted Expansion Estimator

To establish the design consistency of t_2 in equation (2) it is sufficient to assume that the sample design and population values of the y_i are such that

$$\left\{ \sum_{g=1}^G (M_g/m_g) \sum_{i \in S_g} w_i y_i / T \right\} - 1 = O_p(1/\sqrt{m}),$$

and, given any first-phase sample,

$$\left(\sum_{k \in S_g} w_k / \sum_{k \in S_g} w_k \right) (m_g/M_g) - 1 = O_p(1/\sqrt{m}) \quad (A1)$$

for all g . These assumptions justify equation (5) in the text.

We assume in our analysis that G is bounded and that each m_g has the same asymptotic order as m . This is only possible when the S_g are determined *after* the first-phase sample has been drawn. Otherwise, the M_g would be random variables, and a minimum size for each m_g could not be guaranteed for all possible first-phase samples. In principle, we are assuming the existence of a mechanism for determining the S_g and the second-phase sampling fractions given any first-phase sample. By contrast, the exact values of G and the m_g can but need not be fixed before the first-phase sample is drawn.

A Comment on the Asymptotic Framework

Recall that the text showed that the jackknife contains a component that estimates the second-phase variance (i.e., $E_2[(t_2 - t_1)^2]$) in an asymptotically unbiased manner given any first-phase sample (see equation (14)). As a result, that component also estimates the average (i.e., unconditional) second-phase variance across all possible first-phase samples (i.e., $E_1\{E_2[(t_2 - t_1)^2]\}$) in an asymptotically unbiased manner.

In our empirical work, we strayed from the sampling framework described above so that the results could be easily summarized. In particular, we defined the S_g beforehand, and let the M_g be random. When the first-phase sample was such that M_g was less than the desired m_g (say 50) in some second-phase stratum, we planned to choose all the individuals in S_g for the second-phase sample. As a result, there would be no contribution to the mean squared error (or bias) of t_2 from second-phase stratum g when that particular first-phase sample was selected, and so no asymptotic assumptions about m_g would be necessary. As it happened, in no simulation was M_g actually less than 50. Nevertheless, a decision rule about the second-phase sampling fractions was in place for every possible first-phase sample.

Jackknife Replicates

There are (at least) two distinct asymptotic frameworks for the first-phase sample. In the first, there is an arbitrarily large number of first-phase strata each of which is bounded in size; that is, each $1/n_h = O(1)$ while $1/H = O(1/m)$. In the second, all the first-phase strata are arbitrarily large; that is, $1/n_h = O(1/m)$. Under either framework, we assume that the number of elements in each cluster is $O(1)$; that is to say, bounded.

Since every m_g is of the same asymptotic order as m , it is not unreasonable to assume under either regime that, given any first-phase sample,

$$\sum_{i \in S_g} w_{hji} / \sum_{i \in S_g} w_i - 1 = O_p(1/m), \quad (A2)$$

and

$$\sum_{i \in S_g} w_{hji} / \sum_{i \in S_g} w_i - 1 = O_p(1/m), \quad (A3)$$

which can be used to establish equation (9). Similarly, we assume that given any first-phase sample

$$\sum_{i \in S_g} w_{hji} y_i / \sum_{i \in S_g} w_i y_i - 1 = O_p(1/m), \quad (A4)$$

which assures us that $r_{hji} - r_i = O_p(1/m)$.

Equations (12), (13), and (14)

Since the number of elements in each cluster is bounded, say by B . The third term on the right hand side of equation (12) has at most GB^2 terms, a bounded number.

Each of these terms is of order $1/m_g$ (formally, the probability that any one term is of asymptotic order greater than $1/m_g$ is zero). Consequently, the second line of equation (12) is asymptotically ignorable.

Equation (14) holds when each $1/n_h = O(1)$, because if each n_h is less than C (say), then the third term on the right hand side of equation (13) will be the sum of at most $G(BC)^2$ terms, a bounded number. Each of these terms is again of order $1/m_g$. Consequently, the second line of equation (13) is asymptotically ignorable.

Alternatively, suppose each $1/n_h$ were $O(1/m)$. We will assume that the sample design and population is such that, given any first-phase sample,

$$A_h = \sum_{i \in F_h^*} w_i (e_i c_i - 1) r_i / \sum_{i \in F_h^*} w_i y_i = O_p(1/\sqrt{m}) \quad (A5)$$

for all h . To see why this is a reasonable assumption, observe that conditioned on the first-phase sample, the denominator of A_h is a domain total – the sum of the $w_i y_i$ among the elements in F_h^* . Consequently, it is $O(m)$ (without loss of generality we can assume that all the w_i are $O(1)$). The numerator of A_h is the difference between an expansion estimator (the sum of the $w_i e_i c_i r_i$ in F_h^*) based on a stratified simple random sample and its target (the sum of the $w_i r_i$ in F_h^*). Equation (A.5) makes the modest assumption that the sampling design and population is such that this difference is $O_p(\sqrt{m})$ for every possible first-phase sample.

Under assumption (A5), $\sum_{i \in F_h^*} w_i z_i = \sum_{i \in F_h^*} w_i y_i (1 + A_h)$ is approximately equal to $\sum_{i \in F_h^*} w_i y_i$, which implies $E_2[(\sum_{i \in F_h^*} w_i z_i)^2/n_h] \approx (\sum_{i \in F_h^*} w_i y_i)^2/n_h$. Equation (14) follows from this near equality and from equations (11) and (12) (since n_h is large, $n_h/(n_h - 1) \approx 1$).

Counter-examples to the Jackknifes for the Double Expansion Estimator

As a counter-example to the replicate form in equation (16), consider the situation where each cluster contains a single element, $H = G = 1$, and all the y_i values are equal to 1. As a result, $t_3 = T$, which means that t_3 has no variance. Unfortunately $t_{(hj)3} = T[n_j/(n_1 - 1)](m - 1)/m$ when $j \in s$ and $Tn_1/(n_1 - 1)$ otherwise. Thus, $(t_{(hj)3} - T)/T = O_p(1/m)$. Now v_{j3}/T^2 computed from the $t_{(hj)3}$ would also be $O(1/m)$ since it is the sum of n_1 terms of order $O(1/m^2)$.

Although v_{j3}/T^2 is $O(1/m)$, v_{j3} is not close enough to zero for our purposes. To see why, observe that if the y_i were all $N(1,1)$, then the relative variance of t_3 would be $1/m$, which is also $O(1/m)$. Thus, for v_{j3} to be nearly zero, v_{j3}/T^2 would have to be smaller than $O(1/m)$. It is not, and the jackknife variance estimator is not nearly unbiased.

As a counter-example to the replicate form in equation (17), consider the situation where each cluster is again a single element and all y_i values are equal to 1, but now $H = m$, $G = 1$, the population size in each h is N_0 , $n_h = 2$ for all h , and $M_1 = 2m$. As a result, $T = t_3 = mN_0$, so that t_3 has no variance. The replicate $t_{(hj)3}$ can take on four possible values. If $hj \in s$ and $hj' \in s$ ($j \neq j'$), then $t_{(hj)3}^* = [(m/2)(2m - 1)/(m - 1)]N_0$. If $hj \in s$ and $hj' \notin s$, then $t_{(hj)3}^* = [(m - 1)/2](2m - 1)/(m - 1)N_0$. If $hj \notin s$ and $hj' \in s$, then $t_{(hj)3}^* = [(m/2)(2m - 1)/m]N_0$. If $hj \notin s$ and $hj' \notin s$, then $t_{(hj)3}^* = [(m - 1)/2](2m - 1)/mN_0$. In all cases, $(t_{(hj)3}^* - T)/T = O_p(1/m)$, and so the jackknife variance estimator fails to be nearly unbiased.

The Two-phase Regression Estimator

To support the arguments in the text about the regression estimator in equation (21), we assume the sampling design and population values are such that the following asymptotic relationships hold. First,

$$\sum_{i \in S} w_i x_i (\sum_{i \in S} w_i e_i d_i x_i' x_i)^{-1} d_i x_i' - 1 = O_p(1/\sqrt{m}), \quad (A6)$$

which is a generalization of equation (A1). Likewise, equations (A2) and (A3) generalize to

$$\sum_{i \in S_g} w_{hji} d_i q_i / \sum_{i \in S_g} w_i d_i q_i - 1 = O_p(1/m), \quad (A7)$$

and

$$\sum_{i \in S_g} w_{hji} e_i d_i q_i / \sum_{i \in S_g} w_i e_i d_i q_i - 1 = O_p(1/m) \quad (A8)$$

for all q_i , where q_i is an element of the matrix $x_i' x_i$. Finally, the assumption in equation (A4) generalizes to

$$\sum_{i \in S_g} w_{hji} d_i p_i / \sum_{i \in S_g} w_i d_i p_i - 1 = O_p(1/m) \quad (A9)$$

for all p_i , where p_i is an element of the matrix $x_i' y_i$.

REFERENCES

- ISAKI, C.T., and FULLER, W.A. (1982). Survey design under the regression superpopulation model. *Journal of the American Statistical Association*, 77, 89-96.
- KOVAČEVIĆ, M.S., and YUNG, W. (1997). Variance estimation for measures of income inequality and polarization – an empirical study. *Survey Methodology*, 23, 1, 41-52.
- KREWSKI, D., and RAO, J.N.K. (1981). Inferences from stratified samples: properties of linearization, jackknife, and balanced repeated replication methods. *Annals of Statistics*, 9, 1010-1019.
- OH, H.L., and SCHEUREN, F.J. (1983). Weighting adjustment for unit nonresponse. *Incomplete Data and Sample Surveys, Volume 2: Theory and Bibliographies*, (Eds. W.G. Madow, I. Olkin, and D.B. Rubin). New York: Academic Press, 143-184.
- RAO, J.N.K., and SHAO, J. (1992). Jackknife variance estimation with survey data under hot deck imputation. *Biometrika*, 79, 4, 811-822.
- RAO, J.N.K., and WU, C.F.J. (1985). Inferences from stratified samples: Second-order analysis of three methods for nonlinear statistics. *Journal of the American Statistical Association*, 80, 620-630.
- RUST, K. (1985). Variance estimation for complex estimators in sample surveys. *Journal of Official Statistics*, 1, 381-397.
- SÄRNDAL, C.-E., SWENSSON, B., and WRETMAN, J.H. (1992). *Model Assisted Survey Sampling*. New York: Springer-Verlag.
- SINGH, M.P., DREW, J.D., GAMBINO, J.G., and MAYDA, F. (1990). *Methodology of the Canadian Labour Force Survey: 1984-1990*. Catalogue No. 71-526, Statistics Canada.
- STUKEL, D.M., and BOYER, R. (1992). Calibration Estimation: An Application to the Canadian Labour Force Survey. Methodology Branch Working Paper, SSMD, 92-009E. Statistics Canada.
- WOLTER, K. M. (1985). *Introduction to Variance Estimation*. New York: Springer-Verlag.

A Synthetic, Robust and Efficient Method of Making Small Area Population Estimates in France

GEORGES DECAUDIN and JEAN-CLAUDE LABAT¹

ABSTRACT

Since France has no population registers, population censuses are the basis for its socio-demographic information system. However, between two censuses, some data must be updated, in particular at a high level of geographic detail, especially since censuses are tending, for various reasons, to be less frequent. In 1993, the Institut National de la Statistique et des Études Économiques (INSEE) set up a team whose objective was to propose a system to substantially improve the existing mechanism for making small area population estimates. Its task was twofold: to prepare an efficient and robust synthesis of the information available from different administrative sources, and to assemble a sufficient number of "good" sources. The "multi-source" system that it designed, which is reported on here, is flexible and reliable, without being overly complex.

KEY WORDS: Population estimates; Administrative files; Robust estimation.

1. INTRODUCTION

In France, as in all countries that do not have population registers, censuses of the population are the cornerstone of the socio-demographic information system. However, censuses are quite massive operations that cannot at present be carried out more often than once every seven or eight years. In the interval between censuses, it is therefore necessary to update some information, especially at a high level of geographic detail, particularly since for various reasons, censuses are tending to be less frequent. Thus, small area population estimates are a major challenge for the Institut National de la Statistique et des Études Économiques (INSEE).

Despite the progress achieved in this field, the situation in 1993 still seemed fairly unsatisfactory. When figures from the 1990 population census were compared to the population estimates made on the basis of the previous census (1982) for the metropolitan departments, the differences noted were sometimes sizable.

INSEE therefore created a methodology team whose mission was to propose a system that would substantially improve the existing mechanism. Initially, the next census was to take place in 1997. It therefore seemed reasonable to have the new system operate on an experimental basis until the census, so as to see how well it worked before using it in actual production. When the census was postponed to 1999, it became more necessary to bring the project to a successful conclusion quickly, so as to be able to use the new system in 1996.

To achieve its objective, the team devoted itself, with maximum pragmatism, to a twofold task: to develop an efficient and robust synthesis of the information available from different administrative sources, and to assemble a sufficient number of "good" sources. The "multi-source" system that it designed, which is described here, is not overly complex and seems effective. A more detailed description of it is provided in Decaudin and Labat (1996).

2. MAIN CONCLUSIONS

The team's main conclusions are as follows:

- 1) It is impossible to improve total population estimates using sample surveys, unless the survey is conducted on such a scale that it would be similar to a census.
- 2) No single administrative source adequately reflects changes in the population. At the local level, all sources can exhibit drift, breaks, jolts, *etc.*, which are not always easy to detect. Furthermore, even at the local level, it is often quite difficult if not impossible to get the agency responsible to provide explanatory details, much less corrections in the case of errors. In any event, it is unwise to rely on a single administrative source, however good it may be, since its permanency is never guaranteed.
- 3) On the other hand, total population estimates can be improved substantially by simultaneously using several sources. A "multi-source" system, similar to the one presented here but more rudimentary, was tested retrospectively over the intercensal period 1982-1990, for the 96 metropolitan departments. The mean error (mean deviation as an absolute value from the results of the March 1990 census) fell below 0.9%, whereas the mean error registered at the time, with the estimation system then in place, was 1.4%.

3. SIMULTANEOUS USE OF SEVERAL SOURCES

For using several sources jointly, different methods are possible.

A method that is universal – and easy to implement – is multiple regression. In simplified form, this amounts to using, for any area z , the following relationship:

$$P(n+1, z)/P(n, z) = c + \sum_S (k_S N_S(n+1, z)/N_S(n, z)),$$

¹ Georges Decaudin and Jean-Claude Labat, Institut National de la Statistique et des Études Économique, 18, Blvd. Adolphe-Pinard, 75765 Paris, CEDEX 14.

where $P(n, z)$ is the population of area z on January 1 of year n , the values $N_s(n, z)$ are the numbers from each source S on the same date and k_s are coefficients, which are estimated by multiple regression over a past period. Here c is a constant term that is used only in the regression, with calibration on the national population serving to correct any drift.

This method is used in various countries, including Canada and the United States (for example, see Statistics Canada 1987 and Long 1993). Nevertheless, it was not adopted because it has numerous drawbacks:

- it must be possible to estimate the coefficients, which requires data from each source extending back over a fairly long period;
- the coefficients can change over time, without it being possible to control this change;
- as noted above, the administrative sources are, for various reasons (changes in regulations, abrupt shifts in management, errors, etc.), subject to what might be called "anomalies". For each source S , the scope of these anomalies is reflected in part in the coefficient k_s , to an extent that depends on how great their medium-term effect has been over the calibration period [la période d'étalonnage]; but anomalies nevertheless occur in estimates with the same weight as the "good" data from the same source. The estimates are then highly distorted.

Another method is known as the "composite" method. Each source is used to estimate the population in one or more age classes: age class X , which is well-covered by the source, but also sometimes another class that definitely exhibits a pattern very similar to that of class X (for example, the "30-45" age group, if X represents the "under 18" age group). It is then necessary to have appropriate indicators for the other components of the population and correctly manage the consolidation of these estimates "in parts".

This type of method, used in the United States (Long 1993), seemed to us to be problematic, especially because of the difficulty of adequately dealing with "anomalies".

The proposed "multi-source" system is based on a robust synthesis of estimates from different sources. It combines demographic reasoning with purely statistical techniques. It draws on the experiments conducted by the INSEE's regional directorate in Brittany in the early 1970s (Laurent and Guéguen 1971; Guéguen 1972). Should one of the sources fail, such a system is not prevented from functioning, even though its performance may be somewhat diminished.

4. DEMOGRAPHIC BASE

The demographic reasoning which is at the base of the system is elementary: assuming that we know the total population $P(n)$ for an area on January 1 of year n , the population $P(n+1)$ of the area on January 1 of year $n+1$

is deduced by summing the two components of the change during year n : natural increase (births minus deaths), and net migration (immigrants minus emigrants).

$$P(n+1) = P(n) + N(n) - D(n) + I(n) - E(n).$$

In France, natural increase data are provided annually at the commune level by vital statistics. If the latter are not yet available in final form, which is often the case in the third quarter of year $n+1$, it is easy to estimate them with a low margin of uncertainty.

The only unknown, then, is net migration for year n : $SM(n) = I(n) - E(n)$ or what amounts to the same thing, the net migration rate $T(n) = SM(n)/P(n)$. In other words, estimating the population comes down to estimating net migration since the last date on which the population is known (or is assumed to be known), and vice versa.

In France, net migration figures are of some importance, although less so than in other countries such as Canada or the United States. In addition, they generally exhibit a certain inertia, at least at relatively aggregated geographic levels. One way to assess the influence of changes to them from one intercensal period to the next is to measure the errors that would have been committed during each period if the population had been estimated by using the average annual net migration rates for the preceding period. Over the period 1982-1990, for the departments (excluding Corsica), the mean end-of-period error (in 1990, at the end of eight years) would have been only 1.3%. It was not certain, when the team started its work, that much greater accuracy could be achieved. However, both in 1975 and in 1982, the mean error that would have been committed with the trend method would have been much greater: 2.8% and 2.7% respectively (over seven years). It would therefore seem that the period 1982-1990 was exceptional and that in the future the difference will again be more pronounced.

5. ESTIMATES FROM THE DIFFERENT SOURCES

From each source, using an appropriate method, we draw an estimate of annual net migration rate for the population as a whole. The methods that may be used depend on the data available.

For each of the sources tested and found to be "good", at least at the departmental level, a method is proposed. The five sources retained are the following: housing tax; electrical utility customers; children receiving family allowances; educational statistics; electoral file.

The data on the composition of households for tax purposes, which appear in the income tax files, are the sixth source that should provide very good results. However, to date, these data have been analysed for only a few departments, and the methodology for using them is not yet completely defined.

We also propose to integrate a trend estimate of the net migration rate into the system.

Two categories of methods are used. The first concerns the sources relating to households; the second concerns those relating to individuals.

5.1 Sources Relating to Households

Some sources provide information on changes in the number of households. This is the case with the files on *housing taxes* (HT) and *electrical utility customers* (EUC). The housing tax is one of the four main local direct taxes. As its name indicates, it applies to occupied dwellings, with main residences and secondary residences being treated separately. The housing tax file takes account of the situation on January 1 of the taxation year. Starting in the 1980s, the HT source was the basis for the departmental population estimates developed by INSEE (Descours 1992). In the early 1990s, it was replaced by the EUC source, in light of the distortions caused by a change to the HT management system which gradually worked its way through all departments.

The method adopted for using these sources follows classical principles. It leads directly to an estimate of the total population, and it involves three main stages:

- 1) estimating the number of households;
- 2) estimating average household size and from there, estimating the population of households;
- 3) adding the "non-household" population.

In the first stage, it is assumed that the number of households changes in accordance with the data supplied by the source (number of main residences for HT purposes or number of electrical utility customers). The second stage is more delicate. It is based on both the use of statistics on dependants from the HT files and on a trend estimate of average household size.

In the proposed "multi-source" system, we move on to the net migration rate, for comparison with other sources, using vital statistics data (*cf.* Section 4).

5.2 Sources Relating to Individuals

The other sources used concern individuals. Only a certain age group X of the population is generally covered adequately. The method then involves two main stages:

- 1) estimating, from the source, the net migration rate for the population aged X ;
- 2) from there, estimating the net migration rate for the population as a whole.

The second stage is based on the following statistical relationship, observed in the past, between the change, from one period to another, of the overall net migration rate (T) and the change in the net migration rate for the population aged X (TX):

$$T_2 - T_1 = \delta_X (TX_2 - TX_1),$$

where δ_X is a coefficient close to 1, depending on the age group X . This relationship is similar to the one used by

de Guibert-Lantoine (1987) to estimate the population on the basis of educational statistics.

For the corresponding age groups in the different sources used, the values, estimated by linear regression, of the coefficient δ_X (± 2 standard deviations) are shown in tables 1 and 2.

Table 1
Estimates of δ_X on Departments, Excluding Corsica,
Internal Net Migration

Period 1	Period 2	Age at end of period		
		0-19	10-14	35 and over
1962-1968	1968-1975	0.76 (+/- 0.04)	0.69 (+/- 0.06)	1.24 (+/- 0.09)
1968-1975	1975-1982	0.77 (+/- 0.03)	0.88 (+/- 0.06)	1.56 (+/- 0.08)
1975-1982	1982-1990	0.70 (+/- 0.11)	0.49 (+/- 0.10)	1.26 (+/- 0.17)

Table 2
Estimates of δ_X Over the Two Periods 1975-1982 and
1982-1990, Excluding Corsica, Total Net Migration

	Age at end of period		
	0-18	9-15	35 and over
Departments	0.65 (+/- 0.11)	0.57 (+/- 0.10)	1.22 (+/- 0.16)
Department - employment zone	0.65 (+/- 0.04)	0.59 (+/- 0.04)	1.17 (+/- 0.06)

The approach followed in the first stage depends on the source:

Electoral File

Annual migration figures for voters in the selected age group (30 and over) are supplied directly by the electoral file managed by INSEE. We go from the rate of net migration of voters to the residential net migration rate by dividing the former by a coefficient reflecting the magnitude of the change in the electoral file.

Educational Statistics

The net migration figure for those in the 5-9 age group is obtained by subtracting their number in year n from that of the same cohorts the next year (that is, from those in the 6-10 age group in year $n + 1$) and deducting deaths.

Children Receiving Family Allowances

The number of persons in the 0-17 age group is estimated on the assumption that it evolves similarly to the number of children receiving family allowances. From this a figure for the net migration of young persons is obtained by comparing this estimate to a hypothetical change in the youth population without migration, that is, a change due solely to natural increase.

6. SYNTHESIS

6.1 Principles

The different basic estimates of the annual net migration rate are treated statistically in order to obtain a "synthetic rate", to be used as the final estimate. The treatment serves to eliminate outliers, underweight suspect values and, more generally, assign to each source a weight that reflects its performance.

More specifically, since each source can "drift", the different basic estimates are generally biased; they are first corrected for the national bias of the corresponding source for the year considered, a bias that is estimated in advance. In proceeding in this way, we implicitly assume that the difference between the local bias and the national bias is minor in relation to the irreducible unexplained portion of the difference (flou irréductible). Once we have estimates for a number of years, it should be possible to test this hypothesis and if necessary, replace it with one that corresponds more closely to reality, so as to improve the correction of biases at the local level.

It should be noted that such a seemingly simple operation as correcting the national bias nevertheless requires several precautions. The solution that consists in carrying out a gross calibration on the national net migration rate, considered by definition as a good reference, is not very satisfactory, owing to anomalies that may distort the calibration. It is therefore preferable to estimate the biases by means of a process in which we also eliminate anomalies. The process is similar to the one used for synthesis, which is described below. However, the determination of biases, assumed to be national in scope and therefore calculated for 96 departments, is less sensitive to anomalies than the determination of synthetic rates, calculated over a small number of sources. Only major anomalies are likely to significantly throw off the calibration of the rates and must therefore be corrected.

The "synthetic" net migration rate is a weighted mean of the basic estimates thus calibrated. Each source S is assigned an initial weight W_S that is supposed to reflect its medium-term accuracy. But in addition, for a given year and area, this weight is modulated to take account of the plausibility of the corresponding rate. Thus, if a rate is "abnormally distant" from the rates obtained from other sources – in practice, from a central value for all rates for the area – its weight is cancelled or reduced. For this, we look at the distance between the rate obtained from each source and the central value identified, and we compare it to a "norm" of distance NO_S specific to the source, determined empirically on the basis of the data available: if the distance is less than " a times the norm", the weight is not automatically changed; if it is greater than " b times the norm", it is set at 0; between the two, the weight is multiplied by a coefficient, included between 0 and 1, calculated by interpolation.

Note that the trend estimate is formally treated like those from exogenous sources; its weight is cancelled when it is

considered as implausible because it is too far from the other estimates.

The synthesis is achieved automatically, which ensures homogeneity and an explicit logic to the treatments carried out. This does not, however, eliminate the need to control the results obtained.

6.2 Theoretical Presentation

On the theoretical level, we sought to use reasonings and robust estimation techniques, such as described in Hoaglin, Mosteller and Tukey (1983). The method adopted falls within the framework of M -estimators of central tendency and more specifically in the category of W -estimators, which use the reweighted least squares algorithm.

Since the net migration rates for year n and area z obtained from different sources S (and corrected for their national biases) are denoted $TC_S(n, z)$, the synthetic rate $T(n, z)$ solves the implicit equation:

$$\sum_S W_S \cdot NO_S \cdot \Psi\left(\frac{TC_S(n, z) - T(n, z)}{NO_S}\right) = 0,$$

where the function Ψ is of the type that redescends to a finite rejection point:

$$\begin{aligned} \Psi(r) &= r & \text{for } |r| \leq a, \\ \Psi(r) &= r \frac{b - |r|}{b - a} & \text{for } a < |r| \leq b, \\ \Psi(r) &= 0 & \text{otherwise.} \end{aligned}$$

Using an iterative process, we can gradually refine the automatic processing of suspect data.

6.3 First Analysis of the Distances From Each Rate to the Central Value for the Rates

- 1) For each area z we calculate a first central value of the "calibrated" rates $TC_S(n, z)$. The central value used must not be overly sensitive to the possible existence of quite distant values for some sources, but at the same time it must be influenced by a source to the extent that the source is on average more accurate. Under these conditions, rather than choosing the median – which would meet the first condition – we use a statistic of rank that is a little more elaborate but nevertheless simple, owing to the small number of values; this statistic is the mean, weighted by respectively 1/2, 1/4, 1/4, of the three quartiles:
 - the median of the rates $TC_S(n, z)$ weighted by the initial weights W_S ,
 - the lower quartile (Q1) of the weighted rates,
 - the upper quartile (Q3) of the weighted rates.
- 2) The rates $T1(n, z)$ thus obtained are calibrated on the net migration rate for the higher level, by simple translation:

$$TC1(n, z) = T1(n, z) +$$

$$TREF(n) - \sum_z (T1(n, z)P(n, z)) / \sum_z P(n, z)$$

where $P(n, z)$ is the population of area z on January 1 of year n and $TREF(n)$ is the net migration rate for the higher level (the national rate for the departmental synthesis).

- 3) For each area, we calculate the differences between each rate and this calibrated central value:

$$EC1_S(n, z) = |TC_S(n, z) - TC1(n, z)|.$$

- 4) For each source and each area, the size of this difference is assessed in relation to the "norm" of distance NO_S specific to the source. This "norm" is determined empirically on the basis of the available data: theoretically it is the average of the distances observed in the past, excluding anomalies. The result is a first modulation of the weight originally assigned to this source:

- if $EC1_S(n, z) \leq a1 NO_S$, where $a1$ is a parameter to be chosen (in the vicinity of 2), we do not change W_S , the initial weight for S . In other words, if $WM1_S(n, z)$ is the modulation coefficient of W_S (coefficient included between 0 and 1), we take $WM1_S(n, z) = 1$;
- if $EC1_S(n, z) > b1 NO_S$, where $b1$ is another parameter (in the vicinity of 3), we set W_S at 0, meaning that we eliminate source S : $WM1_S(n, z) = 0$;
- if $a1 NO_S < EC1_S(n, z) \leq b1 NO_S$, we interpolate $WM1_S(n, z)$ as a function of the value of $EC1_S(n, z)$:

$$WM1_S(n, z) = (b1 NO_S - EC1_S(n, z)) / ((b1 - a1) NO_S).$$

- 5) At the end of this first phase, we therefore have new weights specific to each source and each area, which would allow us to locally eliminate or underweight suspect rates: $W1_S(n, z) = W_S WM1_S(n, z)$.

6.4 Iterations

- 1) Using the weights thus modified $W1_S(n, z)$, we estimate a new central value for each area, this time taking the weighted average of the rates:

$$T2(n, z) = \sum_S (TC_S(n, z) W1_S(n, z)) / \sum_S W1_S(n, z).$$

- 2) We calibrate each rate $T2(n, z)$ on the net migration rate for the higher level, by translation. We obtain $TC2(n, z)$.
- 3) We calculate, in each area, the differences between each rate and the calibrated average rate: $EC2_S(n, z) = |TC_S(n, z) - TC2(n, z)|$. Using these differences, we calculate new modulation coefficients for the initial weights, using the parameters $a2$ and $b2$, which may be different from $a1$ and $b1$ (theoretically they would be lower). We thus obtain new weights $W2_S(n, z)$ which more effectively take account of anomalies, since the

latter are assessed in relation to a better central tendency. With these weights, we estimate a new synthetic rate $T3(n, z)$, which is calibrated on the higher level to obtain $TC3(n, z)$.

- 4) The operations described in point 3 are repeated with the same parameters $a2$ and $b2$. The tests conducted at the departmental level over the period 1982-1990 show that the convergence is generally rapid; the rates are quite often stabilized by the fourth iteration.

7. IMPLEMENTATION AT THE DEPARTMENTAL LEVEL

The estimation system outlined above, which is operationalized for 1990 and subsequent years, was implemented by the project team for the year 1990 at the departmental level, with the following five sources: housing tax (HT), electrical utility customers (EUC), family allowances (FA), educational statistics (ES), electoral file (EF), plus the trend estimate (TREND).

Figure 1 shows the results obtained for several departments. Table 3 shows the values of the weights and norms used to make the system operate. This table also shows certain statistics obtained from the synthesis of the net migration rates; in particular they concern the differences between the rates obtained from each source and the synthetic rates.

Table 3
Implementation for Year 1990 at Department Level
Parameters and Statistics

	HT	EUC	FA	ES	EF	TEND
Weight	115	100	80	70	80	100
Norm	0.15	0.17	0.19	0.20	0.19	0.12
Number of rates	96	96	89	96	94	96
Average distance	0.55	0.14	0.30	0.19	0.14	0.13
Number of "aberrant" rates	37	2	17	3	1	6
Average of distances without "aberrant" rates	0.15	0.13	0.16	0.16	0.13	0.11

Note: - Coefficients (a ; b) applied to norms: (2,5; 3,5) in the first iteration, then (2; 3).
 - The values of the distances and norms correspond to rates expressed as a %.
 - Distances are calculated in relation to the synthetic rates after three iterations.
 - "Aberrant" rates are those for which the weight is cancelled after three iterations.

The results suggest that the system is even more effective than indicated by the summary retrospective test carried out on the 1982-1990 intercensal period with the same sources. Aside from the HT source, which is still distorted, the estimates from the different sources are more convergent than they were on average in the retrospective test (see Table 4).

There is nothing surprising about this, given the rudimentary state of the system tested on the 1982-1990 intercensal period. The data used were rough or even

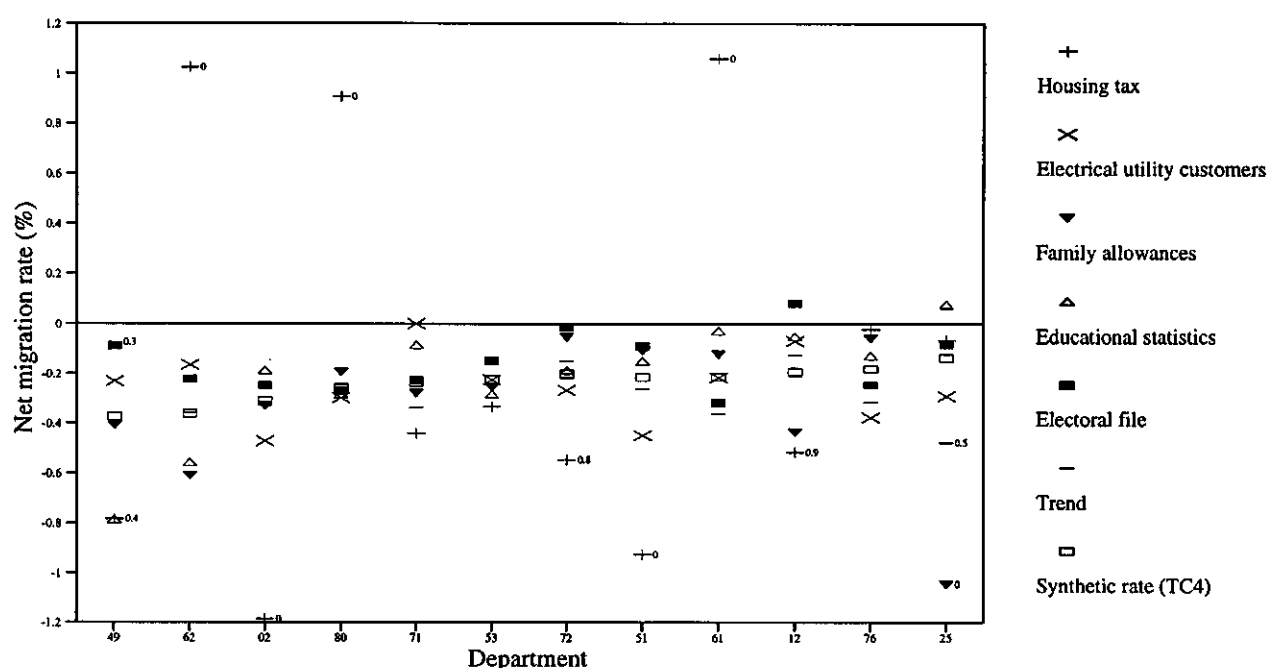


Figure 1: Summary of Net Migration Rates for 1990 for Twelve Departments, Identified by Number (49, 62, etc.). Note: TC4 is the synthetic rate obtained after three iterations. Where the weight for a source has been eliminated or reduced, the value of the modulation coefficient (WM3) is shown.

fragmentary, owing to the difficulty of assembling, in 1993, management data for years past (1982, ...); in addition, the relationships used to draw an estimate of the net migration rate from each source were simplistic; and lastly, the method of synthesis was less elaborate.

It should be noted that the integration of other sources – income tax data in particular – can only further reinforce the effectiveness of the system.

8. SUPPLEMENTS

8.1 Sub-Departmental Levels

The use of some sources may become risky at a geographic level below the departmental level. There are various reasons for this: because the hypotheses on which the method is based become fragile, because the numbers are small, etc. This is especially the case with educational statistics.

However, it should be possible to operate the system for employment areas, or more specifically for cross-tabulations of department and employment area (there are approximately 420 such areas), which serve to ensure consistency with the departmental level. This should not involve too many risks, for the following reasons:

- a certain deterioration of performance in relation to the departmental estimates is acceptable, especially since the departmental estimates should be of good quality;
- the data from the income tax files should be quite useful;
- trend estimation and calibration on estimates at higher geographic levels (in this case the departmental estimates) both act as safeguards.

Of course, there is nothing prohibiting the use of the system to produce estimates for other sub-departmental geographic units.

At the departmental level, it does not seem useful to adapt the parameters (initial weights and norms) to population size; on the other hand, for sub-departmental

Table 4
Mean of Distance in Retrospective Test

	TH	EDF	AF	EN	FE
1982	0.26	0.34	0.50	0.47	0.34
1983	0.28	0.33	0.48	0.47	0.32
1984	0.23	0.28	0.40	0.45	0.34
1985	0.24	0.31	0.48	0.44	0.32
1986	0.23	0.33	0.40	0.33	
1987	0.40	0.28	0.41	0.27	
1988	0.84	0.29	0.30	0.37	0.24
1989	0.97	0.21	0.30	0.33	0.35
Overall mean	0.43	0.30	0.41	0.39	0.32

Notes: -The number of rates per year is generally 96, except for FA (89) and EF (94).
 -The "electoral file" source did not provide rates for 1986 or 1987.
 -The "housing tax" source began to be distorted in 1987.
 -The values of the differences correspond to rates expressed as a %.

levels, such an adaptation appears essential. Otherwise we run the risk of being much too strict for small areas. It would seem that a norm function of the following type might be appropriate:

$$NO_S = \alpha P^\beta,$$

where NO_S is the norm for source S , P is the population of the area and α and β are two parameters that hypothetically depend on source S . The parameter β is obviously negative. If β equals -0.25 , the norm doubles when the population is divided by 16. It also appears that the type of geographic area has an effect: the unexplained portion (le flou) would on average be greater for a commune of 50,000 inhabitants than for an employment area of the same size. The parameters α and β must be defined for each sub-departmental source, and where applicable, for each type of area.

8.2 Timetable

The greater the number of sources, the better the system functions. However, for a given year, data from the different sources become available at different times. Since the system is able to function with a variable number of sources, one can develop, at least at the departmental level, several sets of estimates for January 1 of year n : for example, interim estimates in the third quarter of year n , based on the first sources available, then semi-definitive estimates in the third quarter of year $n + 1$, based on more sources, and then final estimates in the third quarter of year $n + 2$. Different factors must be taken into account: the complexity of an operation, and the magnitude of the changes due to the addition of a source. It will be possible to assess the latter factor by simulations on the first years of implementation of the system.

8.3 Integration of an Additional Source

The system is flexible and modular. Therefore, integrating a new source into it does not pose any particular problem. It is merely a matter of determining the method to be used in order to obtain a good estimate of the net migration rate for each area. The range of methods envisaged by the team is large enough that in most cases, it should be possible to find a type of method that is appropriate to the source.

To determine the parameters (initial weight and norm) to be assigned to the new source in the synthesis, we suggest putting the system through a dry run, with parameters set arbitrarily but reasonably; it is obviously wise to start with a fairly high norm and a fairly low weight. By analysing the differences obtained between the net migration rates obtained from the new source and the synthetic rates, a better norm can be determined. The weight can then be adapted accordingly, using (for lack of anything better) an assumed relationship of quasi-proportionality between the weight and the inverse of the square of the norm. Obviously, this process can be iterated, with the parameters

of the other sources also being changed as required. However, the tests conducted at the departmental level on the period 1982-1990 appear to show that the overall performance of the system is not highly sensitive to changes – even sizable ones – in the initial weights; it is therefore not necessary to determine these weights with great precision – nor, indeed, is it possible to do so – before the next census.

9. CONCLUSION

The “multi-source” population estimation system presented here is robust and flexible, without being overly complex. It can function with a variable number of sources. To integrate a new source into it, no long historical observation period is required. Aberrant data are detected automatically and corrected, so that they do not distort the estimates. The experiments carried out, while still not numerous, indicate that this system is effective. After a debugging and break-in period, it should be possible to use the system in production without too many risks pending the results of the next population census, planned for 1999.

ACKNOWLEDGEMENTS

This article results from the thinking and efforts of a team, led by the authors, which consisted of: Xavier Berne, Michel David, Michel De Bie, Sophie Destandau, Jacques Leclercq, Françoise Lemoine, Catherine Marquis and Marc Simon. The team benefited from the assistance of several departments of INSEE. The Statistical Methods Unit and its chief, Jean-Claude Deville, deserve special mention. The authors also wish to thank Philippe Ravalet for his contribution to the theoretical aspect of this article, as well as the editorial staff of *Survey Methodology* and the members of the editorial jury for their constructive comments.

REFERENCES

- DECAUDIN, G., and LABAT, J.-C. (1996). Une méthode synthétique, robuste et efficace, pour réaliser des estimations locales de population. Document de travail de méthodologie statistique, n° 9601. INSEE. Paris.
- DESCOURS, L. (1992). Estimation de populations locales par la méthode de la taxe d'habitation. *Actes des Journées de méthodologie statistique*, 13 and 14 March 1991. INSEE. Paris.
- GUÉGUEN, Y. (1972). Estimation de la population des villes bretonnes au 1.1.1971. *Sextant*, n° 4. INSEE. Rennes.
- de GUIBERT-LANTOINE, C. (1987). Estimations de population par département en France entre deux recensements. *Population*, 6, 881-910.
- HOAGLIN, D.C., MOSTELLER, F., and TUKEY, J.W. (1983). *Understanding Robust and Exploratory Data Analysis*. New York: John Wiley.

- LAURENT, L., and GUÉGUEN, Y. (1971). Essai d'estimation de la population des villes bretonnes. *Sextant*, n° 1. INSEE, Rennes.
- LONG, J.F. (1993). Postcensal Population Estimates: States, Counties and Places. Population Division. Technical Paper No 3. U.S. Bureau of the Census. Washington DC.
- STATISTICS CANADA (1987). *Population Estimation Methods, Canada*. Catalogue No. 91-528E. Ottawa.

An Adaptive Procedure for the Robust Estimation of the Rate of Change of Investment

PHILIPPE RAVALET¹

ABSTRACT

The presence of outliers in survey data is a recurring problem in applied statistics, and the INSEE survey on industrial investment is not immune from this. The forecasting of the rate of growth of capital investment expenditures in industry therefore comes down to robust estimation of a total in a finite population. The first part of this article analyses the estimator currently used in the Investment Survey. We show that it follows a strategy of reweighting the linear estimator. But the strict dichotomy imposed between outliers – all assumed to be nonrepresentative – and other points is not fully satisfactory from either a theoretical or a practical standpoint. These flaws can be overcome by adopting a model-based approach and estimating by GM-estimators, applied to the case of a finite population. We then construct a robust adaptive procedure that determines the appropriate estimator on the basis of the residuals observed in the sample in cases where the residuals may be assumed to be symmetrical. Lastly, this method is applied to the data from the Investment Survey for the period 1990-1995.

KEY WORDS: Economic surveys; Outliers; Robust estimation; GM estimator; Adaptive procedure.

1. INTRODUCTION

Since 1952, the Institut National de la Statistique et des Études Économiques (INSEE) has been conducting an investment survey that provides estimates of the future trend of capital investment expenditures in industry, well before the National Accounts are released or the findings of exhaustive surveys are published. The estimation of the rate of investment growth is based on the declarations of some 2,500 company heads concerning their intentions to purchase capital goods.

The almost systematic presence of outliers in these data is a major problem. Outliers can seriously distort the estimate of the average growth rate and lead to unacceptable results. According to Chambers (1986), two types of outliers may be distinguished. Nonrepresentative points designate either measurement errors, which survey staff strive to correct during data collection, or unique individuals in the population. By contrast, representative outliers designate individuals which, while somewhat unusual, cannot be considered exceptional. There are undoubtedly similar individuals in the population not questioned, and the information that they contain must be integrated into the estimate.

The problem posed here is that of robust estimation of a total in a finite population with auxiliary information, a problem to which theory provides no definitive answer. Nevertheless, various techniques, reviewed in Lee (1995), can be applied. The estimation method currently used in the Investment Survey follows the logic of reweighting the linear estimator, following Hidiroglou and Srinath (1981). However, the identification and treatment of outliers are not entirely satisfactory. In particular, all outliers are assumed to be nonrepresentative, and the dichotomy between

“normal” points and outliers makes the estimation quite sensitive to the choice of outliers.

The introduction of a linear superpopulation model, which describes the change in investment at the level of individuals, enables us to better assess the unusual nature of an observation and determine how representative it is. Its estimation by means of GM-estimators is then an attractive alternative to the least squares method, whose absence of bias is quite costly in terms of variance. The adjustment of the weight function depends at the outset on characteristics of the population according to criteria now well described in the literature. Since these characteristics can change not only from one stratum to another but also over time, the significance of an adaptive procedure is obvious. On the basis of a first robust estimate, we determine the appearance of the distribution of residuals, and then we choose the estimator to be used according to a predefined rule. Following Hogg, Bril, Han and Yul (1988), we construct an adaptive procedure based on indicators of tail weight and concentration estimated from the sample, since the residuals are not expected to be asymmetrical. This procedure is applied to the data from the Investment Survey for the period 1990-1995.

2. ESTIMATOR FOR THE INVESTMENT SURVEY

2.1 Estimation Principle

In a finite population $U = \{1, \dots, N\}$, which here represents a stratum of the survey, a sample $s = \{1, \dots, n\}$ of size n , is drawn, and $\bar{s} = \{n+1, \dots, N\}$ designates the population not questioned. Each company is questioned on

¹ Philippe Ravalet, Division des enquêtes de conjoncture, INSEE, 15 Bd. G. Péri, BP 100, 92244 MALAKOFF CEDEX.

its investment expenditures for two consecutive years $t - 1$ and t , denoted respectively x and y .

Knowing the total amount X of investments for year $t - 1$ in the population, we can deduce from the estimate \hat{Y} of total investments for year t the average rate of change of equipment expenditures between $t - 1$ and t :

$$\hat{\theta} = \frac{\hat{Y} - X}{X}.$$

To simplify the notations, we define the parameter $\Theta = 1 + \theta = Y/X$, estimated by $\hat{\Theta} = \hat{Y}/X$.

The estimator currently used in the INSEE survey draws on the ratio method, with the level of investment in $t - 1$ as auxiliary information:

$$\hat{Y}_{\text{ratio}} = \frac{X}{\sum_s x_i} \sum_s y_i.$$

This estimator may be written as a weighted linear estimator:

$$\hat{Y}_{\text{ratio}} = \sum_s w_i z_i. \quad (1)$$

In this expression, $w_i = Xx_i/\sum_s x_j$ is the weight of individual i and $z_i = y_i/x_i$ is the annual change in its investment. Such an estimator will be sensitive to the presence of outliers on both z and w . An *atypical point* will exhibit a change z that is very different from that of the others, while an *influential point* will have a weight w that is large enough to attract, by leverage, the average rate of change of the stratum towards its own rate of change. Since the decisive criterion for characterizing an observation as an outlier is that the product wz is large enough to distort the estimate \hat{Y}_{ratio} , the distinction between atypical points and influential points is, of course, arbitrary. The generic term *large investors* (or LI for short) will designate these outliers as a group, while the term *extrapolatables* will refer to the other individuals in the sample.

Having carried out an *a posteriori* partition of the sample $s = \{\text{LI}\} \cup \{\text{extrapolatables}\}$, we estimate the total investments of the rest of the population \bar{s} on the basis of the behaviour of only the extrapolatable individuals according to the ratio method:

$$\hat{Y}_{\text{LI}} = \sum_s y_i + \left(\frac{\sum_{\bar{s}} x_i}{\sum_{\{\text{extra}\}} x_i} \right) \sum_{\{\text{extra}\}} y_i. \quad (2)$$

In (2), the weight of the extrapolatables $1 + \sum_{\bar{s}} x_i / \sum_{\{\text{extra}\}} x_i$ is quite strictly greater than the weight of the large investors, which is equal to 1.

2.2 Selection of Large Investors

The large investors are selected within each stratum on the basis of their influence on the estimation of Θ according to an iterative procedure. At the outset, all individuals are

assumed to be extrapolatable, and for each of them we calculate a not-taken-into-account index, measuring the impact on $\hat{\Theta}$ of its exclusion from the sample, $\text{NTIA} = (\hat{Y}_{\text{LI}}^i - \hat{Y}_{\text{LI}})/X$ where \hat{Y}_{LI}^i is the estimated total without individual i .

The firm with the largest NTIA index in absolute value is said to be a large investor. \hat{Y}_{LI} is then re-estimated with this new partition of U , and then the next large investor is identified. The selection stops when all extrapolatable individuals' have an influence on the estimate that is below a given threshold. The greater the number and mass of observations, the easier it is to verify this condition. Conversely, it will prove impossible to verify the condition if the number of individuals is too small; in that case, the survey manager merely makes sure that no individual has a much greater influence than the others, thus introducing an element of subjectivity into the procedure.

By this iterative mechanism, the usual phases of detection and treatment of outliers are carried out simultaneously. The main problem is that the status of an individual is not an intrinsic characteristic but instead depends on the composition of the sample. This can change from one survey to another. In addition, in certain hypothetical cases (Ravalet 1996), this procedure can lead to the unnecessary exclusion of some individuals, since at no point is the status of large investor called into question.

2.3 Strategy for Reweighting the Linear Estimator

The estimator LI in fact follows from the strategy for reweighting the linear estimator (1) presented by Hidioglou and Srinath (1981) using the example of estimation of a total without auxiliary information. Having already carried out a partition $s = s_1 \cup s_2$ of the sample distinguishing the outliers s_1 (numbering n_1) from the other observations s_2 , the authors propose to reduce, in $\hat{Y} = (N/n) \sum_s y_i$, the weight N/n of the outliers to a lower value λ by positing

$$\hat{Y}_\lambda = \lambda \sum_{s_1} y_i + \frac{N - \lambda n_1}{n - n_1} \sum_{s_2} y_i$$

and

$$\hat{Y}_\lambda = \sum_s y_i + \frac{N - n}{n - n_1} \sum_{s_2} y_i + n_1(\lambda - 1) \left[\frac{1}{n_1} \sum_{s_1} y_i - \frac{1}{n - n_1} \sum_{s_2} y_i \right].$$

The optimal value of λ that minimizes the mean square deviation of this estimator, whether or not conditional on the number of outliers in the sample, depends on several parameters of the population. Without prior information, the choice of λ is a delicate one.

Applied to the case of the estimator of the ratio with auxiliary variable x , this is written as:

$$\hat{Y}_{\text{ratio } \lambda} = \sum_s y_i + \sum_{\bar{s}} x_i \frac{\sum_{s_2} y_i}{\sum_{s_2} x_i} + (\lambda - 1) \left(\frac{\sum_{s_1} y_i}{\sum_{s_1} x_i} - \frac{\sum_{s_2} y_i}{\sum_{s_2} x_i} \right) \sum_{s_1} x_i. \quad (3)$$

The first two terms of the second member of (3) form an estimate of the total Y , under the implicit hypothesis that all outliers are in the sample, and the third is a correction taking account of the possible presence of outliers in the population not questioned. This correction is a function of the λ selected and the difference in average behaviour between the two types of individuals estimated in the sample.

When (2) and (3) are considered together, it may be seen that the estimator LI is formally equivalent to the case $\lambda = 1$. The use of \hat{Y}_{LI} thus implicitly assumes that the outliers have been correctly identified and are all non-representative. In Ravalet (1996), it was shown that these two hypotheses were unfortunately seldom verified in the context of the Investment Survey.

Since the identification procedure is manual and the criterion used is relatively *ad hoc* in the absence of any hypothesis on the population, it is not impossible that some outliers will escape selection. The use of the ratio on the extrapolatables then poses the problem of the robustness of the estimation in relation to the choice of large investors. In addition, it is unlikely that all these points are unique. The atypical points, which are especially numerous among small and medium-sized firms, should instead be considered as representative. However, choosing $\lambda > 1$ would inevitably raise the question of the robustness of the third term of (3).

To try to compensate for these defects, changes to the estimator \hat{Y}_{LI} are possible. For example, the mean of the extrapolatables may be replaced by a more robust estimator, and only the nonrepresentative points are designated as large investors. This technique fits into the more general framework of M-estimators, in which the existence of a model facilitates both the detection and treatment of outliers (Lee 1995). It is then no longer a matter of constructing a strict dichotomy between outliers and other points but rather of defining areas of varying representativeness.

3. ROBUST ESTIMATION BY GM-ESTIMATORS

3.1 The Linear Model and GM-Estimators

Assume the existence of a linear model ξ that links together, for the overall population U , investments x and y on dates $t-1$ and t .

$$\xi: y_i = \beta x_i + \epsilon_i$$

with

$$\begin{aligned} E(\epsilon_i) &= 0 \\ E(\epsilon_i \epsilon_j) &= 0 \quad \forall i \neq j, \\ V(\epsilon_i) &= \sigma^2 \eta(x_i) \end{aligned}$$

Slope β of the regression line passing through the origin in the superpopulation model is interpreted as the rate of change Θ in the population. The variance of y is assumed to be an increasing function of x and η is generally a power function: $\eta(x_i) = x_i^\gamma$.

According to the model, the best unbiased linear estimator (Brewer 1963 and Royall 1970) of the total is $\hat{Y}_{mc} = \sum_s y_i + \hat{\beta}_{mc} \sum_{\bar{s}} x_i$ where $\hat{\beta}_{mc} = (\sum_s x_i y_i / \eta(x_i)) / (\sum_s x_i^2 / \eta(x_i))^{-1}$ is the least squares estimator.

In the particular case $\eta(x) = x$, this expression reduces to $\hat{\beta}_{mc} = \sum_s y_i / \sum_s x_i$, estimator of the ratio. This unbiased estimator is effective only under the hypothesis of normality of the residuals, and it does not prove to be very robust.

The M-estimators (Huber 1981) serve to define a robust version of the least squares by replacing the square function, in the minimization program, with a function p that increases less rapidly:

$$\text{Min} \sum_s p \left(\frac{y_i - \beta_R x_i}{\sigma \sqrt{\eta(x_i)}} \right).$$

The M-estimator $\hat{\beta}_R$ is the solution of the following implicit equation:

$$\sum_s \psi \left(\frac{y_i - \hat{\beta}_R x_i}{\sigma \sqrt{\eta(x_i)}} \right) \frac{x_i}{\sqrt{\eta(x_i)}} = 0$$

where

$$\psi(t) = \frac{\partial p(t)}{\partial t}.$$

The function ψ , like Huber's function $\psi(t) = \text{Max}(-c, \text{Min}(t, c))$, depends on one or more adjustment constants c controlling the portion of observations that must be considered as outliers. This estimator will still be sensitive to the effect of outliers on the explanatory variable x . Therefore a more general class of estimators, called GM-estimators (Hampel, Ronchetti, Rousseeuw and Stahel 1986), is defined by means of the following implicit equation:

$$\sum_s w \left(\frac{x_i}{\sigma \sqrt{\eta(x_i)}} \right) \psi \left(\frac{r_i}{\sigma} v \left(\frac{x_i}{\sigma \sqrt{\eta(x_i)}} \right) \right) \frac{x_i}{\sqrt{\eta(x_i)}} = 0$$

with

$$r_i = \frac{y_i - \hat{\beta}_R x_i}{\sqrt{\eta(x_i)}}.$$

A choice usually made is Mallows' formulation: $v(t) = 1$ and $w(t) = 1/t$. Hence a robust estimator $\hat{\beta}_R$ will verify the implicit equation

$$\sum_s \psi \left(\frac{y_i - \hat{\beta}_R x_i}{\sigma \sqrt{\eta(x_i)}} \right) = 0. \quad (4)$$

In general, the parameter σ is unknown and must be replaced in this expression by a robust estimate $\hat{\sigma}$ of the dispersion of the residuals

$$\sum_s \psi \left(\frac{y_i - \hat{\beta}_R x_i}{\hat{\sigma} \sqrt{\eta(x_i)}} \right) = \sum_i \psi \left(\frac{r_i}{\hat{\sigma}} \right) = 0.$$

The estimator of the total will then be:

$$\hat{Y}_{\beta R} = \sum_s y_i + \hat{\beta}_R \sum_{\bar{s}} x_i. \quad (5)$$

This estimator is studied by Gwet and Rivest (1992). In general, it is not unbiased in relation to the sample design. Chambers (1986) proposes to correct that bias by introducing into (5) a third term that estimates it robustly:

$$\hat{Y}_{\text{Chambers}} = \sum_{i \in s} y_i + \hat{\beta}_R \sum_{i \in \bar{s}} x_i + \left(\frac{\sum_{i \in s} \frac{x_i / \hat{\sigma} \sqrt{\eta(x_i)}}{\sum_{j \in s} x_j^2 / \hat{\sigma}^2 \eta(x_j)} \psi_E \left(\frac{y_i - \hat{\beta}_R x_i}{\hat{\sigma} \sqrt{\eta(x_i)}} \right) \right) \sum_{i \in \bar{s}} x_i.$$

Choosing a bounded function ψ_E seems a good compromise between estimator bias and variance. For example, Welsh and Ronchetti (1994) opt for a Huber's function with a large adjustment constant $c = 15$. But the adjustment of ψ_E , without prior information on the density of the outliers, is always difficult.

3.2 Choice of Estimator

The desirable properties of ψ functions are now well known with reference to the problem of estimating a central tendency. They must be bounded, continuous, and equivalent to an identity in the vicinity of zero. Strictly monotone functions (Huber) are distinguished from redescending functions such as Tukey's biquadratic function, Andrew's sine and the Hampel or Cauchy function. Because their influence function tends toward zero, these estimators will be less sensitive to the presence of outliers than the Huber function. The speed of convergence

toward zero is an essential characteristic of redescending functions. Those that are nil at a finite distance (Hampel, Tukey or Andrew) exclude outliers from the estimation of β , whereas the others assign them low representativeness.

The choice and adjustment of the ψ function are difficult. They greatly depend on the nature of the data and more specifically on the distribution of the residuals (Hoaglin, Mosteller and Tukey 1983, Ch. 11). An idea, however approximate, of the appearance of the distribution of the residuals should make it possible to better target both the choice and the adjustment of the estimator, and hence to make the estimation more efficient. This intuitive remark is at the origin of adaptive procedures, presented in particular by Hogg (1974) and (1982). The idea is to evaluate the nature of the distribution of the residuals, calculated on the basis of an initial robust estimate (of the norm L_1 type, for example), using carefully selected robust indicators (tail weight, asymmetry, concentration, etc.). The existence of these indicators makes it possible, using a predefined decision rule, to select the appropriate estimator for this situation, and the implicit equation (4) is solved by taking the first robust estimate of β as an initial value.

The idea of an adaptive procedure appears all the more attractive since it systematizes the study that must precede the choice and adjustment of an estimator. That study can prove extremely costly if it must be performed manually for each stratum of the sample and repeated for each survey.

4. CONSTRUCTION OF AN ADAPTIVE PROCEDURE

This section describes the construction of an adaptive procedure for calculating the average rate of change of investment on the basis of economic survey data. Consequently, certain choices were made in light of the specific nature and characteristics of those data and are not necessarily transposable to other regression models. In particular, after checking the data, we adopted the hypothesis of a symmetrical distribution of residuals and we excluded the case of light-tail distributions.

The construction of an adaptive procedure, which draws on the works of Moberg, Ramberg and Randles (1980), is carried out in several stages. The first step is to choose the ψ function (or family of functions) to be used. The second is to select the various criteria for characterizing the distribution of residuals. Using these criteria, a classification rule is constructed. Finally, each class is matched with the adjustment of the estimator to be used.

4.1 Choice of ψ Function

Since Huber-type monotone functions do not provide sufficient protection against outliers, only redescending functions were considered. Among them, we selected the generalized Cauchy function (used in particular by Moberg *et al.* 1980 to approximate generalized lambda functions) and the Tukey biquadratic function:

$$\psi_c(r) = \frac{cr}{(b+r)^2 + c}, \quad \forall r$$

and

$$\psi_T(r) = \frac{r}{c} \left(1 - \frac{r^2}{c^2} \right)^2, \quad \forall |r| \leq c.$$

These two estimators are quite different in their treatment of outliers (see Figure 1). The biquadratic function equals zero for longer than the Cauchy function, but on the other hand it has a finite rejection point: the residuals beyond $c \cdot \sigma$ do not enter into the estimate, whereas the Cauchy function assigns them a certain representativeness. The parameter b serves, in principle, to control the asymmetry of ψ according to that of the residuals.

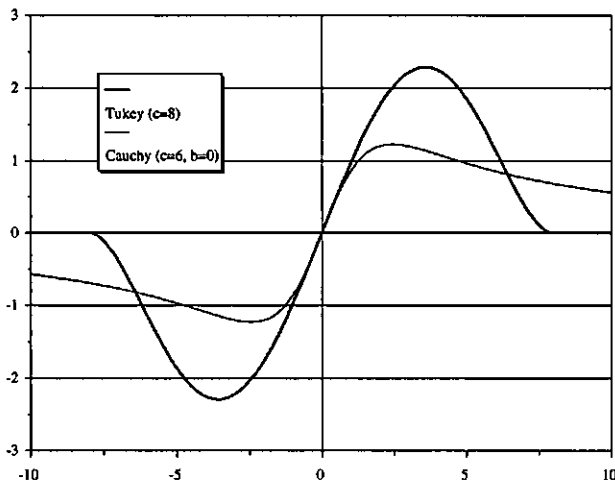


Figure 1. Cauchy and Tukey Functions

4.2 Parameter of Scale, Calculation Algorithm and Selection Criteria

In general an estimator $\hat{\sigma}$ of dispersion is defined by an implicit equation $\sum \chi(r_i/\hat{\sigma}) = 0$, where χ is an even function. It is therefore a matter of solving the system of non-linear equations in $(\hat{\beta}, \hat{\sigma})$ following:

$$\begin{cases} \sum_i \psi \left(\frac{y_i - \hat{\beta}x_i}{\hat{\sigma}\sqrt{\eta(x_i)}} \right) = 0 \\ \sum_i \chi \left(\frac{y_i - \hat{\beta}x_i}{\hat{\sigma}\sqrt{\eta(x_i)}} \right) = 0. \end{cases} \quad (6)$$

Rivest (1989) offers several examples showing that resolving system (6) can pose problems, owing to the fact that there may be a number of solutions, even in the case of a monotone ψ function. Following his recommendations, we will proceed in two stages. First, the parameter of dispersion σ is estimated using the median of the absolute values (MAD) of the residuals defined on the basis of the median of the individual rates of change. Then β is calculated by (4) using the value of σ found previously.

For solving (4), we preferred the reweighting algorithm to the Newton-Raphson algorithm, since it seems to converge more easily, especially when the adjustment constant is small.

Since the effectiveness of an adaptive procedure depends on the effectiveness of the decision-making process, the greatest attention must be paid to the nature, quality and robustness of the information that guides the choice of the estimator. Tail weight is an indispensable indicator, since it provides information on the relative significance of outliers in the sample and thus in the population (see Hoaglin *et al.* 1983, ch. 10). For the tail weight indicator, we adopted the proposal of Hogg (1974):

$$\tau(p) = \frac{\bar{U}(p) - \bar{L}(p)}{\bar{U}(0.5) - \bar{L}(0.5)}$$

$\bar{U}(p)$ (resp. $\bar{L}(p)$) is the mean of the np largest (resp. smallest) order statistics, using a linear interpolation when np is not whole. We chose $p = 0.05$; for the normal distribution $\tau(0.05)$ is equal to 2.59.

In addition, like Hogg *et al.* (1988), we considered it important to test for the possible presence of a distribution of the double exponential type, measuring the concentration of residuals by the following pk indicator:

$$pk = \frac{\bar{X}(1 - \beta, 1 - \alpha) - \bar{X}(\alpha, \beta)}{\bar{X}(0.5, 1 - \beta) - \bar{X}(\beta, 0.5)}$$

where $\bar{X}(a, b)$ is the means of the order statistics between the na -th and the nb -th, with the sizes interpolated if na or nb are not integers. We selected $\alpha = 0.05$ and $\beta = 0.15$, or $pk = 2.7$ for a normal distribution.

Finally, different studies (Moberg *et al.* 1980, Hogg *et al.* 1988) have emphasized the importance of the dissymmetry of distributions. When there are asymmetrical residuals, the bias of robust estimators can be sizable, making it tricky to use them (Chambers *et al.* 1993). In the INSEE Investment Survey, the residuals are theoretically asymmetrical since they are confined to a limited range ($r = y - \beta x \geq -\beta x$). However, we noted empirically that this asymmetry was very slight and could safely be ignored. The failure of the correction of a possible bias by the function ψ_E in Chambers' estimator moreover confirms this observation. Only the symmetrical case is considered here; the bias of the estimators defined by (5) is therefore nil.

4.3 Classification of Distributions and Adjustment of the Estimator

The definition of the decision rule was based on the study of eight specific symmetrical distributions illustrating various tail weight and concentration situations (see Table 1). We were interested in the family of contaminated distributions $CN(\alpha, K)$, with the distribution function $F(x) = (1 - \alpha)\Phi(x) + \alpha\Phi(x/K)$ where Φ is the cumulative function of the distribution $N(0, 1)$, since these distributions give a good representation of real data (Hoaglin *et al.* 1983, ch. 10), especially the data in the Investment Survey (Ravalet 1996). While Gaussian in the middle, they nevertheless contain more outliers than the normal distribution $N(0, 1)$.

Table 1
Eight Specific Distributions

	$\tau(.05)$	pk
1 Normal distribution	2.59	2.76
2 Contaminated dist $CN(.05, 3)$	2.94	2.83
3 Double exponential dist.	3.28	3.41
4 Contaminated dist $CN(.05, 10)$	4.47	2.85
5 Contaminated dist $CN(.10, 10)$	5.42	3.05
6 Contaminated dist $CN(.20, 10)$	5.64	4.44
7 Slash distribution	7.65	4.19
8 Cauchy distribution	7.82	4.78

The two indicators $\tau(0.5)$ and pk were simulated over these eight distributions, for several sample sizes. The graph of $(\tau(0.5), pk)$ serves to distinguish four groups of distributions: light-tailed, relatively unconcentrated distributions of the normal type or $CN(.05, 3)$; heavy-tailed distributions of the type $CN(.05, 10)$, $CN(.10, 10)$, and $CN(.20, 10)$, and very heavy-tailed distributions of the Slash or Cauchy type; and concentrated distributions such as the double exponential distribution. These four classes are defined (see Figure 2) by the following equation boundaries:

$$\text{Class I: } \tau(0.5) \leq 3.6 - \frac{14}{n} \text{ and } pk \leq 3.20$$

$$\text{Class II: } 3.6 - \frac{14}{n} < \tau(0.5) \leq 5.8 - \frac{35}{n}$$

$$\text{Class III: } 5.8 - \frac{35}{n} < \tau(0.5)$$

$$\text{Class IV: } \tau(0.5) \leq 3.6 - \frac{14}{n} \text{ and } pk > 3.20$$

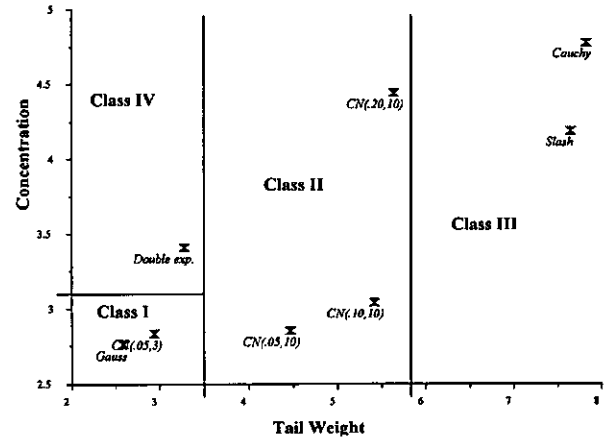


Figure 2. Four Classes of Distributions

The final stage consists in setting the adjustment of the two estimators in each class. Since we are interested only in the symmetrical case, the b parameter of the Cauchy function is nil. By simulations, we determined for the eight reference distributions the optimal constants c of the Tukey and Cauchy functions (*i.e.*, minimizing the variance of these estimators or, what amounts to the same thing here, their mean square deviation). These do indeed diminish with tail weight, except of course for the case of the double exponential distribution, which requires an adjustment similar to those used for the Slash and Cauchy distributions.

Tukey's estimator is more efficient on the normal or contaminated distributions, but it generally requires finer adjustment. Figure 3 shows the example of the contaminated distribution $CN(.10, 10)$. Lastly, while the choice of the constant appears to be relatively critical for the heavy-tailed or concentrated distributions, a wide band of value is possible for distributions close to the normal distribution.

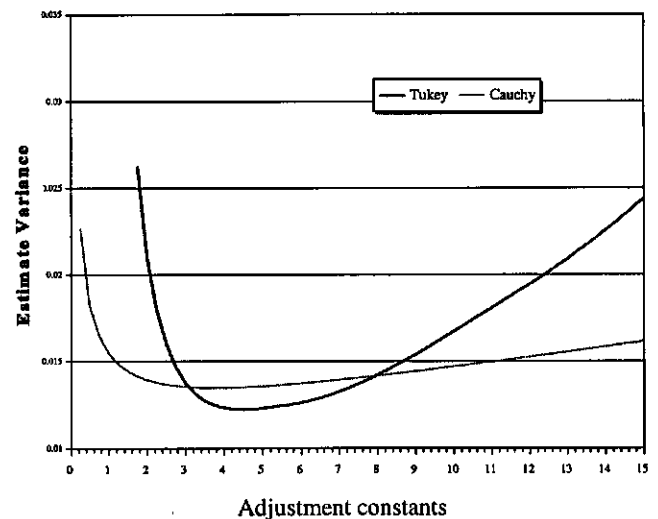


Figure 3. Variance of Tukey and Cauchy Estimators for the Distribution $CN(.10, 10)$ ($n=100$)

The synthesis of these results serves to define the adjustments to be used on each distribution class. These adjustments, established for samples of size 100 (Table 2), remain entirely acceptable for samples sizes between 50 and 150.

Table 2
Adjustment of Estimators by Class of Distribution
of Residuals ($n = 100$)

Class	Tukey	Cauchy
I	7	7
II	4.5	4
III	3	1
IV	3	1

5. APPLICATION TO THE INVESTMENT SURVEY

5.1 The Problem of Stratification

The strata used for the LI estimator are defined by the cross-tabulation of an activity (18 manufacturing sectors) and a company size class (small, medium or large). Among these 54 strata, approximately 20 never contain more than 20 observations. This stratification is therefore too fine for the adaptive procedure to be used correctly, as it assumes a minimum number of observations.

Since small firms are fairly distinct from medium-sized and large firms in terms of dispersion and residuals tail weight, differentiation by size is maintained. Sectors must thus be grouped. We decided not to adopt the method used by Sohre (1995), which consists of grouping after data collection those sectors having the closest parameters (here the average change in investment). Proximity is impossible to assess in small strata, and the groups obtained are likely to change from one survey to another, making comparisons difficult. We preferred to redefine 15 new strata based on a higher classification level distinguishing only four sectors: intermediate goods, professional capital goods, automobile, and consumer goods.

5.2 Characteristics of Strata

The hypothesis of a variance of residuals independent of x in the model ξ cannot be accepted. The choice of γ in the function η is made in such a way that the curve of the residuals (in absolute value) as a function of the regressor, smoothed by the LOESS method, shows no trend (Cleveland 1979). For the stratum representing intermediate goods and medium-sized companies in the April 1995 survey (see Figure 4), $\gamma = 1.3$ is an acceptable compromise between the appearance of a downward trend for small values of x and the cancellation of the upward trend for the larger values of x . A similar examination on the other strata confirmed this choice for the manufacturing industry as a whole.

In each stratum, the distribution of the residuals systematically exhibits a heavier tail than the normal distribution, without being extremely heavy-tailed. Within a given sector, the tail weight indicator decreases with company size. The great majority of the strata representing small and medium-sized firms were assigned to Class 2. Large firms more often exhibit somewhat heavy-tailed distributions, close either to the normal distribution (Class 1), or the double exponential distribution (Class 4). Class 2 is by far the largest and represents 75% of cases. Only 20% of the distributions are recognized as somewhat heavy-tailed and are assigned in equal proportions to classes 1 and 4. On the other hand, very heavy-tailed distributions (Class 3) are unusual (less than 5% of the cases). While there appears to be a certain persistence to the classification, it is not perfect. And the changes are quite real, since they resist a slight modification of the boundaries between classes. Thus this perfectly justifies the use of an adaptive procedure.

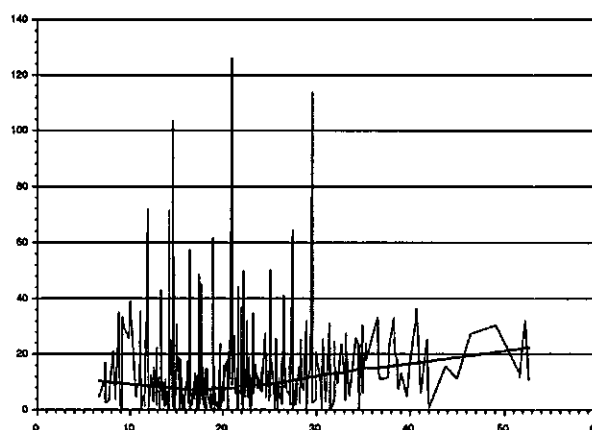


Figure 4. Absolute Value of Residuals ($\gamma = 1.3$, Intermediate Goods, Size 2, April 95)

5.3 Resulting Estimates

The estimation procedure based on (5), applied to the six surveys covering the period 1990-1995, yielded the results shown in Figure 5. Also shown are National Accounts estimates, those obtained with the LI estimator, and those from the Annual Business Survey (ABS), which is exhaustive.

For the manufacturing sector as a whole, the results of the adaptive procedure are comparable to those obtained with the LI estimator. The biquadratic function results in estimates that are consistently lower than those obtained with the Cauchy function. With a finite rejection point, the Tukey function is less influenced by the slight asymmetry toward the right in the distribution of the residuals. These new estimates are closer to those of the ABS than to the National Accounts estimates. This is hardly surprising, considering the excellent correlation between individual

ABS data and the responses obtained in the survey. As yet there is no explanation for the differences in 1991 and 1994 in relation to the National Accounts estimates. Apart from the year 1994, the estimates obtained with the Cauchy function are entirely acceptable in the intermediate goods and automobile sectors and to a lesser extent in the professional capital goods sector. On the other hand, in consumer goods, the results are fairly far from the National Accounts estimates. Here we are likely running up against a problem of sample quality. This sector is quite heterogeneous, and a few activities such as printing are poorly covered by the survey.

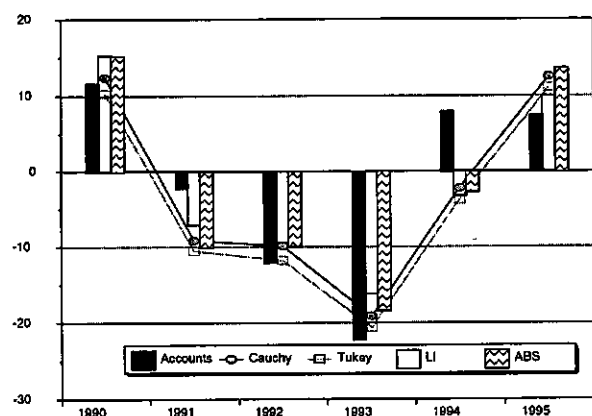


Figure 5. Investment Growth Rate in Value in the Manufacturing Industry

6. CONCLUSIONS

This article presents a theoretical justification of a procedure currently used to process data from the Investment Survey; in particular it offers a justification of the principle of excluding outliers or large investors. However, the strategy of reweighting the linear estimator following Hidirolou and Srinath (1981) shows itself to be insufficient for this purpose in several respects, mainly having to do with the identification and treatment of representative outliers. The dichotomy between extrapolatable individuals and large investors appears too radical and leads to a lack of robustness, since the influence curve of this estimator is not continuous.

On the other hand, the hypothesis of a linear superpopulation model and its estimation by GM-estimators seemed to us to be of great interest from both a methodological and practical standpoint. The insertion of these techniques into an adaptive procedure also makes it possible to have a robust estimator for a variety of situations. Following principles described in the literature, the procedure proposed here uses indicators of tail weight and concentration of the residuals in the linear model calculated from the sample, to decide on the adjustment of the weight function to be used, it being assumed that the residuals are

symmetrical. The estimates made with the Cauchy function yielded satisfactory results on the manufacturing industry, and they largely validate previously published results. The advantages of this method over the one currently used basically have to do with lower implementation costs and greater control over the methodology employed.

The adaptive procedure was constructed independently of the survey, and therefore there is no guarantee that the classification is optimal for the strata content. Furthermore, we did not study the robustness of the rule for assigning values to a class. This issue is important when one carries out several successive measurements and one wants to interpret the revisions. Clearly, further research on these classification methods is required, in order to integrate additional information such as the information yielded by earlier estimates or comprehensive surveys of the population studied.

ACKNOWLEDGEMENTS

The author wishes to thank Michel Hidirolou and Dominique Ladiray for their comments and suggestions during the preparation of this article.

REFERENCES

- BREWER, K.R. (1963). Ratio estimation and finite population: some results deducible from the assumption of an underlying stochastic process. *The Australian Journal of Statistics*, 5, 93-105.
- 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). Outlier robust sample survey inference. *Bulletin of the International Statistical Institute, Proceedings of the 49th Session, Book 2*, 55-72.
- CLEVELAND, W.S. (1979). Robust locally weighted regression and smoothing scatterplots. *Journal of the American Statistical Association*, 74, 829-836.
- 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., ROUSSEUW, P.J., and STAHEL, W.E. (1986). *Robust Statistics: The Approach Based on Influence Function*. New York: John 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.
- HOAGLIN, D.C., MOSTELLER, F., and TUKEY, J.W. (1983). *Understanding Robust and Exploratory Data Analysis*. New York: John Wiley.
- HOGG, R.V. (1974). Adaptive robust procedures: a partial review and some suggestions for future applications and theory. *Journal of American Statistical Association*, 69, 909-923.

- HOGG, R.V. (1982). On adaptive statistical inferences. *Communication in Statistics*, 11, 2531-2542.
- HOGG, R.V., BRIL, G.K., HAN, S.M., and YUL, L. (1988). An argument for adaptive robust estimation. *Probability and Statistics Essays in Honor of Franklin A. Graybill*. Amsterdam: North-Holland/Elsevier, 135-148.
- HUBER, P.J. (1981). *Robust Statistics*. New York: John Wiley.
- LEE, H. (1995). Outliers in business surveys. In *Business Survey Methods*. New York: John Wiley.
- MOBERG, T.F., RAMBERG, J.S., and RANGLES, R.H. (1980). An adaptive multiple regression procedure based on M-estimators. *Technometrics*, 22, 213-224.
- RAVALET, P. (1996). L'estimation du taux d'évolution de l'investissement dans l'enquête de conjoncture: analyse et voie d'amélioration. Document de travail de l'INSEE Méthodologie Statistique, 9604.
- RIVEST, L.P. (1989). De l'unicité des estimateurs robustes en régression lorsque le paramètre d'échelle et le paramètre de régression sont estimés simultanément. *Canadian Journal of Statistics*, 17, 141-153.
- ROYALL, R.M. (1970). On finite population sampling theory under certain linear regression models. *Biometrika*, 57, 377-387.
- SOHRE, P. (1995). The Adaptive KOF Procedure for the Estimation of Industry Investment. 22nd CIRET Conference, Singapore.
- WELSH, A.H., and RONCHETTI, E. (1994). Bias-Calibrated Estimations of Totals and Quantiles From Sample Surveys Containing Outliers. Technical Report, Dept. of Econometrics, University of Geneva, Switzerland.

Sampling and Maintenance of a Stratified Panel of Fixed Size

F. COTTON and C. HESSE¹

ABSTRACT

Statistical agencies often constitute their business panels by Poisson sampling, or by stratified sampling of fixed size and uniform probabilities in each stratum. This sampling corresponds to algorithms which use permanent numbers following a uniform distribution. Since the characteristics of the units change over time, it is necessary to periodically conduct resamplings while endeavouring to conserve the maximum number of units. The solution by Poisson sampling is the simplest and provides the maximum theoretical coverage, but with the disadvantage of a random sample size. On the other hand, in the case of stratified sampling of fixed size, the changes in strata cause difficulties precisely because of these fixed size constraints. An initial difficulty is that the finer the stratification, the more the coverage is decreased. Indeed, this is likely to occur if births constitute separate strata. We show how this effect can be corrected by rendering the numbers equidistant before resampling. The disadvantage, a fairly minor one, is that in each stratum the sampling is no longer a simple random sampling, which makes the estimation of the variance less rigorous. Another difficulty is reconciling the resampling with an eventual rotation of the units in the sample. We present a type of algorithm which extends after resampling the rotation before resampling. It is based on transformations of the random numbers used for the sampling, so as to return to resampling without rotation. These transformations are particularly simple when they involve equidistant numbers, but can also be carried out with the numbers following a uniform distribution.

KEY WORDS: Panel; Stratified sampling of fixed size; Stratified simple random sampling; Maximum coverage; Sample rotation; Equidistant numbers.

1. INTRODUCTION

We consider the successive selection of samples intended to follow the change over time of sums of variables, more generally functions of sums, in a population. For example, this may be a population of businesses or establishments for which we wish to follow monthly sales trends. The ideal would be to be able to conserve a constant sample, but demographic movements make this impossible and it may not be desirable in light of the survey response burden.

The methods for selecting units presented in this article are subject to the following three constraints:

Firstly, it is necessary to regularly introduce births and to take deaths into account.

Secondly, sampling involves characteristics of units which change over time, such as the size or primary activity of businesses. These characteristics can be used to modulate the probabilities of inclusion. Notably, it is often prudent to increase these probabilities with the size of the units if we estimate sums of variables correlated with this size. In addition, these characteristics may eventually be used as stratification criteria. In this article, a stratum will mean a subset of the population within which the sampling is of fixed size, to the nearest rounded digit. However, the criteria used in the stratification of the first sampling, such as the primary activity of the unit, become "inexact" or become less and less correlated with the variables of interest such as size. This results in a progressive increase

in the variance of the estimates. To remedy this, it is appropriate to carry out a resampling of the sample from time to time after updating the stratification and calculating new probabilities of inclusion. This must be done while endeavouring to conserve the maximum number of units. However, fatally, units will be excluded and others will be introduced, mainly because of changes in the probabilities of inclusion, although this would also happen because of the changes of strata, even if the probabilities of inclusion remained constant.

Thirdly, we would like to distribute our survey response burden over a larger number of units. We determined a maximum duration limit for inclusion in the panel. Beyond this limit, the unit is replaced by another unit chosen from those which have never been included, or which have been absent the longest. We call this change of the sample over time rotation. It is generally slow and regular. The various methods for performing this rotation are well known in statistical agencies. They consist mainly in attributing, at the beginning, a permanent random number to each unit of the population. The successive samples are defined by intervals over these numbers or by the ranks induced by these numbers.

We call the chronological sequence of samples resulting from these updating operations a "panel" and the set of updating operations "maintenance" of the panel.

The maintenance scheme presented in this article is analogous to that of Hidioglou, Choudhry and Lavallée (1991). It corresponds to a frequency of updating of the

¹ F. Cotton, Institut National de la Statistique et des Études Économiques, Département de l'Informatique and C. Hesse, Institut National de la Statistique et des Études Économiques, Département "Système Statistique d'Entreprises", 18 boulevard Adolphe-Pinard, 75675, Paris Cedex 14.

stratification and probabilities which is significantly less than the survey frequency. This is generally the case for surveys with an infra-annual periodicity. The speed of demographic movements is not considered large enough to make it worthwhile to reselect the sample every time. The rotation is carried out without changing the probabilities of inclusion and the strata between two resamplings and it is regularly spread over time to conserve a certain continuity of the quality of the estimators of change over time. This also corresponds to a duration of inclusion of which the expected value is constant. In certain algorithms, we could determine a constant duration between two resamplings; otherwise we could set an upper limit. The speed of rotation represents a compromise between the efficiency of the estimators of change over time, which is greater the lower the rate of renewal, and the concern not to keep a unit in the panel for too long. Note that the quest for maximum coverage in the resampling remains meaningful with the rotation: we first remove the fraction to be renewed as if there were no resampling, then we seek the maximum coverage with the residual portion.

We will examine several methods of panel maintenance, with emphasis on maximizing sample coverage during resamplings. We will distinguish more particularly a process which assigns equidistant numbers to the units before each change of stratum.

The article is divided as follows:

After reviewing definitions and describing a few notations in section 2, we briefly indicate in section 3 how Poisson sampling makes it possible to carry out the previous maintenance scheme simply and perfectly. This sampling has the disadvantage of being of random size, but it serves as a reference for the stratified sampling of fixed size which we then consider.

In most instances, in these samplings, we determined probabilities of inclusion at the outset and used a rounded number to determine an entire sample size in each stratum. This problem, examined in section 4, is not negligible when the strata are small, which can occur for strata of births. In addition, rounding is used in the method which we propose to maximize the coverage after resampling.

Section 5 deals with the maximum coverage of samples of fixed size. First, we review two known methods: that of Kish and Scott (1971) and another based on the attribution to each unit of permanent independent numbers following the uniform distribution. The Kish and Scott method (1971) seems poorly suited to an intermediate rotation between resamplings. The other method, which reproduces simple random sampling in each stratum, does not have this disadvantage, but the coverage is less than with the Kish and Scott method (1971). Finally, we propose that the numbers be equidistant before resampling. We then obtain the same coverage as with the Kish and Scott method (1971), at least in the case of proportional distribution, while facilitating intermediate rotations. However, the coverage remains less than the maximum theoretical coverage which we obtain, for example, with Poisson sampling.

In sections 6 and 7, we present the intermediate phases of updating births and deaths and of rotation.

To conclude the topic of maintenance, we show in section 8 how resampling can take place between two phases of rotation. We present a type of algorithm which extends after resampling the rotation before resampling. It is based on transformations of the random numbers used in the sampling, so as to return to resampling without rotation. These transformations are particularly simple when they involve equidistant numbers, but can also be carried out with the uniform beginning numbers if we wish to continue with simple random sampling.

2. REMINDERS, DEFINITIONS AND NOTATIONS

Let there be a population, or finite set of units $i \in U = \{1, \dots, N\}$ where N is the size of the population.

We consider only samples without replacement. A sample is then simply a subset s of U . We call sample size the number n of units which it contains.

A sampling or selection plan is a discrete probability $p(s)$ over the set of samples.

We can generalize to joint sampling of several samples. By limiting ourselves to two samples s_1, s_2 , the joint sampling is the probability $p(s_1, s_2)$ over the set of pairs (s_1, s_2) .

The first-order probability of inclusion of an individual i is defined by:

$$\pi_i = \sum_{s \ni i} p(s).$$

$E(.)$ being the expected value with respect to the sampling, this yields:

$$E(n) = \sum_{i \in U} \pi_i.$$

In the case of two samples with first-order probabilities of inclusion $\pi_{i,1}, \pi_{i,2}$, we can define the joint probability of inclusion:

$$\pi_{i,1,2} = \sum_{s_1 \ni i, s_2 \ni i} p(s_1, s_2).$$

This yields the constraint:

$$\pi_{i,1,2} \leq \min(\pi_{i,1}, \pi_{i,2}). \quad (2.1)$$

If $i \in s_1$, the probability of reselection in s_2 is $\pi_{i,2}/\pi_{i,1} \leq \min(1, \pi_{i,2}/\pi_{i,1})$.

In Poisson sampling, the selection of the units is independent and the sample size is random. Except in section 3, we will instead consider sampling where the size is fixed to the nearest rounded digit.

Simple random sampling (SRS) is sampling of fixed size where the samples are equiprobable. This yields $\pi_i = n/N$.

The population is partitioned into strata $U_h, h = 1, \dots, H$ of sizes N_h . In this article, we will call a set of H independent samples of fixed size n_h in each stratum

“stratified sampling of fixed size” and we will limit ourselves to samplings with a uniform first-order probability of inclusion in each stratum. We will then use the notation $f_h = \pi_i$. We will call a stratified sampling of fixed size with simple random sampling in each stratum “stratified simple random sampling” (SSRS).

We will call the number of consecutive surveys where a unit is included in the panel “duration of inclusion of a unit.” We will notate it D_i , or D_h in the particular case where it is the same for all units of a stratum h . When $\pi_i \geq 0.5$, this duration cannot be less than $\pi_i/(1 - \pi_i)$. For example, if $\pi_i = 0.7$, the duration of inclusion is at least 3. In practice, we will not rotate units whose π_i exceeds a certain threshold.

In addition, the previous variables are indexed by survey wave t . The population U_t of size N_t and the sample s_t of size n_t vary because of births and deaths, and the sample also varies as a result of the stipulated rotation. Moreover, we will consider samples at particular times $t = t_1$ of the first sampling and $t = t_2$ of the first resampling. For the sake of simplicity, they will be notated s_1, s_2 instead of s_{t_1}, s_{t_2} . The algorithms described for the pair (s_1, s_2) will be valid for the following resampling pairs.

3. SOLUTION BY POISSON SAMPLING

It is enlightening to examine how we can observe the panel maintenance scheme by Poisson sampling. This is the model which we will endeavour to approximate in order to choose a selection method.

We attribute to each unit i , at its birth, a number which is a random number ω_i selected according to the uniform distribution in $[0, 1)$. It is implicit in the formulae where these numbers appear that the results of the operations are *modulo 1*.

During the first sampling, at date $t = t_1$, we select the units such that ω_i belongs to the interval $[0, \pi_{i,1})$ where $\pi_{i,1}$ are the probabilities of inclusion given. In the absence of rotation, we keep this interval at the following dates until resampling. Births as well as deaths are distributed at random in this interval. The resampling, at date $t = t_2$ is carried out by selecting the units of the interval $[0, \pi_{i,2})$ where $\pi_{i,2}$ are new probabilities of inclusion. The joint probability of inclusion is equal to the length of the common interval, *i.e.*, $\min(\pi_{i,1}, \pi_{i,2})$ which is the maximum theoretically possible according to the formula (2.1). The expected value of the coverage is therefore itself maximal.

Let us now consider a rotation between the sampling and the resampling. We maintain the probability $\pi_{i,1}$ and we can determine a duration of inclusion $D_{i,1}$, which is variable depending on the units, but fixed until the resampling. This constraint is realized by defining the sample at date $t (t_1 < t < t_2)$ by the interval

$$\omega_i \in [(t - t_1)\pi_{i,1}/D_{i,1}, (t - t_1)\pi_{i,1}/D_{i,1} + \pi_{i,1}).$$

The rate of rotation is a random variable. Its expected value results from $D_{i,1}$. It is equal, for any subset V of the population, to $\sum_{i \in V} (\pi_{i,1}/D_{i,1}) / \sum_{i \in V} \pi_{i,1}$.

At the first resampling at date $t = t_2$, we could define the sample by

$$\omega_i \in [(t_2 - t_1)\pi_{i,1}/D_{i,1}, (t_2 - t_1)\pi_{i,1}/D_{i,1} + \pi_{i,2}).$$

However, we encounter a difficulty for units such that

$$\pi_{i,2} < \pi_{i,1} \left(1 - \frac{1}{D_{i,1}} \right),$$

and if ω_i belongs to the interval

$$\left[(t_2 - t_1)\pi_{i,1}/D_{i,1} + \pi_{i,2}, (t_2 - t_1)\pi_{i,1}/D_{i,1} + \pi_{i,1} \left(1 - \frac{1}{D_{i,1}} \right) \right).$$

These units, which were previously in the sample, are excluded but will be reincluded in a future rotation. If we wish to avoid this, we must make the limit of the new interval coincide with that of the old interval, and the sample at date $t = t_2$ is finally defined by:

$$\omega_i \in [a_{i,1}, a_{i,1} + \pi_{i,2}),$$

where:

$$a_{i,1} = (t_2 - t_1)\pi_{i,1}/D_{i,1} + \max \left[0, \pi_{i,1} \left(1 - \frac{1}{D_{i,1}} \right) - \pi_{i,2} \right].$$

The joint probability of inclusion is equal to the length of the common interval, *i.e.*,

$$\min \left(\pi_{i,1} \left(1 - \frac{1}{D_{i,1}} \right), \pi_{i,2} \right).$$

This is also the maximum compatible with the rotation.

If we continue the rotation with durations of inclusion $D_{i,2}$ the interval at date $t > t_2$ is:

$$[a_{i,1} + (t - t_2)\pi_{i,2}/D_{i,2}, a_{i,1} + (t - t_2)\pi_{i,2}/D_{i,2} + \pi_{i,2}).$$

Poisson sampling controls exactly the duration of inclusion and maximizes, as an expected value, the coverage during resampling but with the disadvantage of a random sample size, regardless of the subpopulation. In the following pages, we will endeavour to devise algorithms similar to those just described for Poisson sampling in order to apply them to stratified sampling of fixed size. We will try to control the duration of inclusion in the rotation, as for Poisson sampling, and to approximate the same rate of coverage during resampling. We will begin with the problem of coverage during resampling in section 5, but first, it is useful to clarify certain concepts concerning the rounding of sample sizes by stratum.

4. ROUNDING OF SAMPLE SIZES BY STRATUM

This problem is related to the estimation formulae. These formulae use the first-order probabilities of inclusion, either in the unbiased Horvitz-Thompson estimator or in adjusted estimators. Let f_h be the probability of inclusion by stratum, and let $v_h = N_h f_h$. We must have a whole number n_h per stratum. An initial method for accomplishing this consists in restricting the choice of the f_h in such a way that v_h is an integer. In each stratum where we would have had $v_h < 1$, we must take $v_h = 1$ so that $f_h > 0$. However, if the stratification is very fine vis-à-vis the sample size, this occurs in numerous strata. This makes it necessary either to increase the sample size or to decrease the sampling rate in the other strata, to the detriment of efficiency.

We will use a second method, which consists in linking the probability f_h more loosely to n_h . We apply a rounding process such that $E(n_h) = v_h$, where v_h is no longer necessarily an integer.

Let us assume that $I(\cdot)$ is the integer part function. We must have

$$\Pr[n_h = I(v_h) + 1] = \varphi_h,$$

$$\Pr[n_h = I(v_h)] = 1 - \varphi_h,$$

where $\varphi_h = v_h - I(v_h)$.

It is then no longer necessary that $n_h > 0$ in order for $f_h > 0$. Note that the first method can be considered a particular case of the second. This rounding can be done independently by stratum, in a linked way by systematic rounding or by the Cox method (1987). We describe only systematic rounding.

Let us first order all of the strata, and index them by their rank. Let $c_0 = 0$ and $c_h = \sum_{j=1}^h \varphi_j$; we select a number θ in the interval $[0, 1)$, according to the uniform distribution and we take $n_h = I(v_h) + 1$ in the strata such that $c_{h-1} \leq m - 1 + \theta < c_h$ for m entirely.

This implies that

$$|(n_{j_1} + \dots + n_{j_2}) - (v_{j_1} + \dots + v_{j_2})| < 1,$$

for any j_1, j_2 such as $1 \leq j_1 \leq j_2 \leq H$.

In particular, the global size differs by less than one unit from its expected value. This is obviously not the case with independent roundings.

5. ALGORITHMS FOR THE MAXIMUM COVERAGE OF SAMPLES OF FIXED SIZE

The maintenance algorithms which we propose are based on the attribution of equidistant numbers. This is not necessary during the first sampling, nor in the rotation, but is used to maximize the coverage during updates of the

stratification. That is why we examine this maintenance phase first.

Let us begin by describing all the notations and making a few useful observations.

We select a first sample s_1 stratified according to criterion h_1 . After a certain time has elapsed, we select a new sample s_2 with an updated stratification h_2 . The first-order probabilities of inclusion are respectively f_{h_1}, f_{h_2} and the sample sizes required by stratum are respectively n_{h_1}, n_{h_2} . It is sufficient to consider what happens in any stratum $h_2 = g$. Let $s_{g,1}$ be the part of the first sample s_1 in this new stratum, of which the size $n_{g,1}$ is generally random. Let $s_{g,2}$ be the part of the second sample s_2 in this new stratum, of which the size is fixed to the nearest rounded digit. The size $n_{g,1,2}$ of the coverage cannot exceed the limit $n_{g,1,2}^+ = \min(n_{g,1}, n_{g,2})$. We can hope to devise $s_{g,2}$ a resampling process with a uniform first-order probability of inclusion in $s_{g,1}$ which makes it possible to attain this limit, at least when the first-order probabilities of inclusion in $s_{g,1}$ are also equal to a single value $f_{h_1} = f_1$. Note that, even if this limit is attained, the fixed size constraints decrease the coverage. The finer the stratification, the greater this effect. In fact, the smaller the population of stratum g , the greater the likelihood that the coefficient of variation of $n_{g,1}$ will be large, as well as the proportion of units not reselected in the case $n_{g,1} > n_{g,2}$.

There is an obvious way of attaining the limit $n_{g,1,2}^+$. Let us assume first of all that the first-order probabilities of inclusion in $s_{g,1}$ are uniform. If $n_{g,1} < n_{g,2}$, we add $n_{g,2} - n_{g,1}$ units to $s_{g,1}$ selected at random in the complement of $s_{g,1}$. If $n_{g,1} > n_{g,2}$, we remove $n_{g,2} - n_{g,1}$ units from $s_{g,1}$ selected at random. By construction this yields $s_{g,2} \subseteq s_{g,1}$ or $s_{g,2} \supseteq s_{g,1}$, and $n_{g,1,2} = n_{g,1,2}^+$. If the first-order probabilities of inclusion in $s_{g,1}$ are not uniform, we apply the same method within subsets where these probabilities are uniform. This is the method proposed by Kish and Scott (1971) on page 468 of their article. They do not stipulate the procedure for random selection, but we assume that it is SRS.

As Kish and Scott point out, the second-order probabilities of inclusion are not uniform and if the first sampling is a SSRS, the second sampling no longer meets this definition. The first-order probability of inclusion, itself, is not strictly uniform when it includes elements of strata from the previous sampling: see an example in the appendix. However, there is another method which verifies this condition. It is well known to statistical agencies which practise coordination of samples. For the sake of convenience, we will call it "method 1".

Method 1:

Use of independent numbers following the uniform distribution

We attribute to the units, at their birth, ω_i numbers which follow the uniform distribution in $[0, 1)$ and are independent, as in Poisson sampling. The first sample s_1 is obtained by selecting, for example, the n_{h_1} units of lower rank according to ω_i in each stratum. With this algorithm, the maximum coverage is also obtained by selecting the n_{h_2}

units of lower rank according to ω_i in each stratum h_2 . Moreover, it is obvious that these two samplings are SSRS.

It is also obvious that we cannot obtain greater coverage with this algorithm. In addition, we conjecture that it is not possible to do better, for SSRS, regardless of the algorithm.

On the other hand, the coverage is poorer as an expected value than with the Kish and Scott method (1971), at least in the particular case where the first-order probabilities of inclusion in s_1 are uniform. In fact, at that point the relations $g s_{g,2} \subseteq s_{g,1}$ or $s_{g,2} \supseteq s_{g,1}$, $n_{g,1,2} = n_{g,1,2}^+$, are not necessarily true and the loss of coverage is greater, the smaller the strata during the first sampling.

We shall demonstrate this, again in the particular case of a uniform probability of inclusion f_1 in s_1 . Let us assume that ω_{h_1} is the greatest value of ω_i for the units of s_1 in stratum h_1 , and ω_g the greatest value of ω_i for the units of s_2 in stratum g . Let $\omega_1^- = \min(\omega_{h_1})$ and $\omega_1^+ = \max(\omega_{h_1})$. If $\omega_g \leq \omega_1^-$ then $s_{g,2} \subseteq s_{g,1}$ and if $\omega_g \geq \omega_1^+$, then $s_{g,2} \supseteq s_{g,1}$. In both cases $n_{g,1,2} = n_{g,1,2}^+$. The risk of not attaining the limit exists only if $\omega_1^- \leq \omega_g \leq \omega_1^+$. In this case, the relation $s_{g,2} \subseteq s_{g,1}$ or $s_{g,2} \supseteq s_{g,1}$ is no longer necessarily true: see Figure 1, where we considered only 2 strata h_1 . The loss of coverage is greater where the quantity $\omega_1^+ - \omega_1^-$ is greater as an expected value, and therefore where the strata h_1 are smaller.

Method 2:

Use of equidistant numbers

If we accept not to conserve a SSRS, how can we modify the previous method to obtain the same coverage as the Kish and Scott method (1971), at least when we have the uniform probability of inclusion f_1 in s_1 ? We have seen that the loss of coverage was the result of the deviation between the ω_{h_1} . It is sufficient to transform the ω_i into new numbers $\rho_{i,1}$ in such a way that the ρ_{h_1} which correspond to the ω_{h_1} are as close as possible to a common value, i.e., f_1 . More specifically, we would like to have the equivalence:

$$\{i \in s_1 \rightarrow R_{h_1}(i) \in [1, \dots, n_{h_1}]\} \rightarrow \rho_{i,1} \in [0, f_{h_1}),$$

where $R_{h_1}(i)$ is the rank according to ω_i in h_1 of unit i . A solution is given by the transformation:

$$\rho_{i,1} = \frac{R_{h_1}(i) - 1 + \theta_{h_1}}{N_{h_1}} \quad (5.1)$$

where θ_{h_1} is a real number which verifies:

$$\begin{cases} \theta_{h_1} \in [0, \varphi_{h_1}), n_{h_1} = I(v_{h_1}) + 1, \\ \theta_{h_1} \in [\varphi_{h_1}, 1), n_{h_1} = I(v_{h_1}). \end{cases}$$

The transformation therefore involves the rounded number of the v_{h_1} examined in section 4. The sampling of s_2 is carried out like that of s_1 except that the $\rho_{i,1}$ now play the role of the ω_i : in each new stratum g we define rounded sizes $n_{g,2}$ and we select the $n_{g,2}$ units of lower rank

according to $\rho_{i,1}$. Note that these ranks are different from those induced by ω_i .

Let us assume that the probability of inclusion in s_1 is still uniform. Let ρ_g be the value of $\rho_{i,1}$ for the unit of rank $n_{g,2}$ in g . If $\rho_g \in [0, f_1)$, then $s_{g,2} \subseteq s_{g,1}$. Otherwise $s_{g,2} \supseteq s_{g,1}$. In this particular case, we therefore attain the maximum coverage $n_{g,1,2}^+$ as in the Kish and Scott method (1971), and unlike method 1. We illustrate in Figures 1 and 2 how the transformation into equidistant numbers makes it possible to increase the coverage compared to method 1.

We apply the same algorithm when the probabilities of inclusion in s_1 are not uniform. Unlike the Kish and Scott method (1971), we do not need to fix the size of the new sample within subsets where these probabilities are uniform. This is another advantage and we think that it increases the coverage.

Nonetheless, the coverage obtained by this algorithm remains lower, as an expected value, than that of a Poisson sampling with the same probabilities of inclusion. In order to have, as an expected value, the same coverage as with Poisson sampling, it would be sufficient to define $s_{g,2}$ by $\rho_{i,1} \in [0, f_g)$. In fact, we would then have $\Pr(i \in s_1 \cap s_2) = \min(f_{h_1}, f_g)$, but the sampling so obtained would no longer be of fixed size.

The following resamplings, after new updates, are carried out by repeating the process. For example, before selecting s_3 we calculate equidistant numbers $\rho_{i,2}$ based on $\rho_{i,1}$ (and not ω_i) in each stratum h_2 .

The resulting sampling plan in the new strata is no longer a SRS. In particular, the probabilities of inclusion of the pairs of units vary generally as a function of the former strata. In other words, the resampling keeps a "trace" of the stratification of the first sampling. Moreover, the probabilities of inclusion of the units in $s_{g,2}$ are not exactly equivalent to f_g , except for the sample defined by $\rho_{i,1} \in [0, f_g)$. For the sample of fixed size $n_{g,2}$ this probability varies as a function of the size of the former strata. As in the Kish and Scott method (1971), we do not strictly control these probabilities. However, the deviation between f_g and the true probability becomes negligible when $n_{g,2}$ is sufficiently large.

Note 1. The transformation of numbers which independently follow the uniform distribution in equidistant numbers was proposed by Brewer, Early and Hanif (1984) as a way of rotating samples in the same manner as Poisson sampling, with the advantage of a smaller variance of the sample size. However, this transformation is performed by taking the set of the population, and therefore they did not address the problem of maximum coverage during changes of stratum. The numbers change only when births and deaths are updated, according to a procedure which is also quite different from that which we propose for changes of stratum.

Note 2. In the demonstration we just provided, it is not necessary that the numbers be completely equidistant. It is sufficient that the n_{h_1} units of s_1 and the $N_{h_1} - n_{h_1}$ complementary units have their new numbers respectively in $[0, f_{h_1})$, $[f_{h_1}, 1)$. We could attribute these new numbers

in such a way that they independently follow the uniform distribution in these intervals.

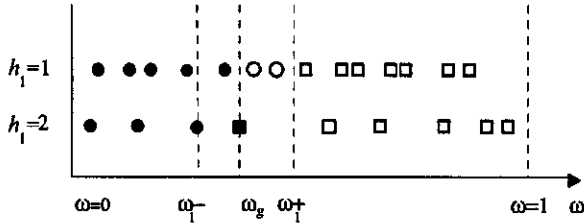


Figure 1. Coverage with method 1 (numbers following the uniform distribution).

We have represented the units in g according to the value of the number ω (on the abscissa) and the stratum h_1 of the first sampling (on the ordinate). We assume that there are only two strata. The circles correspond to $s_{g,1}$ and the squares to the complementary part. The solids correspond to $s_{g,2}$ and the blanks to the complementary part. The size of $s_{g,2}$ was fixed at 9 which defines ω_g . In this example, we see that two units are not reselected (in $h_1 = 1$) and that another is new (in $h_1 = 2$). The size of the coverage is 8, while the Kish and Scott method would make it possible to reselect the 9 units in $s_{g,1}$.

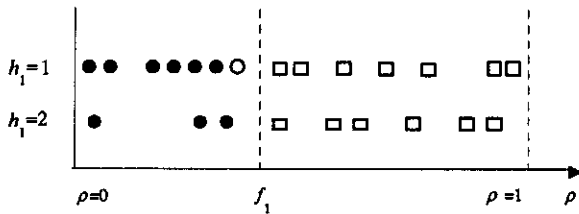


Figure 2. Coverage with method 2 (equidistant numbers).

We are in the same situation as in Figure (1), but this time the equidistant numbers ρ serve as the abscissa of the units. This equidistance is defined in each of the whole strata h_1 and the gaps we see in the sequence of numbers correspond to the units which are not in g . The first sample $s_{g,1}$ is composed of the units for which this number is less than the probability of inclusion f_1 , regardless of the stratum. The second sample $s_{g,2}$ is composed of the 9 units with the smallest ρ and the coverage is 9, as with the Kish and Scott method (1971).

6. UPDATING BIRTHS AND DEATHS WITHIN STRATA

In this section and the following one, we consider the stratification (h) without reference to the period. The updating of births and deaths within strata is essentially a particular case of change of the strata of units. It is exactly as if the births entered the strata and the deaths left. We can therefore apply the previous methods. Let us take a look, in particular, at method 2.

In a stratum, the population $U_{h,t}$ of size $N_{h,t}$ varies with each updating carried out at time t . We will notate the births as $B_{h,t+1}$ and the deaths $D_{h,t+1}$ between t and $t+1$, this yields $U_{h,t+1} = U_{h,t} + B_{h,t+1} - D_{h,t+1}$.

We consider the simple case where the probabilities of inclusion f_h remain uniform in $U_{h,t}$ and constant. The size $n_{h,t}$ of the sample $s_{h,t}$ is a rounded number to the integer of $N_{h,t}f_h$. The numbers $\rho_{i,t}$ change with each updating. Just before updating $s_{h,t}$, leading to $s_{h,t+1}$:

- a) we make equidistant the numbers $\rho_{i,t-1}$ in $U_{h,t}$.
- b) we attribute equidistant numbers to the units of $B_{h,t+1}$.

Let $\rho_{i,t}$ be the number so obtained. An initial solution would consist in selecting the $n_{h,t+1}$ units of $U_{h,t+1}$ with the smallest $\rho_{i,t}$. Note that these are no longer equidistant because we removed the deaths situated at random.

However, units with numbers close to f_h can leave the sample and then return on a future occasion. We remedy this by a rightward shift of the selection interval. Let $\rho_{h,d}$ be the number of the beginning unit of the selection interval for $s_{h,t}$ and $\rho_{h,e}$ that of the unit immediately following the end unit of this interval in $U_{h,t}$. In other words, the sample $s_{h,t}$ consists of the interval closed to the left and open to the right $[\rho_{h,d}, \rho_{h,e})$. Between t and $t+1$, the number of units of $U_{h,t+1}$ belonging to this interval becomes $m_{h,t+1}$. If $n_{h,t+1} \geq m_{h,t+1}$, the beginning of the interval for $s_{h,t+1}$ is fixed to the unit of number $\rho_{h,d}$, otherwise we shift the interval in such a way that its end is the unit of number $\rho_{h,e}$. We therefore have a slight involuntary rotation.

7. ROTATION BETWEEN TWO RESAMPLINGS

7.1 Rotation Without Updating of Births and Deaths

We can then stipulate a time of inclusion D_h whole and constant in the stratum. We have two variants, depending on whether we keep the same rounded number or vary it.

7.1.1 Fixed Rounded Number

We therefore have a size n_h strictly fixed during the rotation. We divide n_h into D_h whole numbers $n_{h,l}$ ($l = 1, \dots, D_h$) such that $|n_{h,l} - n_h/D_h| < 1$. Let q_h be the quotient and r_h the remainder of the division of $t - t_1$ by D_h and let $n_{h,0} = 0$. The sample at time t includes the units ranging from rank $1 + q_h n_h + \sum_{l=0}^{r_h-1} n_{h,l}$ to rank $(q_h + 1)n_h + \sum_{l=r_h}^{D_h-1} n_{h,l}$. If $D_h = D$, we can stipulate in addition

$$|\sum_{h=1}^H n_{h,l} - \frac{n}{D}| < 1, l = 1, \dots, D_h.$$

The variance of the rate of rotation is then practically nil.

However, the duration of inclusion is not controlled when $v_h < 1$: this yields $n_h = 0$ or $n_h = 1$. In the first case, there is no rotation, and in the second case, on the contrary, the time of exclusion can be considered too short. The following method makes it possible to obtain a rotation which corresponds to v_h .

7.1.2 Variable Rounded Number

The sample $s_{h,t}$ is defined based on the numbers rendered equidistant:

$$i \in s_{h,t} \Leftrightarrow \rho_{i,1} \in \left[f_h \frac{t-t_1}{D_h}, f_h \frac{t-t_1}{D_h} + f_h \right).$$

The sample size varies between $I(v_h)$ and $I(v_h) + 1$ in the stratum, and it is independent of the sizes in the other strata. This shows us what the result would be of the sample rotation advocated by Brewer *et al.* (1984) in the case of stratified sampling of fixed sized and uniform probability in each stratum.

7.2 Rotation With Updating of Births and Deaths

To simplify, we assume that each new survey wave is accompanied by the introduction of the births since the previous wave and a rotation. The method bifurcates into two procedures depending on whether or not we wish to respect exactly the durations of inclusion D_h between two resamplings.

7.2.1 Procedure A

The births are isolated in separate strata, and we wait for the resampling before subtracting the deaths. In this case each wave of births is dealt with exactly like an initial sampling after attributing the numbers ω_i . The sampling is carried out by stratifying with the same nomenclature (h), or with another more scattered or more confined. To simplify the notations, but without loss of generality, we assume that this is the same nomenclature. The index of stratification can then be written (b, h), where b crossed with h indicates the wave of births with a particular modality $b = 1$ corresponding to the units already existing during the first sampling or a previous resampling. This brings us back to the case of section 7.1 in each stratum (b, h) and the duration of inclusion is respected exactly.

The number of strata, and therefore of rounded numbers, is multiplied by the number of waves of births. The sample size can become fairly random with independent roundings (but less so than with Poisson sampling). It may therefore be worthwhile to link, at least partially, the rounded numbers. For example, we carry out a systematic rounding in the dimension h for each b or the reverse. We then keep these roundings and this is the 7.1.1 method which then applies rather than the 7.1.2 method.

7.2.2 Procedure B

In procedure B, we subtract the deaths at each survey wave. This is the type of updating presented in section 6. We would prefer a fixed duration of inclusion, but that is made difficult by the random number of deaths. At most, we can try to control a maximum duration of inclusion DM_h . We may also wish to prevent the units which have just left the sample from returning on a future occasion, which can occur if the rotation is slow. The idea is to get back to the algorithm described in section 6 by removing first of all from $s_{h,t}$ the units of which the previous duration

of inclusion in $s_{h,t}$ attained DM_h . They are found the farthest to the left of the interval $[\rho_{h,d}, \rho_{h,e})$ and are mixed with the births too recent to have attained DM_h . However, these must still be removed in order for the distribution of the sample according to the generations to be correct. For that, it is sufficient to attribute to the births a fictitious previous duration of inclusion which falls between 1 and DM_h , just after defining the sample. For example, after defining $s_{h,t}$, we assign to each unit of $B_{h,t}$ belonging to the sample the same previous duration of inclusion in the sample as that of the unit of $U_{h,t-1}$ situated immediately to the left. Then let $R_{h,d}$ be the highest rank among the ranks according to $\rho_{i,t}$ of the units of the interval associated with $s_{h,t}$ which have been included DM_h times in the sample; we discard the first units of $s_{h,t}$ up to and including rank $R_{h,d}$. Finally, this brings us back to the algorithm described in section 6 with, for $\rho_{h,d}$, the number of the unit of rank $R_{h,t} + 1, \rho_{h,e}$, remaining that of the unit which follows the unit of last rank in $s_{h,t}$.

8. RESAMPLING AFTER ROTATION

We now reselect the indices of strata h_1, h_2 . We define the stratification h_1 as a function of the procedure used for the updates of the births. With procedure A, we place the births in separate strata, this is the stratification defined by crossing the waves of births b with the nomenclature h_1 . With procedure B, h_1 is identical to h_1 . However, we keep the notations of the independent quantities of b as f_{h_1}, D_{h_1} .

The selection of the new sample s_2 , in a new stratification h_2 must be carried out at period $t = t_2$.

We begin by removing from the previous sample (at period $t = t_2 - 1$) the units which have attained the maximum authorized duration of inclusion. There remains a sample s'_1 of size n'_1 , of which we would like to conserve the maximum number of units in the resampling.

In the case without rotation examined in section 5, it was easy to define the resampling because the sample s_1 was composed of the units of lower rank according to ω_i in each stratum after a real number independent of the ω_i . In this instance, this number is 0. The resampling took place in the same manner by selecting the units of lower rank according to $\rho_{i,1}$, after this number, in the new strata.

After rotation this no longer works: there is no longer any real independent of the numbers such that the sample s'_1 is composed of units of lower rank after it. This is true even in the case where $f_{h_1} = f_1$. The problem is obviously aggravated with f_{h_1} varying by stratum. The idea which then comes to mind is to first carry out a transformation of the numbers in such a way that those from s'_1 find themselves at the beginning of $[0, 1)$. This will then bring us back to the case without rotation. This is the same kind of idea which is presented by Hidiroglou, Choudhry and Lavallée (1991).

This transformation is fairly immediate in the particular case where the updates are done with procedure A and with the variable rounded number from section 7.1.2. Without

resampling, the selection interval at time t_2 would have been:

$$\rho_{i,1} \in [(t_2 - t_1)f_{h_1}/D_{h_1}, (t_2 - t_1)f_{h_1}/D_{h_1} + f_{h_1}].$$

The resampling results in new strata with probabilities f_{h_2} . These include the creations of units between the dates $t_2 - 1$ and t_2 , to which we attribute equidistant numbers $\rho_{i,1}$, in each stratum h_2 , independently of the survivors. They still contain units whose death has occurred since the previous sampling. It is possible to define a new sample s_2 in the same way as for Poisson sampling, by the interval, i.e.,

$$\rho_{i,1} \in [a_{h_1}, a_{h_1} + f_{h_2}],$$

where:

$$a_{h_1} = (t_2 - t_1)f_{h_1}/D_{h_1} + \max\left[0, f_{h_1}\left(1 - \frac{1}{D_{h_1}}\right) - f_{h_2}\right].$$

Let us recall that we shift from the supplementary quantity

$$f_{h_1}\left(1 - \frac{1}{D_{h_1}}\right) - f_{h_2}, \quad \text{if } f_{h_1}\left(1 - \frac{1}{D_{h_1}}\right) - f_{h_2} > 0,$$

to prevent the units which have just left the sample from returning too quickly.

As for Poisson sampling, the probability of a survivor being in the old and the new sample is then the maximum possible, namely:

$$\min\left(f_{h_1}\left(1 - \frac{1}{D_{h_1}}\right), f_{h_2}\right).$$

However the size n'_{h_2} of this sample is random, whereas we want a sample of fixed size n_{h_2} . We obtain it by selecting, in each new stratum h_2 , after having removed the deaths, the n_{h_2} units of lower rank according to $\eta_{i,1} = \rho_{i,1} - a_{h_1}$. This number therefore plays, for the resampling, the same role that ω_i played during the first sampling.

If, on the other hand, we chose procedure A with a fixed rounded number in the rotation or if we chose procedure B, we must begin again with the rank of the units of h_1 during the last updating. This is the rank according to ω_i with procedure A or the rank according to $\rho_{i,1} - 1$ with procedure B. Let us assume that N_{h_1} is the size of the population at date $t_2 - 1$. Let $R_{h_1,d}$ be the rank of the unit preceding the one of lower rank in s'_1 and $R_{h_1}(i)$ the rank of unit i . The number used to classify the units in the new strata becomes:

$$\eta_{i,1} = \frac{R_{h_1}(i) - 1 - a_{h_1} + \delta_{h_1}}{N_{h_1}} \text{ modulo } 1,$$

where:

$$a_{h_1} = R_{h_1,d} + \max\left(0, n'_{h_1}/N_{h_1} - f_{h_2}\right).$$

With procedure A we can keep $\delta_{h_1} = \theta_{h_1}$ while we make a choice of δ_{h_1} consistent with the last rounded number if procedure B is applied. However, because of the rotation, this choice has a minor impact on the coverage and it would be almost as well to select at random in $[0, 1)$.

9. CONCLUSION

Algorithms based on equidistant numbers do not produce SRS. The first-order probabilities of inclusion are not exactly controlled and the second-order probabilities are unknown. During the changes of stratum, there remains a "trace" of the former strata in the new strata. The application of the SRS formulae to estimate the variance leads to biased results, generally in the direction of over-estimation. However, we think that the improvement in coverage during resamplings provided by the algorithms based on equidistant numbers outweighs the disadvantage of biased estimation of the variance and of the confidence intervals. According to section 5, the finer the stratification the greater this advantage. In particular, the use of equidistant numbers seems to be quite indicated with procedure A where the strata (b, h) are likely to be very small for the waves of births $(b > 1)$. The advantage of equidistant numbers is not as great with procedure B. However, making the numbers of births equidistant renders both the number of survivors reselected at each updating of the sample and the duration of inclusion less random.

However, let's take a quick look at what would change in the maintenance if we wanted to conserve SSRS. At each stage we must conserve the independent and uniform distribution of the ω_i . First of all, the phases of updating the births and of rotation between resamplings described in sections 6 and 7 apply while still conserving the same ω_i and the procedure is even simpler. The most delicate part is the resampling after the intermediate phase of rotation. The objective is to obtain not only a SSRS but also, if possible, the same coverage as for method 1 in section 5.

Let us assume that $\alpha_{h_1}(j)$ is the number ω of the unit of rank j in a former stratum h_1 .

Let us assume first of all that, in a former stratum, all the units are such that $f_{h_2} \geq n'_{h_1}/N_{h_1}$. In particular, this occurs in all the strata for a sampling with a single rate in the sampled part, if we do not lower this rate. We then endeavour to find a transformation such that the numbers of the units of the sample are at the beginning of $[0, 1)$. The simplest is the permutation:

$$\begin{cases} \beta_{h_1}(j) = \alpha_{h_1}(j + N_{h_1} - R_{h_1,d}), & j \leq R_{h_1,d}, \\ \beta_{h_1}(j) = \alpha_{h_1}(j - R_{h_1,d}), & j > R_{h_1,d}. \end{cases}$$

However, a less costly transformation is:

$$\begin{cases} \beta_{h_1}(j) = \alpha_{h_1}(j) + \alpha_{h_1}(N_{h_1}) - \alpha_{h_1}(R_{h_1,d}), & j \leq R_{h_1,d}, \\ \beta_{h_1}(j) = \alpha_{h_1}(j) - \alpha_{h_1}(R_{h_1,d}), & j > R_{h_1,d}. \end{cases}$$

It is sufficient to find the result of $\alpha_{h_1}(R_{h_1,d})$ and $\alpha_{h_1}(N_{h_1})$, after which a simple sequential calculation makes it possible to deduct β from α .

The Jacobian of the transformation is equal to 1 and consequently the numbers conserve their uniform distribution. Moreover, the joint distribution $p(s_1, s_2)$ is the same as if there had been no rotation. The demonstration is provided in Cotton and Hesse (1992, page 55). We therefore have the maximum coverage of SSRS.

If this yields units with $f_{h_2} < n'_{h_1}/N_{h_1}$ in the stratum and we apply the transformation, the units whose rank falls approximately between $N_{h_1}f_{h_2}$ and n'_{h_1} are not reselected during the resampling but will be reintroduced during a future rotation. It is therefore preferable to use, for these units, a transformation which is situated just before f_{h_2} the new numbers. We must proceed by subsets according to the value of f_{h_2} . However, that tends to decrease the coverage.

ACKNOWLEDGEMENTS

The starting point of our work is an internal document from the Business Survey Methods Division of Statistics Canada: Hidioglou, M.A., Srinath, K.P. (1990), Methods of integrated sampling for sub-annual business surveys.

We would like to thank a co-writer and an anonymous referee for their assistance in the drafting of this article.

Some of the methods proposed have been applied to the INSEE, but the opinions expressed are solely those of the authors.

APPENDIX

Probabilities of Inclusion in the Kish and Scott Method (1971)

Let us consider an example where the first-order probability of inclusion is not strictly controlled.

The population is divided into three parts A , B and C of equal size N . The first sampling is a SRS of $2a$ units in $A + B$ and a SRS of a units in C . During the second sampling, we wish to select a units in A and $2a$ units in $B + C$, while retaining the maximum number of units from the first sample and with uniform probability of inclusion a/N . The Kish and Scott method consists in adding or removing by SRS the appropriate number of units separately in A and in $B + C$. In A , the second marginal sampling is a SRS and the probability of inclusion is quite

uniform. We will show that this is not the case in $B + C$. Let n_1 and n_2 be the sizes of the two successive samples in B . By symmetry, the probability of inclusion during the second sampling is uniform in B . It is equal to:

$$\begin{aligned} E(n_2)/N &= [E(n_1) + E(n_2 - n_1)]/N \\ &= a/N + E(n_2 - n_1)/N. \end{aligned}$$

If $n_1 = a, n_2 - n_1 = 0$; otherwise the expected value of $n_2 - n_1$ conditional on n_1 differs depending on the sign of $a - n_1$:

If $a - n_1 > 0, E[(n_2 - n_1) | n_1] = (a - n_1)(N - n_1)/(2N - n_1 - a)$.

If $a - n_1 < 0, E[(n_2 - n_1) | n_1] = (a - n_1)n_1/(n_1 + a)$.

Note $p(n_1)$ the probability that the first sample will have the size n_1 in B . This yields:

$$E(n_2 - n_1) = \sum_{n_1} p(n_1) E[(n_2 - n_1) | n_1].$$

Since the sizes of A and B are equal, $p(n_1) = p(2a - n_1)$, therefore:

$$\begin{aligned} E(n_2 - n_1) &= \sum_{n_1 < a} p(n_1) \{E[(n_2 - n_1) | n_1] + E[(n_2 - n_1) | (2a - n_1)]\} \\ &= \sum_{n_1 < a} p(n_1)(a - n_1)[(N - n_1)/(2N - n_1 - a) - (2a - n_1)/(3a - n_1)] \\ &= (2a - N) \sum_{n_1 < a} p(n_1)(a - n_1)^2 / [(2N - n_1 - a)(3a - n_1)] \\ &= (2a - N)K, K > 0. \end{aligned}$$

Except in the case $2a - N = 0, E(n_2 - n_1)$ is not nil and $E(n_2)/N$ is different from a/N . The probability of inclusion is therefore not uniform in $B + C$.

REFERENCES

- BREWER, K.R.W., EARLY, L.J., and HANIF, M. (1984). Poisson, modified Poisson and collocated sampling. *Journal of Statistical Planning and Inference*, 10, 15-30.
- COTTON, F., and HESSE C. (1992). Tirages coordonnés d'échantillons. INSEE working paper E9206.
- COX, L.H. (1987). A constructive procedure for unbiased controlled rounding. *Journal of the American Statistical Association*, 82, 520-524.
- HIDIROGLOU, M.A., CHOUDHRY, G.H., and LAVALLÉE, P. (1991). A sampling and estimation methodology for sub-annual business surveys. *Survey Methodology*, 17, 195-210.
- KISH, L., and SCOTT, A. (1971). Retaining units after changing strata and probabilities. *Journal of the American Statistical Association*, 66, 461-470.

Empirical Bayes Estimation of Small Area Proportions Based on Ordinal Outcome Variables

PATRICK J. FARRELL¹

ABSTRACT

Much research has been conducted into the modelling of ordinal responses. Some authors argue that, when the response variable is ordinal, inclusion of ordinality in the model to be estimated should improve model performance. Under the condition of ordinality, Campbell and Donner (1989) compared the asymptotic classification error rate of the multinomial logistic model to that of the ordinal logistic model of Anderson (1984). They showed that the ordinal logistic model had a lower expected asymptotic error rate than the multinomial logistic model. This paper also aims to compare the performance of ordinal and multinomial logistic models for ordinal responses. However, rather than focussing on classification efficiency, the assessment is made in the context of an application where the objective is to estimate small area proportions. More specifically, using multinomial and ordinal logistic models, the empirical Bayes approach proposed by Farrell, MacGibbon and Tomberlin (1997a) for estimating small area proportions based on binomial outcome data is extended to response variables consisting of more than two outcome categories. The properties of estimators based on these two models are compared via a simulation study in which the empirical Bayes methods proposed here are applied to data from the 1950 United States Census with the objective of predicting, for a small area, the proportion of individuals who belong to the various categories of an ordinal response variable representing income level.

KEY WORDS: Bootstrap; Complex survey design; Logistic regression; Random effects models; Small area summary statistics; Taylor series.

1. INTRODUCTION

Much research has been conducted into the modelling of ordinal responses (see Albert and Chib 1993, Anderson 1984, Crouchley 1995, and McCullagh 1980). Some authors argue that, when the response variable is ordinal, inclusion of ordinality in the model to be estimated should improve model performance. Under the condition of ordinality, Campbell and Donner (1989) theoretically compared the asymptotic classification error rate of the multinomial logistic model to that of the ordinal logistic model of Anderson (1984), demonstrating that the ordinal model had a lower expected asymptotic error rate. However, in a subsequent simulation study, Campbell, Donner, and Webster (1991) illustrated that ordinal models classify less accurately than multinomial models under a variety of circumstances, and concluded that ordinal models confer no advantage when the main purpose of an analysis is classification.

This paper also aims to compare the performance of ordinal and multinomial logistic models for ordinal responses. However, rather than focussing on classification efficiency, the assessment is made in the context of an application where the objective is to estimate small area proportions.

The estimation of small area parameters is a finite population sampling problem which has received considerable attention. An excellent review of such research appears in Ghosh and Rao (1994). These authors demonstrate that as a compromise between synthetic and direct

survey estimators, estimators based on empirical or hierarchical Bayes procedures are not subject to the large bias that is sometimes associated with a synthetic estimator (see Gonzales 1973), nor are they as variable as a direct survey estimator. A similar conclusion was drawn by Farrell, MacGibbon, and Tomberlin (1997a) in a study of the properties of an empirical Bayes estimator for small area proportions based on a binomial outcome variable.

Despite the numerous studies aimed at predicting small area proportions based on binomial response variables (see Dempster and Tomberlin 1980, MacGibbon and Tomberlin 1989, Farrell 1991, Farrell *et al.* 1997a, Malec, Sedransk, and Tompkins 1993, Stroud 1991, and Wong and Mason 1985), little attention has been given to estimating proportions based on response variables with more than two outcome categories. This paper extends the empirical Bayes approach of Farrell *et al.*, (1997a), to such response variables by basing the estimates on multinomial and ordinal logistic models. To compare the estimates of small area proportions based on an ordinal outcome variable using multinomial and ordinal models, the proposed empirical Bayes methods are applied to data from the 1950 United States Census in order to predict, for a given small area, the proportion of individuals who belong to the various categories of an ordinal response variable representing income level.

For such an estimation problem, there are many issues which require attention. They include the selection of predictor variables for the model, model diagnostics, the sample design, and the properties of the estimators

¹ Patrick J. Farrell, Assistant Professor, Department of Mathematics and Statistics, Acadia University, Wolfville, Nova Scotia, B0P 1X0.

employed. For example, among the model diagnostics for the multinomial and ordinal models was an assessment of model fit which was based on residuals. For a description of this diagnostic and others, see Farrell (1991). The findings did not appear to indicate a lack of fit for either model. In this study, the focus is on investigating the properties of empirical Bayes estimators over repeated realizations of the sample design using a simulation. For many survey practitioners, such properties are of prime importance.

One concern associated with using an empirical Bayes estimation approach is that interval estimates do not attain the desired level of coverage, since the uncertainty that arises from having to estimate the parameters of the prior distribution is not accounted for. This study incorporates the suggestion of Laird and Louis (1987) to use bootstrap techniques for adjusting naive estimates of accuracy. Alternatively, Prasad and Rao (1990) have developed a procedure which attempts to account for the uncertainty not captured by the naive estimates. Although their approach was designed for three specific linear models containing random effects, Cressie (1992) has made certain conjectures as to when the procedure is appropriate. Of importance is the constraint that the outcome variable must follow a normal distribution.

The proposed empirical Bayes procedures based on multinomial and ordinal logistic models are presented in Section 2. The simulation study to compare multinomial and ordinal logistic models for ordinal responses is described in Section 3, while the conclusions and discussion are presented in Section 4.

2. ESTIMATION PROCEDURES

Consider a discrete small area characteristic of interest with M possible outcomes. The subscript m will reference these categories, where $m = 1, \dots, M-1$ and $m^* = 1, \dots, M$. In addition, underlined lower case and capital letters will designate vectors, while bold capital letters will represent matrices.

The estimation procedures are illustrated under a two stage sample design, where individuals are sampled from selected local areas. Thus, local areas are the primary sampling units here. Let p_{im^*} be the proportion of individuals in the i -th local area that belong to category m^* of the response variable. Then

$$p_{im^*} = \sum_j y_{ijm^*} / N_i, \quad (2.1)$$

where y_{ijm^*} is either zero or one, depending upon whether the j -th individual in local area i belongs to category m^* of the characteristic of interest, and N_i is the population size of the i -th local area.

The approach employed by Farrell *et al.*, (1997a), to estimate small area proportions based on binomial outcome variables is extended here to allow for the estimation of p_{im^*} . The procedure follows the explicitly model-based

approach proposed by Dempster and Tomberlin (1980). Let π_{ijm^*} represent the probability that the j -th individual within the i -th local area belongs to category m^* of the response variable. Then, according to Royall (1970), p_{im^*} in (2.1) is estimated by

$$\hat{p}_{im^*} = \left(\sum_{j \in S} y_{ijm^*} + \sum_{j \in S'} \hat{\pi}_{ijm^*} \right) / N_i, \quad (2.2)$$

where S is the set of n_i sampled individuals from local area i , and S' is the set of individuals in local area i not included in the sample. Values for the $\hat{\pi}_{ijm^*}$ are required. To obtain these estimates, logistic regression models are used to describe the probabilities associated with individuals in the population.

Under a multinomial logistic model, the π_{ijm^*} are described as follows:

$$\begin{aligned} \log(\pi_{ijm} / \pi_{ijM}) &= \mathbf{X}_{ij}^T \boldsymbol{\beta}_m + \delta_{im}, \\ \delta_i &\sim \text{i.i.d. Normal}(\mathbf{0}, \mathbf{D}), \end{aligned} \quad (2.3)$$

where $\delta_i^T = (\delta_{i1}, \dots, \delta_{i(M-1)})$, $i = 1, \dots, I$, and \mathbf{D} is an unknown covariance matrix. In this model, \mathbf{X}_{ij} is a vector of fixed effects predictor variables, the vector $\boldsymbol{\beta}_m$ contains the fixed effects parameters associated with the m -th category of the outcome variable of interest, and δ_{im} is a normally distributed random effect associated with the m -th category of the characteristic of interest in the i -th local area. The vector \mathbf{X}_{ij} may include covariates at both the individual and aggregate levels. For sample designs of more than two stages, an analogous model would contain random effects for the sampling units at each stage, excluding the final one.

Note that the model in (2.3), unlike a similar model proposed by Malec *et al.*, (1993), does not contain interaction terms between the local area effects and the fixed effects predictor variables. However, terms to acknowledge such interaction could be included if they were deemed necessary.

To obtain Bayes estimates of the model parameters, values are assumed for the unknown parameters of the random effects distribution. Let $\mathbf{y}_{ij}^T = (y_{ij1}, \dots, y_{ijM})$ be a vector for the ij -th sampled individual where the component associated with the category of the outcome variable to which the individual belongs has a value of one. The remaining entries are zero. If \mathbf{Y} is a matrix with rows \mathbf{y}_{ij}^T , then the data are distributed as:

$$f(\mathbf{Y} | \boldsymbol{\beta}, \boldsymbol{\delta}_c) \propto \prod_{ij} \pi_{ij1}^{y_{ij1}} \pi_{ij2}^{y_{ij2}} \dots \pi_{ijM}^{y_{ijM}},$$

where $\boldsymbol{\beta}^T = (\boldsymbol{\beta}_1^T, \dots, \boldsymbol{\beta}_{M-1}^T)$, and $\boldsymbol{\delta}_c^T = (\boldsymbol{\delta}_1^T, \dots, \boldsymbol{\delta}_I^T)$. If a flat distribution is specified for the fixed effects, the distribution of the parameters is $f(\boldsymbol{\beta}, \boldsymbol{\delta}_c | \mathbf{D}_c) \propto \exp(-\frac{1}{2} \boldsymbol{\delta}_c^T \mathbf{D}_c \boldsymbol{\delta}_c)$, where $\mathbf{D}_c = \text{diag}(\mathbf{D}, \mathbf{D}, \dots, \mathbf{D})$. The joint distribution of the data and the parameters is determined using $f(\mathbf{Y} | \boldsymbol{\beta}, \boldsymbol{\delta}_c)$ and $f(\boldsymbol{\beta}, \boldsymbol{\delta}_c | \mathbf{D}_c)$, and subsequently employed to obtain the posterior distribution of the parameters. Unfortunately, a

closed form for this posterior distribution cannot be derived due to the intractable integration required to obtain the marginal distribution of Y . A possible approach could be a stochastic integration method such as Gibbs sampling (see Zeger and Karim 1991). Ripley and Kirkland (1990) indicate that the drawbacks of such an approach include the intensive computations and questions about when the sampling process has achieved equilibrium. Since computing time is of particular concern for the simulation discussed in Section 3, this approach will not be pursued here. Alternatively, Breslow and Clayton (1993) state that there is still room for simple, approximate methods. Many authors have found that a multivariate normal approximation of the posterior works very well in practice (see Farrell *et al.* 1997a, Laird 1978, Tomberlin 1988, and Wong and Mason 1985). Breslow and Lin (1995) warn, however, that such an approach might yield inconsistent estimates for the fixed effects parameters. Thus, if \hat{p}_{im+} is to be based on fixed effects estimates obtained in this manner, the same might apply to the consistency of \hat{p}_{im+} as an estimator for p_{im+} .

Following Farrell *et al.* (1997a), the posterior distribution of the parameters is approximated as a multivariate normal distribution having its mean at the mode and covariance matrix equal to the inverse of the information matrix evaluated at the mode. The information matrix here is simply the second derivative of the posterior distribution taken with respect to β and δ . When values are specified for the unknown parameters of the random effects distribution, the resulting mode and covariance matrix constitute an initial set of estimates of the model parameters. Empirical Bayes estimates are then obtained by using the EM algorithm described by Dempster, Laird, and Rubin (1977) to determine estimates for the parameters of the random effects distribution. The algorithm converges quickly, taking only a few minutes in real time. For details on how the empirical Bayes estimates are obtained for a model based on a two stage sample design and a binomial response variable, see MacGibbon and Tomberlin (1989).

The empirical Bayes estimates of the model parameters are used in (2.2) to determine \hat{p}_{im+} . In developing an expression for the uncertainty of \hat{p}_{im+} , N_i is assumed to be known. Since the approach being used is model-based and predictive in nature, the uncertainty in \hat{p}_{im+} arises solely from the $\sum \hat{\pi}_{ijm+}$ term; the $\sum y_{ijm+}$ term has zero variance. Thus, the mean square error of \hat{p}_{im+} as a predictor for p_{im+} can be estimated as

$$\widehat{\text{MSE}}(\hat{p}_{im+}) = \widehat{\text{Var}}\left(\frac{\sum_{j \in S'} \hat{\pi}_{ijm+}}{N_i}\right) + \frac{\sum_{j \in S'} \hat{\pi}_{ijm+}(1 - \hat{\pi}_{ijm+})}{N_i^2}. \quad (2.4)$$

For sampled local areas, where n_i is greater than zero, the first term of (2.4) is of order $1/n_i$, while the second term is of order $1/N_i$. In this study, the approximation of the mean square error of \hat{p}_{im+} is based on the first term only, which yields a useful approximation provided that N_i is large

compared to n_i . For nonsampled local areas, the first term in (2.4) is of order 1; therefore it always dominates the second term.

To estimate the uncertainty of \hat{p}_{im+} , which is expressed as a non-linear function of the estimators of the fixed and random effects, the expression for \hat{p}_{im+} is linearized by taking a first order multivariate Taylor series expansion about the realized values of the fixed and random effects. The variance of the resulting expression, call it $\widehat{\text{Var}}(\hat{p}_{im+})$, is taken as an estimate of the uncertainty of \hat{p}_{im+} . Details of the Taylor series expansion are given in Farrell *et al.*, (1997a), for a binomial outcome variable.

When population micro-data for auxiliary variables are not available, \hat{p}_{im+} in (2.2) cannot be determined. For non-linear models such as (2.3), prediction is not straightforward in this situation. However, an alternative estimator to \hat{p}_{im+} , say \tilde{p}_{im+} , which requires only local area summary statistics (a mean vector and finite population covariance matrix) for both continuous and categorical variables can be obtained by extending the approach proposed by Farrell, MacGibbon, and Tomberlin (1997b) for achieving this objective when estimating binomial small area parameters. The same Taylor series expansion that was used to estimate the accuracy of \hat{p}_{im+} can be employed to obtain a measure of the uncertainty for \tilde{p}_{im+} , $\widehat{\text{Var}}(\tilde{p}_{im+})$.

The approach described in this section can also be used to develop point and interval estimates for small area proportions based on \hat{p}_{im+} and \tilde{p}_{im+} when an ordinal model is used. In this study, a fixed and random effects model is proposed for the π_{ijm+} which is based on the ordinal model proposed by McCullagh (1980)

$$\log\left(\frac{\pi_{ij1} + \dots + \pi_{ijm}}{\pi_{ij(m+1)} + \dots + \pi_{ijm}}\right) = \beta_{0m} - X_{ij}^T \beta + \delta_{im}, \quad (2.5)$$

$$\delta_i \sim \text{i.i.d. Normal}(0, D).$$

The vector X_{ij} contains the values of the fixed effects predictor variables for the ij -th individual, while β represents a vector of fixed effects parameters. Associated with the m -th category of the response variable is a constant term, β_{0m} . The random effects are again assumed to be normally distributed. Note that an important feature of the model in (2.5) is that the restriction $\beta_{0(m+1)} - \beta_{0m} \geq \delta_{im} - \delta_{i(m+1)}$ must hold in order for $\pi_{ij(m+1)} \geq 0$. A discussion concerning this constraint is given in Section 3.

The approach used to approximate the uncertainty in \hat{p}_{im+} and \tilde{p}_{im+} when π_{ijm+} is based on either (2.3) or (2.5) can be described as naive, since $\widehat{\text{Var}}(\hat{p}_{im+})$ and $\widehat{\text{Var}}(\tilde{p}_{im+})$ do not account for the uncertainty which results from estimating the parameters of the random effects distribution. Thus, interval estimates for p_{im+} that are based on $\widehat{\text{Var}}(\hat{p}_{im+})$ and $\widehat{\text{Var}}(\tilde{p}_{im+})$ are typically too short. Many approaches have been proposed for addressing this issue (see Carlin and Gelfand 1990, and Laird and Louis 1987). In this study, the Type III bootstrap proposed by Laird and Louis (1987) is used to adjust naively-estimated measures of uncertainty. The procedure is described in Farrell *et al.*,

(1997a), for a binomial outcome variable. It can be extended to (2.3) and (2.5), and is applicable regardless of whether estimation is based on \hat{p}_{im+} or \bar{p}_{im+} .

The procedure requires that a number of bootstrap samples, N_B , be generated from a given set of data. Suppose that small area estimation is to be based on \hat{p}_{im+} . For the b -th bootstrap sample, an estimate \hat{p}_{bim+} for p_{im+} based on (2.3) or (2.5), along with a naive estimate of the variability of \hat{p}_{bim+} , $\widehat{\text{Var}}(\hat{p}_{bim+})$ are obtained. The quantities \hat{p}_{bim+} and $\widehat{\text{Var}}(\hat{p}_{bim+})$ are determined for each of N_B bootstrap samples, and used to calculate a bootstrap-adjusted estimate of the variability associated with \hat{p}_{im+} :

$$\widehat{\text{Var}}^{(B)}(\hat{p}_{im+}) = \frac{\sum_b \widehat{\text{Var}}(\hat{p}_{bim+})}{N_B} + \frac{\sum_b (\hat{p}_{bim+} - \hat{p}_{im+}^{(B)})^2}{N_B - 1},$$

$$\text{where } \hat{p}_{im+}^{(B)} = \frac{\sum_b \hat{p}_{bim+}}{N_B}.$$

Note that even though individuals are not selected by simple random sampling without replacement in this study, survey weights have not been attached to the records. However, in practice, the weights attached to a record will vary due to features of the survey design, such as differential nonresponse and clustering. In this study, the models account for the effects of these features. Further research is necessary to determine what impact the incorporation of survey weights into the models would have on the bootstrapping procedure.

3. A DATA EXAMPLE

A comparison of the estimates for small area proportions based on multinomial and ordinal logistic models was carried out using a simulation study where the response variable was ordinal. The data set is based on a 1% sample of the 1950 United States Census (United States Bureau of the Census 1984). Data based on the 1950 Census is used since it constitutes a public use microdata sample, and none of the more recent census data is available in this form. Thus, the results below for the multinomial and ordinal models are obtained by using predictor variable data for each individual within a local area. For a discussion of the difficulties encountered in obtaining microdata, see Bethlehem, Keller, and Pannekoek (1990).

The application considered is the estimation of the proportion of individuals in a given local area associated with each of the three categories of an ordinal outcome variable representing total personal income, where a local area is typically specified to be a state. This variable encompasses all sources of income, including wages and salaries, business income, and net income from other sources. An individual is regarded as having a low (less

than \$2,500), medium (\$2,500 to under \$10,000) or high (\$10,000 and over) level of total personal income during 1949. Thus, $m = 1$ for low income (Category 1), $m = 2$ for medium income (Category 2), and $m = 3$ for high income (Category 3). The multinomial and ordinal models were each used to obtain point and interval estimates for 42 local areas. Twenty of these areas were sampled, the others were not. Note that individuals with no income were included in Category 1. An alternative approach would have been a two stage model; a first stage logistic model for the probability of non-zero income, and a second stage multinomial or ordinal model for income category conditional on non-zero income.

In practice, historical data are often available for survey planning purposes. For example, variable selection for purposes of model predictions could be based on previous census data. To emulate this situation, a random sample of size 2,000 was selected from the 1% sample. Variables for model prediction were determined by applying a stepwise logistic regression procedure. The variables selected were age, gender, and race. With regards to race, individuals were categorized as white, negro, or other.

Thus, the multinomial and ordinal models used in this study included four individual level predictor variables for age, gender, and race (two indicator variables were required to code the various races). However, they also contained four local area variables representing average age, the proportion of males, the proportion of whites, and the proportion of negroes. Regardless of which model is considered, these local area variables are necessary since, when they are excluded, a relationship is noted between the expected value of \hat{p}_{im+} and its bias, where as the expected value increases, the bias increases from large negative to large positive values. The inclusion of domain level covariates removes this correlation. Therefore, since local area variables are also included in the models, the multinomial model contains eighteen fixed effects parameters (two for each of the individual level and local area predictor variables, and two constant terms) and forty random effects (two for each of the twenty sampled local areas), while the ordinal model contains ten fixed effects parameters (one for each of the individual level and local area predictor variables, and two constant terms) and forty random effects (two for each of the twenty sampled local areas). For a detailed study comparing logistic regression models for estimating small area proportions with and without domain level covariates which uses binomial outcome data, see Farrell *et al.*, (1997a).

The data for estimating the proportions of individuals in each local area belonging to the various income level categories were obtained from the 1% sample using a self-weighting two stage sample design. In the first stage, 20 out of 42 local areas were selected, without replacement, using probabilities proportional to size (PPS). More specifically, the approach used to select these local areas was randomized systematic selection of primary sampling units with PPS (see Kish 1965, p. 230). Then, at the second stage, 50 individuals were randomly selected from each

chosen local area. A total of 500 samples were drawn using this two stage design; however, resampling was not performed at the local area selection stage. Thus, the same 20 local areas were sampled in each of the 500 replicates. For these 20 sampled local areas, the average local area proportions for Categories 1, 2, and 3 of income level are 0.7142, 0.2260, and 0.0598.

Note that for the ordinal model, the constraint $\beta_{02} - \beta_{01} \geq \delta_{i1} - \delta_{i2}$ must hold in order for $\pi_{ij2} \geq 0$. A check of this constraint for each of the 500 samples using the estimates for the constant terms and the random effects indicated that it held at all times. In fact, it was discovered that in each of the 500 samples taken, the difference in the estimates for the constant terms was always positive, at least two orders of magnitude larger than the majority of the absolute differences of the random effects estimates, and always one order of magnitude bigger. Thus, the constant terms in the model dominate over the random effects.

To compare the properties of estimators for small area proportions over repeated realizations of the sample design, for each of the 500 samples selected the quantities \hat{p}_{im+} , $\widehat{\text{Var}}(\hat{p}_{im+})$, and $\widehat{\text{Var}}^{(B)}(\hat{p}_{im+})$ associated with each income level category were obtained for each local area, sampled or not, using both the multinomial and ordinal models. For each model, the estimates for $\widehat{\text{Var}}(\hat{p}_{im+})$ and $\widehat{\text{Var}}^{(B)}(\hat{p}_{im+})$ were used to construct naive and bootstrap-adjusted empirical Bayes symmetric 95% confidence intervals, respectively. Estimates for $\widehat{\text{Var}}^{(B)}(\hat{p}_{im+})$ were obtained by using the bootstrap procedure to generate 100 bootstrap samples from each of the 500 simulation samples.

Note that for the ordinal model, the constraint $\beta_{02} - \beta_{01} \geq \delta_{i1} - \delta_{i2}$ must also hold in the bootstrap procedure for random effects generated from an estimated distribution; otherwise negative estimates for some of the probabilities π_{ijm+} will result when creating bootstrap samples. Over the course of the simulation for the application considered here, no negative probabilities were encountered when bootstrapping. One approach for assessing the likelihood of negative probabilities during the bootstrap procedure is to consider the ratio of the difference $\hat{\beta}_{02} - \hat{\beta}_{01}$ to the estimated prior standard deviation of the difference $\hat{\delta}_{i1} - \hat{\delta}_{i2}$. This ratio was determined for each sampled local area in each of the 500 simulation samples taken. The average of this entire set of ratios was 6.8, and none were found to be less than 5.8. Thus, the difference $\hat{\beta}_{02} - \hat{\beta}_{01}$ was determined to always be at least 5.8 times the estimated standard deviation of the difference $\hat{\delta}_{i1} - \hat{\delta}_{i2}$. Based on the empirical rule, a rule of thumb would be to conclude that when the ratio described above is at least three, it is highly unlikely that negative probabilities will arise when bootstrapping.

Table 1 presents average summary statistics over the 500 simulation samples obtained for the multinomial and ordinal models across all sampled local areas for each of three income level categories. A study of the stability of these statistics was conducted by investigating how they changed as additional samples were taken. Only slight

changes were observed once 150 samples had been reached. Table 1 includes the summary statistics obtained for the first 200 samples in brackets for comparative purposes.

For each income category, two summary statistics shown in Table 1 were evaluated to compare the design bias of \hat{p}_{im+} for the multinomial and ordinal models; the average bias of \hat{p}_{im+} , and the average absolute bias of \hat{p}_{im+} . The average bias is simply the mean over all sampled local areas of the differences obtained when the actual proportion, p_{im+} , for the i -th local area is subtracted from the average point estimate for the area over the 500 simulation samples. The average absolute bias is defined similarly, except that the absolute value of each difference is used. Generally speaking, the results obtained for these two summary statistics were slightly better for the ordinal model, regardless of the income category considered. However, the multinomial model did result in a somewhat smaller average bias for \hat{p}_{im+} for the low income category.

For each sampled local area, empirical root mean square errors (RMSE's) were computed over the 500 simulation samples under each model for the three income categories. For each model and income level combination, the appropriate empirical RMSE's were averaged over all sampled local areas, resulting in the average empirical RMSE's presented in Table 1. Once again, the performance of the ordinal model is slightly better for all three income level categories.

To study the reduction in empirical RMSE when a model-based approach to estimation is used instead of a classical design unbiased method, average empirical RMSE's analogous to those in Table 1 based on the 500 samples were computed using the observed local area sample proportions in place of \hat{p}_{im+} . The average empirical RMSE's obtained were substantially larger (0.0617, 0.0564, and 0.0311 for the low, medium, and high income level categories) than those based on \hat{p}_{im+} under either model.

Table 1 also includes summary statistics over all sampled local areas which relate naive and bootstrap measures of variability in \hat{p}_{im+} to average empirical RMSE. For each income level category, the average relative bias and the average absolute relative bias of the square root of $\widehat{\text{Var}}(\hat{p}_{im+})$ as an estimate of empirical RMSE are shown in Table 1 for the multinomial and ordinal models. The average relative bias is simply the mean over all sampled local areas of the values obtained when the difference resulting from the subtraction of the empirical RMSE for the i -th local area from the average of the square root of $\widehat{\text{Var}}(\hat{p}_{im+})$ for the area over the 500 simulation samples is divided by the empirical RMSE. The average absolute bias is defined similarly, except that the absolute value of each difference is used. The table also presents similar averages for the bootstrap-adjusted measures of variability, $\widehat{\text{Var}}^{(B)}(\hat{p}_{im+})$. For both the multinomial and ordinal logistic models, the average relative bias and average absolute relative bias of the bootstrap-adjusted estimates of variability are substantially smaller in magnitude than their naive counterparts for all three income level categories. In

Table 1

Average Summary Statistics based on 500 Simulation Samples for the Multinomial and Ordinal Logistic Models across all Sampled Local Areas for each Income Level Category.

The average summary statistics obtained over the first 200 simulation samples are included in brackets for comparative purposes

Average	Low Income Level		Medium Income Level		High Income Level	
	Multinomial	Ordinal	Multinomial	Ordinal	Multinomial	Ordinal
Bias of \hat{p}_{im+}	-0.0004 (-0.0004)	-0.0005 (-0.0006)	-0.0007 (-0.0006)	-0.0004 (-0.0003)	0.0011 (0.0010)	0.0009 (0.0009)
Absolute Bias of \hat{p}_{im+}	0.0076 (0.0078)	0.0051 (0.0055)	0.0089 (0.0085)	0.0048 (0.0046)	0.0108 (0.0106)	0.0074 (0.0073)
Empirical RMSE	0.0479 (0.0483)	0.0467 (0.0469)	0.0417 (0.0414)	0.0401 (0.0402)	0.0236 (0.0233)	0.0231 (0.0229)
Relative Bias of $\sqrt{\text{Var}(\hat{p}_{im+})}$	-0.1192 (-0.1197)	-0.1125 (-0.1128)	-0.1273 (-0.1276)	-0.1180 (-0.1186)	-0.1524 (-0.1521)	-0.1376 (-0.1372)
Absolute Relative Bias of $\sqrt{\text{Var}(\hat{p}_{im+})}$	0.1192 (0.1197)	0.1125 (0.1128)	0.1273 (0.1276)	0.1180 (0.1186)	0.1524 (0.1521)	0.1376 (0.1372)
Relative Bias of $\sqrt{\text{Var}^{(B)}(\hat{p}_{im+})}$	-0.0275 (-0.0272)	-0.0173 (-0.0175)	-0.0309 (-0.0314)	-0.0204 (-0.0207)	-0.0391 (-0.0393)	-0.0273 (-0.0269)
Absolute Relative Bias of $\sqrt{\text{Var}^{(B)}(\hat{p}_{im+})}$	0.0294 (0.0290)	0.0227 (0.0228)	0.0349 (0.0343)	0.0263 (0.0265)	0.0450 (0.0446)	0.0353 (0.0347)
Naive Coverage Rate	91.35 (91.325)	91.91 (91.875)	91.19 (91.225)	91.78 (91.750)	90.67 (90.650)	91.26 (91.300)
Absolute Deviation of Naive Coverage from the 95% Nominal Rate	3.65 (3.675)	3.09 (3.125)	3.81 (3.775)	3.22 (3.250)	4.33 (4.350)	3.74 (3.700)
Adjusted Coverage Rate	94.44 (94.400)	94.75 (94.775)	94.37 (94.350)	94.68 (94.650)	93.91 (93.925)	94.40 (94.375)
Absolute Deviation of Adjusted Coverage from the 95% Nominal Rate	1.58 (1.600)	1.43 (1.425)	1.71 (1.725)	1.50 (1.525)	1.91 (1.900)	1.62 (1.650)

addition, these bootstrap-adjusted average summary statistics are all very small, which indicates that the bootstrap-adjusted estimates of variability are capable of incorporating most of the uncertainty that arises from having to estimate the distribution of the random effects.

For each sampled local area, naive and bootstrap-adjusted coverage rates based on 95% interval estimates were computed over the 500 samples under each model for the three income level categories. Over all income level and model combinations, the bootstrap-adjusted coverage rates for individual local areas ranged from 92.2% to 97.6%. Since an approximate bound for the Monte Carlo error is $3\sqrt{(0.95)(0.05)/500}$, or 0.029, all bootstrap-adjusted coverage rates are within 3 standard errors of 95%.

For each model and income level combination, the appropriate coverage rates were averaged over all sampled local areas, resulting in the average naive and bootstrap-adjusted coverage rates in Table 1. A number of observations can be made which hold for each income level category. For both multinomial and ordinal models, the average coverage rates for the bootstrap-adjusted intervals are much closer to the 95% nominal rate than those associated with the naive intervals. However, both the average naive and bootstrap-adjusted coverage rates for the

ordinal model are slightly better than counterparts for the multinomial model. This is also the case for the average absolute deviation of both the naive and bootstrap-adjusted coverage rates from the 95% nominal rate. The average absolute deviation of the naive coverage rates from the 95% nominal rate is simply the mean over all sampled local areas of the absolute values of the differences obtained when the 95% nominal rate is subtracted from the naive coverage rates for the sampled local areas over the 500 simulation samples. The average absolute deviation of the bootstrap-adjusted coverage rates from the 95% nominal rate is defined analogously.

Twenty-two local areas were not sampled. Estimates for the proportion of individuals associated with each income level category were also obtained for these areas using the multinomial and ordinal models. The findings were similar to those for sampled local areas. However, the performance of the models deteriorated somewhat, since nonsampled local areas constitute a holdout sample. For a detailed evaluation of results associated with nonsampled local areas, see Farrell *et al.* (1997a).

A comparison of the estimates for the three income level categories based on micro-data, \hat{p}_{im+} , with those based on local area summary statistics, \bar{p}_{im+} , was also made for each

model. For both models, the results obtained for \tilde{p}_{im*} were gratifyingly close to those obtained using \hat{p}_{im*} , although those obtained for \tilde{p}_{im*} were slightly better. Similar findings were obtained by Farrell *et al.*, (1997b) in a detailed comparison of \hat{p}_{im*} and \tilde{p}_{im*} for a binomial outcome variable.

4. CONCLUSION

Using multinomial and ordinal logistic models, the empirical Bayes approach proposed by Farrell *et al.*, (1997a), for estimating small area proportions based on binomial outcome data has been extended to accommodate outcome variables with more than two categories. It was found that the performance of the approach is preserved for multicategorical outcome data.

To compare the estimates of small area proportions based on an ordinal outcome variable using multinomial and ordinal logistic models, the proposed empirical Bayes methods based on these two models were applied to data from the 1950 United States Census with the objective of predicting, for a small area, the proportion of individuals who belong to the various categories of an ordinal response variable representing income level. The estimates based on the ordinal model were only slightly better in terms of design bias, empirical RMSE, and coverage rates. In addition, an important feature of the ordinal logistic model is that the constraint $\beta_{0(m+1)} - \beta_{0m} \geq \delta_{im} - \delta_{i(m+1)}$ must hold in order for $\pi_{ij(m+1)} \geq 0$. Since the results for the multinomial and ordinal models in the simulation were very similar, a multinomial model could be used for estimating small area proportions based on ordinal outcome variables when there is concern that fitting an ordinal model may result in negative estimates for some of these probabilities.

ACKNOWLEDGEMENTS

This research was supported by NSERC of Canada. The author is grateful to the associate editor and the referees for their valuable comments and suggestions.

REFERENCES

- ALBERT, J.H., and CHIB, S. (1993). Bayesian analysis of binary and polytomous response data. *Journal of the American Statistical Association*, 88, 669-679.
- ANDERSON, J.A. (1984). Regression and ordered categorical variables. *Journal of the Royal Statistical Society, Series B*, 46, 1-30.
- BETHLEHEM, J.G., KELLER, W.J., and PANNEKOEK, J. (1990). Disclosure control of microdata. *Journal of the American Statistical Association*, 85, 38-45.
- BRESLOW, N.E., and CLAYTON, D.G. (1993). Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association*, 88, 9-25.
- BRESLOW, N.E., and LIN, X. (1995). Bias correction in generalised linear mixed models with a single component of dispersion. *Biometrika*, 82, 81-91.
- CAMPBELL, M.K., and DONNER, A. (1989). Classification efficiency of multinomial logistic regression relative to ordinal logistic regression. *Journal of the American Statistical Association*, 84, 587-591.
- CAMPBELL, M.K., DONNER, A., and WEBSTER, K.M. (1991). Are ordinal models useful for classification? *Statistics in Medicine*, 10, 383-394.
- CARLIN, B.P., and GELFAND, A.E. (1990). Approaches for empirical Bayes confidence intervals. *Journal of the American Statistical Association*, 85, 105-114.
- CRESSIE, N. (1992). REML Estimation in empirical Bayes smoothing of census undercount. *Survey Methodology*, 18, 75-94.
- CROUCHLEY, R. (1995). A random-effects model for ordered categorical data. *Journal of the American Statistical Association*, 90, 489-498.
- DEMPSTER, A.P., LAIRD, N.M., and RUBIN, D.B. (1977). Maximum likelihood estimation from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B*, 39, 1-38.
- DEMPSTER, A.P., and TOMBERLIN, T.J. (1980). The analysis of census undercount from a postenumeration survey. *Proceedings of the Conference on Census Undercount*, Arlington, VA, 88-94.
- FARRELL, P.J. (1991). Empirical Bayes Estimation of Small Area Proportions. PhD. dissertation, Department of Management Science, McGill University, Montreal, Quebec, Canada.
- FARRELL, P.J., MacGIBBON, B., and TOMBERLIN, T.J. (1997a). Empirical Bayes estimators of small area proportions in multistage designs. *Statistica Sinica*, 7, 1065-1083.
- FARRELL, P.J., MacGIBBON, B., and TOMBERLIN, T.J. (1997b). Empirical Bayes small area estimation using logistic regression models and summary statistics. *Journal of Business and Economic Statistics*, 15, 101-108.
- GHOSH, M., and RAO, J.N.K. (1994). Small area estimation: an appraisal. *Statistical Science*, 9, 55-93.
- GONZALES, M.E. (1973). Use and evaluation of synthetic estimation. *Proceedings of the Social Statistics Section, American Statistical Association*, 33-36.
- KISH, L. (1965). *Survey Sampling*. New York: John Wiley & Sons Inc.
- LAIRD, N.M. (1978). Empirical Bayes methods for two-way contingency tables. *Biometrika*, 65, 581-590.
- LAIRD, N.M., and LOUIS, T.A. (1987). Empirical Bayes confidence intervals based on bootstrap samples. *Journal of the American Statistical Association*, 82, 739-750.
- MacGIBBON, B., and TOMBERLIN, T.J. (1989). Small area estimates of proportions via empirical Bayes techniques. *Survey Methodology*, 15, 237-252.
- MALEC, D., SEDRANSK, J., and TOMPKINS, L. (1993). Bayesian predictive inference for small areas for binary variables in the National Health Interview Survey. In *Case Studies in Bayesian Statistics*, (Eds. C. Gatsonis, J.S. Hodges, R. Kasf, and N.D. Singpurwalla). New York: Springer Verlag.

- McCULLAGH, P. (1980). Regression models for ordinal data. *Journal of the Royal Statistical Society, Series B*, 42, 109-142.
- PRASAD, N.G.N., and RAO, J.N.K. (1990). On the estimation of mean square error of small area predictors. *Journal of the American Statistical Association*, 85, 163-171.
- RIPLEY, B.D., and KIRKLAND, M.D. (1990). Iterative simulation methods. *Journal of Computational and Applied Mathematics*, 31, 165-172.
- ROYALL, R.M. (1970). On finite population sampling theory under certain linear regression models. *Biometrika*, 74, 1-12.
- STROUD, T.W.F. (1991). Hierarchical Bayes predictive means and variances with application to sample survey inference. *Communications in Statistics, Theory and Methods*, 20, 13-36.
- TOMBERLIN, T.J. (1988). Predicting accident frequencies for drivers classified by two factors. *Journal of the American Statistical Association*, 83, 309-321.
- UNITED STATES BUREAU OF THE CENSUS (1984). Census of the Population, 1950: Public Use Microdata Sample Technical Documentation, edited by J.G. Keane, Washington, D.C.
- WONG, G.Y., and MASON, W.M. (1985). The hierarchical logistic regression model for multilevel analysis. *Journal of the American Statistical Association*, 80, 513-524.
- ZEGER, S.L., and KARIM, M.R. (1991). Generalized linear models with random effects; a Gibbs sampling approach. *Journal of the American Statistical Association*, 86, 79-86.

Poststratification Into Many Categories Using Hierarchical Logistic Regression

ANDREW GELMAN and THOMAS C. LITTLE¹

ABSTRACT

A standard method for correcting for unequal sampling probabilities and nonresponse in sample surveys is poststratification: that is, dividing the population into several categories, estimating the distribution of responses in each category, and then counting each category in proportion to its size in the population. We consider poststratification as a general framework that includes many weighting schemes used in survey analysis (see Little 1993). We construct a hierarchical logistic regression model for the mean of a binary response variable conditional on poststratification cells. The hierarchical model allows us to fit many more cells than is possible using classical methods, and thus to include much more population-level information, while at the same time including all the information used in standard survey sampling inferences. We are thus combining the modeling approach often used in small-area estimation with the population information used in poststratification. We apply the method to a set of U.S. pre-election polls, poststratified by state as well as the usual demographic variables. We evaluate the models graphically by comparing to state-level election outcomes.

KEY WORDS: Bayesian inference; Election forecasting; Nonresponse; Opinion polls; Sample surveys.

1. INTRODUCTION

It is standard practice for weighting in opinion polls to be based entirely or primarily on poststratification, which we use generally to refer to any estimation scheme that adjusts to population totals. The basic approach is to divide the population into a number of categories, within each of which the survey is analyzed as simple random sampling. The poststratification step is to estimate population quantities by averaging estimates in the categories, counting each category in proportion to its size in the population. Poststratification categories are typically based on demographic characteristics (sex, age, *etc.*) as well as any variables used in stratification. Another level of complication, which we do not address here, would occur under cluster sampling.

There is a fundamental difficulty in setting up poststratification categories. It is desirable to divide the population into many small categories in order for the assumption of simple random sampling within categories to be reasonable. But if the number of respondents per category is small, it is difficult to accurately estimate the average response within each category. For example, if we poststratify by sex, ethnicity, age, education, and region of the U.S., some cells may be empty in the sample, whereas others may have only one or two respondents.

A general solution to this problem is to model the responses conditional on the poststratification variables (see Little 1993). For example, the standard approach to adjusting for several demographic variables is to rake across one-way or two-way margins (*i.e.*, iterative proportional fitting, Deming and Stephan 1940), which essentially corresponds to poststratification on the complete multi-way table, but with a model of the responses,

conditional on the demographic variables, that sets higher-level interactions to zero. Methods based on smoothing weights can also be viewed as poststratification, with corresponding models on the responses (see Little 1991). When the poststratification categories follow a hierarchical structure (for example, persons within states in the U.S.), one can improve efficiency of estimation by fitting a hierarchical model (*e.g.*, Lazzeroni and Little 1997). In the related context of regression estimation, Longford (1996) demonstrates the potential for hierarchical linear models to improve the precision of small area estimates based on sample survey data.

In this paper, we set up a hierarchical logistic regression model to be used for poststratification estimates for a binary variable. The advantage of the model, compared to standard poststratification, is that it allows for the use of many more categories, and thus much more detailed population information. The practical gains from this method are greatest for small subgroups of the population. We apply the method to the state-level results of a set of U.S. pre-election polls. This example has the nice feature that we can check our inferences externally by comparing to state-level election outcomes. Details appear in an appendix for computing the hierarchical model using an approximate EM algorithm.

2. MODEL

2.1 Sampling and Poststratification Information

Consider a partition of the population into R categorical variables, where the r -th variable has J_r levels, for a total of $J = \prod_{r=1}^R J_r$ categories (cells), which we label $j = 1, \dots, J$.

¹ Andrew Gelman, Department of Statistics, Columbia University, New York, NY 10027 and Thomas C. Little, Morgan Stanley Dean Witter, New York, NY.

Assume that N_j , the number of units in the population in category j , is known for all j . Let y be a binary response of interest; label the population mean response in each category j as π_j . Then the overall population mean is $\bar{Y} = \sum_j N_j \pi_j / \sum_j N_j$. Assume that the population is large enough that we can ignore all finite-population corrections.

A sample survey is now conducted in order to estimate \bar{Y} (and perhaps some other combinations of the π_j 's). For each j , let n_j be the number of units in category j in the sample. Conditional on the R explanatory variables, assume that nonresponse is ignorable (Rubin 1976). Thus, the R variables should include all information used to construct survey weights, as well as any other variables that might be informative about y .

For the example we shall consider in Section 3, we categorize the population of adults in the 48 contiguous U.S. states by $R = 5$ variables: state of residence, sex, ethnicity, age, and education, with $(J_1, \dots, J_5) = (48, 2, 2, 4, 4)$. (Ethnicity, age, and education are discretized into 4 categories each, as described in Section 3.1.) The $J = 3,072$ categories range from "Alabama, male, black, 18-29, not high school graduate" to "Wyoming, female, nonblack, 65+, college graduate," and, from the U.S. Census, we have good estimates of N_j in each of these categories. We shall consider population estimates (summing over all 3,072 categories) and also estimates within individual states (separately summing over the 64 categories for each state). It is impossible for a reasonably-sized sample survey to allow independent estimates of the mean responses π_j for each category j (in fact, the vast majority of categories will be empty or contain just one respondent), and so it is necessary to model the π_j 's in order to poststratify and thus make use of the known category sizes N_j . The (potential) advantage of poststratification is to correct for differential nonresponse rates among the categories.

2.2 Regression Modelling in the Context of Poststratification

One can set up a logistic regression model for the probability π_j of a "yes" for respondents in category j :

$$\text{logit}(\pi_j) = X_j \beta, \quad (1)$$

where X is a matrix of indicator variables, and X_j is the j -th row of X . If we were to assume a uniform prior distribution on β , then Bayesian inference, for different choices of X , under this model corresponds closely to various classical weighting schemes. These correspondences, which we present below, are general and rely on the linearity of the assumed model (that is, $X_j \beta$ in (1)). (In the case of binary data, which we are considering in this paper, the classical and uniform-prior-Bayesian estimates are not identical, because of the nonlinear logistic transformation in (1), but for large samples the differences are minor.)

The following models correspond to the most commonly used classical poststratification estimates.

- Setting X to the $J \times J$ identity matrix corresponds to weighting each unit in cell j by N_j/n_j ; that is, simple poststratification. This method is well known to work well only if the n_j 's are reasonably large (and it will not work at all if $n_j = 0$ for any j).
- If we set X to the $J \times (\sum_{r=1}^R J_r)$ matrix of indicators for each individual variable, then the estimate of \bar{Y} corresponds approximately to that obtained by raking across all R one-way margins.
- Including various interactions in X corresponds to including these same interactions in the raking. To put it most generally, assuming "structure" of any kind in X corresponds to pooling the poststratification across cells in some way.
- Including no explanatory variables in the model (that is letting X be simply a vector of 1's) leads to the sample mean estimate \bar{y} .

See Holt and Smith (1979) and Little (1993) for more discussion of the relation between weighting estimates and poststratification.

2.3 Hierarchical Regression Modelling for Partial Pooling

When the number of cells is large, none of the above options makes efficient use of the information provided by the categories (for example, simple poststratification gives estimates that are too variable, but if we exclude explanatory variables with many categories, we are discarding important information). Instead, we allow partial pooling across cells by setting up a mixed-effects model (see, e.g., Clayton 1996). We write the vector β as $(\alpha, \gamma_1, \dots, \gamma_L)$, where α is a subvector of unpooled coefficients and each γ_l , for $l = 1, \dots, L$, is a subvector of coefficients (γ_{kl}) to which we fit a hierarchical model:

$$\gamma_{kl} \sim N(0, \tau_l^2), \quad k = 1, \dots, K_l, \quad \text{ind}$$

Setting τ_l to zero corresponds to excluding a set of variables; setting τ_l to ∞ corresponds to a noninformative prior distribution on the γ_{kl} parameters.

Given the responses y_i in categories j , we construct an $n \times J$ categorization matrix C , for which $C_{ij} = 1$ if respondent i is in cell j . Let $Z = CX$. The model (1) then can be written in the standard form of a hierarchical logistic regression model as

$$y_i \sim \text{Bernoulli}(p_i)$$

$$\text{logit}(p_i) = Z\beta$$

$$\beta \sim N(0, \sum_{\beta}),$$

where \sum_{β}^{-1} is a diagonal matrix with 0 for each element of α , followed by τ_l^{-2} for each element of γ_l , for each l . We use the notation p_i , for the probability corresponding to the unit i , as distinguished from π_j , the aggregate probability corresponding to the category j . See Nordberg (1989) and

Belin, Diffendal, Mack, Rubin, Schafer and Zaslavsky (1993) for general discussions of hierarchical logistic regression models for survey data.

2.4 Inference Under the Model

To perform inferences about population quantities, we use the following empirical Bayes strategy: first, estimate the hyperparameters τ_j , given the data y ; second, perform Bayesian inference for the regression coefficients β , given y and the estimated τ_j 's; third, compute inferences for the vector of cell means $\pi = \text{logit}^{-1}(X\beta)$; fourth, compute inferences for population quantities by summing $N_j\pi_j$'s. We view this approach as an approximation to the full Bayesian analysis, which averages over the parameters τ_j . The two approaches will differ the most when components τ_j are imprecisely estimated or are indistinguishable from 0 (see for example, Gelman, Carlin, Stern and Rubin (1995), Section 5.5). In the example we consider here, this is not a problem because the various components are clearly estimated to be different from 0. If this were not the case, it would probably be worth putting in the additional programming effort for a full Bayes analysis. The focus of this paper, however, is on the effectiveness of combining hierarchical modeling with poststratification, not on the relatively minor technical differences between Bayes and empirical Bayes analyses.

The shrinkage of the cell estimates comes in the second step, and the amount of shrinkage depends both on the sample sizes n_j and the data \bar{y}_j . More shrinkage occurs for smaller values of n_j and for values of \bar{y}_j far from the predictions based on the logistic regression model. In addition, more shrinkage occurs if the parameters τ_j are small. A batch of coefficients γ_i with little predictive power will be shrunk toward zero in the estimation, because τ_i will be estimated to have a small value. This is how we can include a large number of coefficients in the hierarchical model without the estimates of population quantities becoming too variable.

3. APPLICATION: BREAKING DOWN NATIONAL SURVEYS BY STATE

3.1 Survey Data

We apply the above methodology to state-by-state results from seven national opinion polls of registered voters conducted by the CBS television network during the two weeks immediately preceding the 1988 U.S. Presidential election. To follow our general notation, we assign $y_i = 1$ to supporters of Bush and $y_i = 0$ to supporters of Dukakis; we discard the respondents who expressed no opinion (about 15% of the total; we follow standard practice and count respondents who "lean" toward one of the candidates as full supporters). Since no data were collected from Hawaii and Alaska, only the 48 contiguous states are included in the model. Washington, D.C., although included in the surveys, was excluded from this analysis

because its voting preferences are so different from the other states that a generalized linear model that fit the 48 states would not fit D.C. well, and as a result, the data from D.C. would unduly influence the results for the states. Since there are few observations for the smaller states and the between-poll variation in the estimated support for Bush is within binomial sampling variability (as measured by a χ^2 test of equality of the proportions of support for Bush in the seven polls), we combine the data from all the polls.

CBS creates survey weights by raking on the following variables, with default classifications for item nonresponse shown in brackets:

Census region:	Northeast, South, North Central, West
sex:	male, female
ethnicity:	black, [white/other]
age:	18-29, 30-44, [45-64], 65+
education:	not high school grad, [high school grad], some college, college grad.

The raking includes all main effects plus the interactions of sex \times ethnicity and age \times education. We include all these variables as fixed effects in our logistic regression model, excluding from our analysis the relatively few respondents with nonresponse in any of the demographic variables. The CBS weights also correct for number of telephone lines and number of adults in household, which affect sampling probabilities; these have minor effects on estimates for Presidential preference (see Little 1996, chapter 3), and we do not include them in our model. Further details of the CBS survey methodology and adjustment appear in Voss, Gelman, and King (1995).

Our model goes beyond the CBS analysis by including indicators for the 48 states as random effects, clustered into four batches corresponding to the four census regions. We check the performance of the model by comparing estimates for each state to the observed Presidential election. (Opinion polls just before the election are reliable indicators of the actual election outcome; see, e.g., Gelman and King 1993.) We also compare the stability of estimates based on different polls over a short period of time.

3.2 Population Data for Poststratification

In order to poststratify on all the variables listed above, along with state, we need the joint population distribution of the demographic variables within each state: that is, population totals N_j for each of the $2 \times 2 \times 4 \times 48$ cells of sex \times ethnicity \times age \times state. Since the target population is registered voters, we should use the population distribution of registered voters. As an approximation to that distribution we use the crosstabulations available in the Public Use Micro Survey (PUMS) data for all citizens of age 18 and over. The PUMS data contain records for 5% of the housing units in the U.S. and the persons in them, including over 12 million persons and over 5 million housing units. These data are a stratified sample of the approximately 15.9% of housing units that received long-form questionnaires in the 1990 Census. Persons in

institutions and other group quarters are also included in the sample. Weights are given for both the housing unit and persons within the unit based on sampling probabilities and adjustment to Census totals for variables included in the short-form questionnaire. We use the weighted PUMS data to estimate N_j for each poststratification category and ignore sampling error in these numbers. The weighted PUMS numbers are very similar to the poststratification numbers used by CBS in their raking (see Little 1996, chapter 3).

3.3 Results

We present results for four methods applied to the combined data from the seven surveys:

1. Classical estimate based on raking by demographic variables (region, sex, ethnicity, age, education, sex \times ethnicity, and age \times education). This is very close to the weighting method used by CBS. For estimates of results by states, we perform weighted averages within each state, using the weights obtained by the raking.
2. Regression estimate using the demographic variables and also indicators for the states, with no hierarchical model (*i.e.*, "fixed-effects" regression). This is very similar to using iterative proportional fitting to rake on states as well as demographics. The state-by-state estimates from this model should improve upon those obtained by raking on demographics because the estimates of π_j 's are weighted by the population numbers N_j rather than the sample numbers n_j within each state.
3. Regression estimate using only the demographic variables, with the state effects set to zero. This model allows the average responses within states to differ only because of demographic variation; to the extent that the demographics do not explain all the variation in opinion, the model should underestimate the variability between states.
4. Regression estimate using the demographic variables, with the 48 state effects estimated with a hierarchical model (in the notation of Section 2, $L = 4$ and $K_1, K_2, K_3, K_4 = 12, 13, 12, 11$). We expect this model to perform best, both because of the flexibility of the hierarchical regression model and because the poststratification uses the population numbers N_j .

We fit each of the regression models to the survey data, obtain posterior simulation draws for each coefficient (conditional on the estimated $\tau_1, \tau_2, \tau_3, \tau_4$), and reweight based on the PUMS data to obtain poststratified estimates for the proportion of registered voters in each state who support Bush for President.

Table 1 presents the raking estimate and the posterior medians and interquartile ranges for the three models, along with data on the survey responses and the actual election outcome. Table 2 gives the nationwide and mean absolute statewide prediction errors for the raking and the three models. The four methods give almost identical results at the national level; the real gain from the model-based

estimates occurs in estimating the individual states. The reduction in mean absolute prediction error from about 6% to 5% can be attributed to using the poststratification information, with the further reduction to 3.5% attributable to the hierarchical modeling. In addition, the last two lines of Table 2 show that the uncertainty estimates from the hierarchical model are short and relatively well calibrated (slightly less than half of the true values fall inside the 50% intervals, which is reasonable since these intervals account only for sampling error and not for nonsampling errors and changes in opinion).

Figure 1 plots, by state, the actual election outcomes vs. the raking estimates and the posterior medians for the three models. As one would expect, the hierarchical model reduces variance, and thus estimation error, by shrinkage. Although the four methods correct the bias of the nationwide estimate by about the same amount, they act differently on the individual states, with the hierarchical model performing best. Figure 2 compares the prediction errors for the hierarchical and raking estimates for the states.

Interestingly, the hierarchical model does not seem to shrink the data enough to the nationwide mean: we can tell this because, in Figure 1d, the actual election outcome is higher than predicted for low-predicted values, and lower than predicted for high-predicted values. *Undershrinkage* means that the estimated parameters $\hat{\tau}_i$ are probably *higher* than their true values, which could be caused by a pattern of nonignorable nonresponse that varies between states so that observed variability in the state proportions is caused by varying nonresponse patterns as well as actual variation in average opinions (see Little and Gelman 1996, for a discussion of this example and Krieger and Pfeffermann 1992, for a more general treatment). The undershrinkage could be quantified by comparing the estimated to the optimal level of shrinkage, but this comparison can only be made after the true values are observed.

It is also possible to compare the models by fitting each separately to each survey and examining the stability of estimates over a short period of time. This would be a more reasonable way to study the models in the common situation that the true population means never become known. Figure 3 displays, for each of our seven surveys, the estimates from raking and from the hierarchical model. (When modeling the surveys individually, we fit a common hierarchical variance for all 48 states because there was not enough data to obtain reliable maximum likelihood estimates for the four regions separately from the data in each poll.) Results are shown for the entire United States and for three representative states: California (a large state), Washington (mid-sized), and Nevada (small). For convenience, the plot also shows the estimates based on the seven surveys pooled and the actual election outcomes. For all the individual states, the hierarchical estimate is less variable over time than the raking estimate. The pattern is clearest in Nevada, where the sample size for the individual surveys was so low that the raking estimate degenerated to 0 or 1 in most cases, but the better performance of the hierarchical model is clear in the other states as well. For

Table 1

By state: election results (proportion of the two-party vote in 1988 received by Bush); survey data (unweighted mean and sample size) from the combined surveys; raking estimate using CBS variables; and posterior median (and interquartile range; that is, width of the central 50% uncertainty interval) of poststratified estimates based on state effects unsmoothed, set to zero, and fit by a hierarchical model.

Estimates are labelled 1, 2, 3, 4 corresponding to the descriptions in Section 3.3.

State	Election result	Sample size	Unweighted mean	Poststratification estimates (and IQRs)			
				1: Raking estimate	2: State effects unsmoothed	3: State effects set to 0	4: Hierarchical model
AL	0.60	134	0.72	0.67	0.63 (0.05)	0.56 (0.01)	0.62 (0.05)
AR	0.57	86	0.57	0.53	0.53 (0.06)	0.60 (0.01)	0.55 (0.06)
AZ	0.61	141	0.62	0.61	0.62 (0.05)	0.56 (0.02)	0.61 (0.05)
CA	0.52	1075	0.57	0.53	0.55 (0.02)	0.53 (0.01)	0.55 (0.02)
CO	0.54	126	0.59	0.59	0.58 (0.06)	0.57 (0.01)	0.57 (0.05)
CT	0.53	103	0.53	0.55	0.52 (0.06)	0.49 (0.02)	0.51 (0.06)
DE	0.56	30	0.40	0.37	0.42 (0.11)	0.60 (0.01)	0.52 (0.08)
FL	0.61	553	0.64	0.62	0.61 (0.03)	0.62 (0.01)	0.61 (0.03)
GA	0.60	211	0.62	0.58	0.56 (0.04)	0.56 (0.01)	0.56 (0.04)
IA	0.45	102	0.38	0.38	0.38 (0.06)	0.59 (0.01)	0.41 (0.06)
ID	0.63	31	0.52	0.58	0.52 (0.12)	0.59 (0.02)	0.55 (0.08)
IL	0.51	429	0.55	0.52	0.53 (0.03)	0.52 (0.01)	0.52 (0.03)
IN	0.60	215	0.75	0.73	0.74 (0.04)	0.56 (0.01)	0.72 (0.04)
KS	0.57	105	0.72	0.71	0.71 (0.06)	0.57 (0.01)	0.68 (0.05)
KY	0.56	146	0.57	0.53	0.56 (0.05)	0.64 (0.01)	0.57 (0.05)
LA	0.55	153	0.62	0.60	0.61 (0.05)	0.54 (0.01)	0.59 (0.04)
MA	0.46	277	0.47	0.41	0.46 (0.04)	0.50 (0.02)	0.47 (0.04)
MD	0.51	207	0.52	0.50	0.49 (0.04)	0.56 (0.01)	0.50 (0.04)
ME	0.56	44	0.52	0.52	0.55 (0.10)	0.52 (0.02)	0.54 (0.08)
MI	0.54	399	0.58	0.55	0.57 (0.03)	0.54 (0.01)	0.57 (0.03)
MN	0.46	210	0.54	0.53	0.53 (0.05)	0.59 (0.01)	0.53 (0.04)
MO	0.52	235	0.46	0.43	0.46 (0.04)	0.55 (0.01)	0.47 (0.04)
MS	0.61	170	0.69	0.70	0.65 (0.04)	0.53 (0.01)	0.63 (0.04)
MT	0.53	31	0.39	0.40	0.40 (0.12)	0.58 (0.02)	0.50 (0.09)
NC	0.58	239	0.59	0.60	0.55 (0.04)	0.58 (0.01)	0.55 (0.04)
ND	0.57	54	0.56	0.56	0.55 (0.09)	0.58 (0.01)	0.56 (0.08)
NE	0.61	90	0.58	0.60	0.56 (0.07)	0.58 (0.01)	0.56 (0.06)
NH	0.63	20	0.70	0.68	0.73 (0.13)	0.53 (0.02)	0.61 (0.10)
NJ	0.57	301	0.57	0.60	0.53 (0.04)	0.46 (0.01)	0.53 (0.03)
NM	0.53	87	0.55	0.54	0.57 (0.07)	0.54 (0.02)	0.56 (0.06)
NV	0.61	19	0.68	0.80	0.67 (0.13)	0.56 (0.02)	0.60 (0.09)
NY	0.48	639	0.42	0.37	0.41 (0.03)	0.45 (0.01)	0.41 (0.02)
OH	0.55	454	0.62	0.63	0.58 (0.03)	0.55 (0.01)	0.58 (0.03)
OK	0.58	93	0.57	0.62	0.59 (0.07)	0.63 (0.01)	0.60 (0.06)
OR	0.48	111	0.50	0.47	0.50 (0.06)	0.58 (0.02)	0.52 (0.06)
PA	0.51	431	0.54	0.54	0.52 (0.03)	0.48 (0.02)	0.52 (0.03)
RI	0.44	65	0.28	0.29	0.27 (0.07)	0.50 (0.02)	0.34 (0.06)
SC	0.62	151	0.70	0.67	0.66 (0.05)	0.55 (0.01)	0.64 (0.04)
SD	0.53	52	0.54	0.51	0.53 (0.09)	0.58 (0.01)	0.54 (0.08)
TN	0.58	252	0.68	0.69	0.66 (0.04)	0.60 (0.01)	0.65 (0.03)
TX	0.56	594	0.58	0.52	0.56 (0.03)	0.60 (0.01)	0.56 (0.02)
UT	0.67	61	0.80	0.85	0.79 (0.07)	0.60 (0.02)	0.72 (0.06)
VA	0.60	255	0.69	0.72	0.67 (0.04)	0.59 (0.01)	0.66 (0.03)
VT	0.52	12	0.54	0.58	0.60 (0.19)	0.53 (0.02)	0.55 (0.11)
WA	0.49	269	0.47	0.41	0.46 (0.04)	0.58 (0.01)	0.48 (0.04)
WI	0.48	264	0.49	0.53	0.48 (0.04)	0.57 (0.01)	0.49 (0.04)
WV	0.48	79	0.48	0.52	0.48 (0.07)	0.65 (0.01)	0.53 (0.06)
WY	0.61	13	0.50	0.36	0.59 (0.17)	0.59 (0.02)	0.59 (0.10)

Table 2

Summary statistics for raw mean of responses, raking estimate, and three poststratified estimates from the combined surveys. Summaries given are the estimated mean of the 48 state vote proportions weighted by state voter turnout (thus, estimated national popular vote proportion for Bush excluding Alaska, Hawaii, and the District of Columbia); the mean absolute error of the 48 state estimates; the average width of the 50% intervals for the states; and the number of the 48 states whose true values fall within the 50% intervals.

Summary	Actual result	Unweighted mean	Raking estimate	State effects unsmoothed	State effects set to 0	Hierarchical model
Mean of national popular vote	0.539	0.568	0.549	0.548	0.547	0.550
Mean absolute error of states	-	0.056	0.066	0.049	0.048	0.035
Average width of 50% intervals	-	-	-	(0.069)	(0.016)	(0.057)
Number of states contained in 50% interval	-	-	-	18	3	20

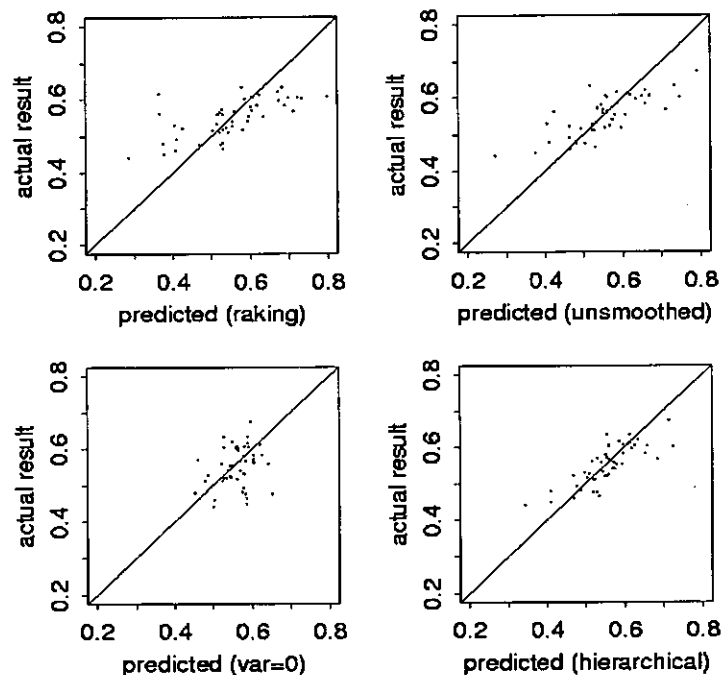


Figure 1. Election result by state vs. posterior median estimate for (a) raking on demographics, (b) regression model including state indicators with no hierarchical model, (c) regression model setting state effects to zero, (d) regression model with hierarchical model for state effects.

example, it was not reasonable to assign Bush only 46% of the support in California (in the poll 3 days before the election) or only 30% of the support in the state of Washington. For the United States as a whole, however, the two estimates are quite similar (in fact, when all seven polls are combined, the raking estimate performs very slightly better), indicating once again that the benefits from the modelling approach appear when studying subsets of the population.

The results for Washington have the surprising property that the regression estimate based on the combined surveys (shown at time “-1” on the graph) is lower than the seven estimates from the original surveys. This occurs because the data from the combined surveys show that the state of Washington supports Bush less than would be predicted merely by controlling for the demographic covariates (that prediction would be the estimate for Washington from the model with state effects set to zero, which from Table 1 is

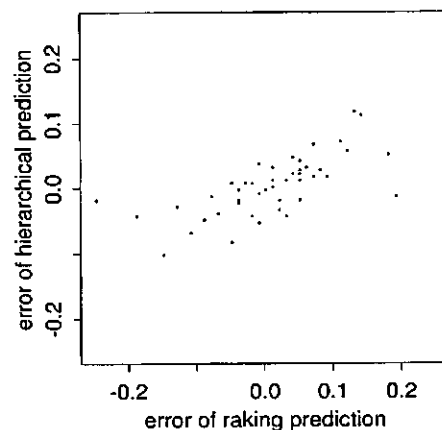


Figure 2. Scatterplot of prediction errors by state for the hierarchical model vs. the raking estimate. The errors of the hierarchical model are lower for most states.

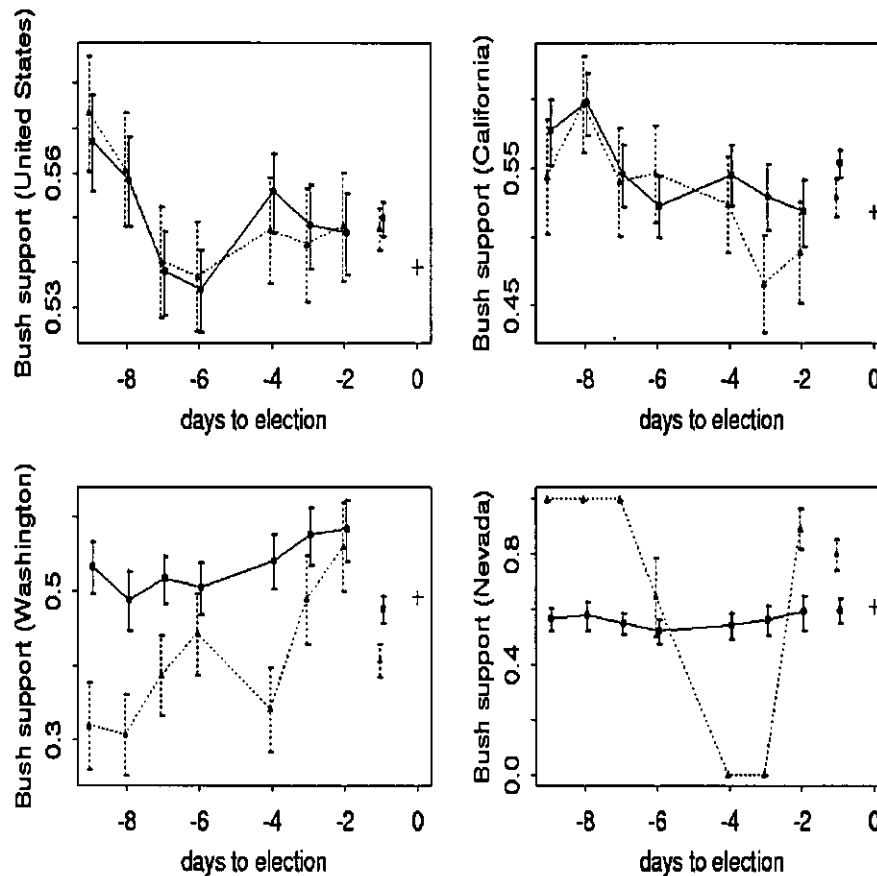


Figure 3. Estimated Bush support estimated separately from seven individual polls taken shortly before the election for (a) the entire U.S. (excluding Alaska, Hawaii, and the District of Columbia), (b) a large state (California), (c) a medium-sized state (Washington), and (d) a small state (Nevada). Each plot shows the raking estimates as a dotted line and the estimates from hierarchical model as a solid line, with error bars indicating 50% confidence bounds for the raking and 50% posterior intervals for the model-based estimates. The polls were taken between nine and two days before the election. Estimates based on the combined surveys are shown at time “-1”, and the actual election result is shown at time “0” on each plot.

0.58). But none of the individual surveys, taken alone, had enough data to make a convincing case that Washington was so far from the national mean, and so the Bayes estimate shrunk their estimates to a greater extent. This behavior, while it may seem strange at first, is in fact appropriate: with a smaller survey, there is less information about the individual poststratification categories, and the model-based estimate produces an estimate for each category that is closer to the sample mean. When all seven surveys are combined, more information is available, and the model relies more strongly on the data in each category. This is how the Bayes procedure essentially balances the concerns of poststratifying on too few or too many categories.

4. DISCUSSION

Poststratification is the standard method of correcting for unequal probabilities of selection and for nonresponse in sample surveys. From the modelling perspective, raking or poststratification on a set of covariates is closely related to

a regression model of responses conditional on those covariates, with population quantities estimated by summing over the known distribution of covariates in the population. Conditioning on more fully-observed covariates allows one to include more information in forming population estimates, but it is well known that raking on too large a set of covariates yields unacceptably variable inferences. We propose a method of poststratification on a large set of variables while fitting the resulting regression with a hierarchical model, thus harnessing the well-known strengths of Bayesian inference for models with large numbers of exchangeable parameters.

The Bayesian poststratification is most useful for estimation in subsets of the population (e.g., individual states in the U.S. polls) for which sample sizes are small. A related area in which modeling should be effective is in combining surveys conducted by different organizations, modeling conditional on all variables that might affect nonresponse in either survey. In addition, the methods in this paper can obviously be applied to continuous responses by replacing logistic regressions by other generalized linear models.

Our purpose in Bayesian modeling is not to fit a subjectively “true” model to the data or the underlying responses, but rather to estimate with reasonable accuracy the average response conditional on a large set of fully-observed covariates. More accurate models of the responses should allow more accurate inferences – but even the simple exchangeable mixed effects model we have fit, with hyperparameters estimated from the data, should perform better than the extremes of the fixed effects model or setting coefficients to zero. Ultimately, the goal of probability modeling and Bayesian inference in a sample survey context is to allow one to make use of abundant poststratification information (e.g., census data classified by sex, ethnicity, age, education, and state) to adjust a relatively small sample survey.

Difficulties with modeling approaches such as ours could arise in several ways. If one adjusts to a large number of categories using too weak a model (such as the model with unsmoothed state effects), the resulting estimates can be too variable. If the population distributions of the variables used in the poststratification are not available (for example, adjusting to a variable that is not measured or is measured inaccurately by the Census), then the N_j 's must be modeled also, which requires additional work. Of course, such additional work would be required to rake on these variables as well. Since all of the methods, including raking and regression methods, assume ignorable models, they will yield incorrect inferences when unmeasured variables affect nonresponse and are correlated with the outcome of interest.

The methods described here are intended as an improvement upon raking-type poststratification adjustments and are not intended to, by themselves, correct for nonignorable nonresponse. However, by allowing one to adjust for more variables, the Bayesian poststratification should allow the use of models for which the ignorability assumption is more reasonable. Having a large number of poststratification categories (e.g., in 48 states) creates problems with classical weighting methods because many categories will have few or even no respondents. Interestingly, however, having many categories can make Bayesian modeling more reliable: more categories means more random effects in the regression, which can make it easier to estimate variance components.

ACKNOWLEDGEMENTS

We thank Xiao-Li Meng and several reviewers for helpful comments and the National Science Foundation for grant DMS-9404305 and Young Investigator Award DMS-9457824.

APPENDIX: COMPUTATION

We use an EM-type algorithm to estimate the hyperparameters τ_i ; given these, we sample from the posterior distribution of the coefficients β using a normal approxi-

mation to the logistic regression likelihood. We use this approximation for its simplicity and because it is reasonable for fairly large surveys, as in our application in Section application; if desired, more exact computations can be performed using the Gibbs sampler and Metropolis algorithm (see Clayton 1996), perhaps using the algorithm described here as a starting point.

When the data distribution is normal and the means are linear in the regression coefficients, the EM algorithm can be used to obtain estimates of the variance components (Dempster, Laird, and Rubin 1977), treating the vector of coefficients β as “missing data.” In this framework, the “complete-data” loglikelihood for τ_i is

$$L(\tau_i | \gamma_i) = \text{const} - K_i \log \tau_i - \frac{1}{2\tau_i^2} \sum_{k=1}^{K_i} \gamma_{ki}^2,$$

so the sufficient statistic for τ_i is $t(\gamma_i) = \sum_{k=1}^{K_i} \gamma_{ki}^2$. Given the current estimate τ^{old} , the expected sufficient statistic is

$$E(t(\gamma_i) | y, \tau^{\text{old}}) =$$

$$\| E(\gamma_i | y, \tau^{\text{old}}) \|^2 + \text{trace}(\text{var}(\gamma_i | y, \tau^{\text{old}})).$$

Since these two terms are not analytically tractable for our model, we use the following approximations which are easily obtained: (1) approximate $E(\gamma_i | y, \tau^{\text{old}})$ with an estimate $\hat{\gamma}_i$, based on y and the estimate τ^{old} , and (2) approximate $\text{var}(\gamma_i | y, \tau^{\text{old}})$ from the curvature of the log-likelihood at the estimate, $\hat{V}_{\gamma_i} = (-L''(\hat{\gamma}_i))^{-1}$. We update these approximations iteratively for all $i = 1, \dots, L$ simultaneously, converging to an approximate maximum likelihood estimate $(\hat{\tau}_1, \dots, \hat{\tau}_L)$. Given an initial guess τ^{old} , the algorithm proceeds by iterating the following two steps to convergence.

Approximate E-step. Solve the likelihood equations iteratively, as described below. Use the estimate $\hat{\beta}$ to obtain an approximation to $E(t(\gamma_i) | y, \tau^{\text{old}})$, for each $i = 1, \dots, L$.

We solve the likelihood equations $d/d\beta L(\beta | y, \tau) = 0$ using iteratively weighted least squares, involving a normal approximation to the likelihood $p(y | \beta) = \prod_i p(y_i | \beta)$, based on locally approximating the logistic regression model by a linear regression model (see Gelman *et al.* 1995, p. 391). Let $\eta_i = (Z\beta)_i$ be the linear predictor for the i -th observation. Starting with the current guess of $\hat{\beta}$, let $\hat{\eta} = Z\hat{\beta}$. Then a Taylor series expansion to $L(y_i | \eta_i)$ gives $z_i \approx N(\eta_i, \sigma_i^2)$, where

$$z_i = \hat{\eta}_i + \frac{(1 + \exp(\hat{\eta}_i))^2}{\exp(\hat{\eta}_i)} \left(y_i - \frac{\exp(\hat{\eta}_i)}{1 + \exp(\hat{\eta}_i)} \right)$$

$$\sigma_i^2 = \frac{(1 + \exp(\hat{\eta}_i))^2}{\exp(\hat{\eta}_i)}.$$

Let $\hat{\Sigma}_{\beta}$ denote the value of Σ_{β} based on plugging in the current estimate $\hat{\tau}$, and let $\hat{\Sigma}_z = \text{diag}(\sigma_i^2)$. Then we obtain an updated estimate and variance matrix using weighted

least squares based on the normal prior distribution and the normal approximation to the logistic regression likelihood:

$$\hat{\beta} = (Z' \sum_z^{-1} Z + \sum_{\beta}^{-1})^{-1} Z' \sum_z^{-1} z \quad (2)$$

$$\hat{V}_{\beta} = (Z' \sum_z^{-1} Z + \sum_{\beta}^{-1})^{-1}. \quad (3)$$

We iterate until convergence and then use $\hat{\beta}$ and the appropriate elements of \hat{V}_{β} to estimate $\text{var}(\gamma_l | y, \tau^{\text{old}})$.

M-step. Maximize over the parameters τ_l to obtain $\tau_l^{\text{new}} = (\hat{E}(t(\gamma_l) | y, \tau^{\text{old}}) / K_l)^{1/2}$, for each $l = 1, \dots, L$. Set τ^{old} to τ^{new} and return to the approximate E-step.

Once the approximate EM algorithm has converged to an estimate $\hat{\tau}$, we draw β from a normal approximation to the conditional posterior distribution $p(\beta | y, \hat{\tau})$, using the values from equations (2) and (3) at the last EM step as the mean and variance matrix in the normal approximation. For each draw of the vector parameter β , we compute the category means, $\pi = \text{logit}^{-1}(X\beta)$, and any population totals of interest, counting each category j as N_j units in the population.

REFERENCES

- BELIN, T.R., DIFFENDAL, G.J., MACK, S., RUBIN, D.B., SCHAFER, J.L., and ZASLAVSKY, A.M. (1993). Hierarchical logistic regression models for imputation of unresolved enumeration status in undercount estimation (with discussion). *Journal of the American Statistical Association*, 88, 1149-1166.
- CLAYTON, D.G. (1996). Generalized linear mixed models. In *Practical Markov Chain Monte Carlo*. (Eds. W. Gilks, S. Richardson, and D. Spiegelhalter), 275-301. New York: Chapman & Hall.
- DEMING, W., and STEPHAN, F. (1940). On a least squares adjustment of a sampled frequency table when the expected marginal tables are known. *Annals of Mathematical Statistics*, 11, 427-444.
- DEMPSTER, A.P., LAIRD, N.M., and RUBIN, D.B. (1977). Maximum likelihood from incomplete data via the EM algorithm (with discussion). *Journal of the Royal Statistical Society B*, 39, 1-38.
- GELMAN, A., CARLIN, J.B., STERN, H.S., and RUBIN, D.B. (1995). *Bayesian Data Analysis*. London: Chapman and Hall.
- GELMAN, A., and KING, G. (1993). Why are American presidential election campaign polls so variable when votes are so predictable? *British Journal of Political Science*, 23, 409-451.
- HOLT, D., and SMITH, T.M.F. (1979). Post stratification. *Journal of the Royal Statistical Society*, 142, 33-46.
- KRIEGER, A.M., and PFEFFERMANN, D. (1992). Maximum likelihood estimation from complex sample surveys. *Survey Methodology*, 18, 225-239.
- LAZZERONI, L.C., and LITTLE, R.J.A. (1997). Random-effects models for smoothing post-stratification weights. *Journal of Official Statistics*, to appear.
- LITTLE, R.J.A. (1991). Inference with survey weights. *Journal of Official Statistics*, 7, 405-424.
- LITTLE, R.J.A. (1993). Post-stratification: a modeler's perspective. *Journal of the American Statistical Association*, 88, 1001-1012.
- LITTLE, T.C. (1996). Models for nonresponse adjustment in sample surveys. Ph.D. thesis, Department of Statistics, University of California, Berkeley.
- LITTLE, T.C., and GELMAN, A. (1996). A model for differential nonresponse in sample surveys. Technical report.
- LONGFORD, N.T. (1996). Small-area estimation using adjustment by covariates. *Quèstió*, 20, to appear.
- NORDBERG, L. (1989). Generalized linear modeling of sample survey data. *Journal of Official Statistics*, 5, 223-239.
- RUBIN, D.B. (1976). Inference and missing data. *Biometrika*, 63, 581-592.
- VOSS, D.S., GELMAN, A., and KING, G. (1995). Pre-election survey methodology: details from nine polling organizations, 1988 and 1992. *Public Opinion Quarterly*, 59, 98-132.

Estimating the Population and Characteristics of Health Facilities and Client Populations Using a Linked Multi-Stage Sample Survey Design

K.K. SINGH, A.O. TSUI, C.M. SUCHINDRAN and G. NARAYANA¹

ABSTRACT

This paper demonstrates the utility of a multi-stage sample survey design that obtains a total count of health facilities and of the potential client population in an area. The design has been used for a state-level survey conducted in mid-1995 in Uttar Pradesh, India. The design involves a multi-stage, areal cluster sample, wherein the primary sampling unit is either an urban block or rural village. All health service delivery points, either self-standing facilities or distribution agents, in or formally assigned to the primary sampling unit are mapped, listed, and selected. A systematic sample of households is selected, and all resident females meeting predetermined eligibility criteria are interviewed. Sample weights for facilities and individuals are applied. For facilities, the weights are adjusted for multiplicity of secondary sampling units served by selected facilities. For individuals, the weights are adjusted for survey response levels. The survey estimate of the total number of government facilities compares well against the total published counts. Similarly the female client population estimated in the survey compares well with the total enumerated in the 1991 census.

KEY WORDS: Sample survey; Program evaluation; Health services; Developing country.

1. INTRODUCTION

The evaluation of the impact of health programs on population-level health outcomes often requires knowledge of the number and characteristics of facilities and potential clients. Such information is frequently lacking in developing countries where program record keeping and vital registration systems tend to be incomplete and poorly maintained.

To obtain current information on health status, health service use, service performance, and client needs, programs have resorted to occasional sample surveys, often designed and conducted independently and subareally (Aday 1991; Ross and McNamara 1983). Some demographic and health surveys (Macro International 1996), however, do provide a national profile of population-level health outcomes, such as fertility, child mortality, and nutritional well-being. The distinct advantage of a national population sample for planning health programs is its ability to measure the attitudes and behaviors of clients as well as non-clients. Program service statistics are limited to actual clients and may not yield the most current or accurate picture of service use.

In addition to client behaviours, it is useful to monitor the accessibility and quality of services, but this requires a separate review of service provision at health facilities or related outlets. Efforts in developing countries, like the situation analysis studies (Miller, Ndhlovu, Gachara and Fisher 1991), involve probability surveys of health facilities

and can provide a national overview of program performance. However, often they are restricted to reviewing public health programs because of incomplete registration of private health providers, such as private clinics or pharmacies. The lack of complete and accurate registration of private-sector service providers prevents probability sample surveys from being used to monitor health care patterns through this sector.

Constraints on available resources to expand and improve the delivery of health care in developing, as well as developed, countries are increasing. This suggests that a more efficient use of resources available for monitoring and evaluation, particularly through surveys, is a consideration for all concerned. Innovative approaches to sample surveys should be developed to provide health planners and managers with a maximum of information at a minimum of precision loss.

We present results from a multi-stage, cluster sample survey designed to estimate the population and characteristics of health facilities and target client populations. The cluster sample for the survey, conducted in the large northern Indian state of Uttar Pradesh, is used as a basis for selecting health facilities and households, with subsequent selection of service staff from the facilities and of married women of childbearing age from the households. The survey was designed to generate independent samples of health facilities, staff, households, and client populations for the health services.

The next section of this paper will describe the survey design, its contents, and fieldwork procedures as applied in

¹ Kaushalendra K. Singh, Carolina Population Center, University of North Carolina at Chapel Hill, CB #8120 University Square, Chapel Hill, NC 27516-3997 and Department of Statistics, Faculty of Science, Banaras Hindu University, Varanasi 221005 India; Amy O. Tsui, Director, Carolina Population Center, University of North Carolina at Chapel Hill, CB #8120 University Square, Chapel Hill, NC 27516-3997 and Department of Maternal and Child Health, School of Public Health, University of North Carolina at Chapel Hill, CB #7400 Rosenau Hall, Chapel Hill, NC 27599-7400; Chirayath M. Suchindran, Carolina Population Center, University of North Carolina at Chapel Hill, CB #8120 University Square, Chapel Hill, NC 27516-3997 and Department of Biostatistics, School of Public Health, University of North Carolina at Chapel Hill, CB #7400 Rosenau Hall, Chapel Hill, NC 27599-7400; Gaade Narayana, The Futures Group International, 1050 17th Street, N.W., Suite 1000, Washington, DC 20036.

Uttar Pradesh. The following section presents the comparative results on health facilities and population, and the last section will discuss lessons learned for survey design from the Uttar Pradesh application. These lessons will be important specifically for this survey's planned replication in two years but generally informative for other countries that may adopt the linked design.

2. THE PERFORM SURVEY IN UTTAR PRADESH

The PERFORM (Project Evaluation Review For Organizational Resource Management) Survey was designed to measure benchmark indicators for a large family planning project called the Innovations in Family Planning Services (IFPS) project sited in Uttar Pradesh and co-funded by the Government of India and the U.S. Agency for International Development. Uttar Pradesh has a population of over 140 million and by itself would rank as the fifth largest developing country.

2.1 Content

Indicator estimates for IFPS are needed at three levels: (1) public and private service delivery points (SDPs), (2) service providers staffing the SDPs or facilities, and (3) client population, represented by women of reproductive age. As IFPS seeks to improve the family planning service environment, it is imperative to obtain measures of indicators at this level but in such a way as to be relatable to the women resident in those environments.

As a result, the PERFORM survey developed seven questionnaires:

- 1-2) An urban block and village questionnaire to inventory all potential and actual providers of health services in the sampled village or urban block;
- 3) A fixed service delivery point (FSDP) questionnaire to gather information on the staff, services, equipment, supplies, and education and motivation activities at sampled public and private facilities.
- 4) A staff questionnaire administered to all FSDP staff involved in family planning services (identified from the FSDP questionnaire) to assess their capabilities and service experiences;
- 5) An individual service agent (ISA) questionnaire to all individuals working outside of self-standing facilities (FSDPs) who currently or potentially can provide health planning services, such as private doctors, pharmacists, midwives, lay health workers, and retailers;
- 6) A household questionnaire to be administered to heads of the sampled households to enumerate household members and selected demographic and social characteristics;
- 7) An individual questionnaire for currently married women between the ages of 13 to 49 (identified from the household questionnaire) to collect information on knowledge of and past, current, and intended use of

health services, recent pregnancy and contraceptive behaviors, and additional background characteristics.

2.2 Sampling Design

PERFORM was designed to provide estimates of facility and population characteristics at the state, regional, divisional, and district levels. The district was important since it was the focal point for introducing innovative approaches and additional IFPS inputs. At the time of the survey design, Uttar Pradesh had 14 administrative divisions; two districts were selected from each using probability proportional to size (PPS) procedures. These areal units have administrative-political boundaries and thus public administration utility. The districts were also aggregated into five regional groupings.

In each district, the total number of households to be sampled was fixed at 1,500. A sample of 1,500 households per district was determined to be sufficient to provide estimates for the main population level indicators. An overall target sample size of 1,627 ever-married women aged 13-49 was required to detect a change of 5 per cent point in contraceptive prevalence (with $\alpha = 0.05$ and $1 - \beta = 0.90$) at district level. It is expected that the number of ever-married women aged 13-49 per household would be 1.15 and therefore, by visiting a sample of 1,415 households the required number of ever-married women would be obtained. Allowing for an increase of 5 per cent to accommodate non-response and non-availability, a target sample of 1,725 ever-married women aged 13-49 from the 1,500 households was considered to be sufficient. The schematic diagram of the sample design is given in Figure 1.

The districts were further stratified into rural and urban areas. According to the Census of India, all places with a municipality, a municipal corporation, a cantonment board, a notified area committee, or all other places with a minimum of 5,000 population, with at least 75 percent of the male working population engaged in non-agricultural pursuits and a population density of at least 400 persons per square kilometer, are classified as urban areas. Urban blocks and rural villages served as the secondary sampling units (SSUs). The 1,500 households to be sampled from each district were allocated to the rural and urban areas in proportion to the size of population within the district. However, if the allocated proportion of urban population was less than 20 percent, the allocation of households in the urban area was fixed at 20 percent. This allocation was prescribed to ensure coverage of a sufficient number of health delivery points.

Households within rural areas were selected using a stratified two-stage sampling plan. The villages in the rural areas were first stratified into four strata depending on the size of the of the population as follows:

Stratum	Population size of the village
I	100 - 499
II	500 - 1,999
III	2,000 - 4,999
IV	5,000 and above.

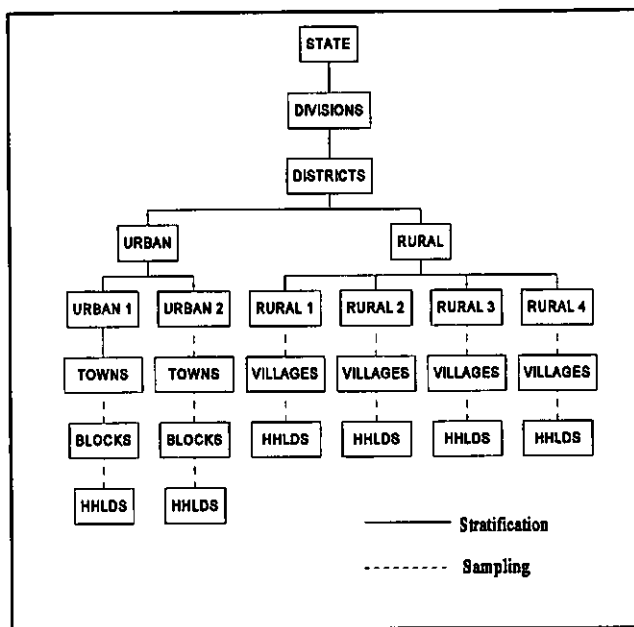


Figure 1. Schematic Diagram of PERFORM Sample Design

Villages with fewer than 100 residents or 20 households were excluded from the list (such villages were rare in the present study). The number of villages to be selected from each district was allocated proportionally to each of the four strata. Villages were selected by first arranging them within the stratum by the female literacy rates and then selecting the required number of villages by a PPS sampling procedure. All households in the selected villages were listed and mapped, and a target number of 20 households was drawn from each selected village using systematic sampling. Villages with more than 500 households or with a population size of 2,500 or more (some in stratum III and all in stratum IV) were segmented into four parts, and two segments were selected for household listing and selection. The required 20 households were selected taking ten households from each segment using systematic random sampling.

Households in urban areas were also selected using a stratified two-stage sampling plan. The towns in the urban areas of a district were stratified into two strata according to population size as follows:

Stratum	Population size of the town
I	100,000 and more
II	Fewer than 100,000.

All towns within stratum I were selected with certainty. Towns in stratum II were arranged according to population size and the required number of towns were selected by PPS. From each sampled town a minimum of two blocks were selected using PPS methods. All households in the selected blocks were listed and mapped, and 15 households were selected from each urban block using systematic random sampling.

2.2.1 District Selection Probability

Let m_k denote the population of the k -th district within a division. Because two districts must be selected from each division, the probability of selecting the k -th district from a division r_k is obtained as

$$r_k = 2 * \frac{m_k}{M}$$

where M is the total population of the division ($M = \sum_{k=1}^t m_k$) and t is the total number of districts in the division.

2.2.2 Village and Household Selection Probability

Let n_{ijk} denote the number of households in the i -th village, j -th stratum and k -th district. Then, p_{ijk} , the probability of selecting village i from the j -th stratum and k -th district is obtained as,

$$p_{ijk} = a_{jk} * \frac{n_{ijk}}{N_{jk}} * r_k$$

where a_{jk} and N_{jk} are, respectively, the number of villages selected and the total number of households in the j -th stratum and k -th district.

Let q_{ijk} be the probability of selecting a household from the rural areas of a selected district. Then q_{ijk} may be given as

$$q_{ijk} = p_{ijk} * \frac{20}{n_{ijk}}$$

where 20 is the number of households drawn from the selected village.

The weights for villages and households are then the inverse of their selection probabilities, i.e., $1/p_{ijk}$ and $1/q_{ijk}$, and are denoted as VW_{ijk} and HW_{ijk} respectively.

2.2.3 Town, Urban Block and Household Selection Probability

The probability of selecting the j -th town from the k -th district, t_{jk} , is obtained as

$$t_{jk} = 1 \quad \text{if the population of the town is } > 100,000$$

$$t_{jk} = c_k \frac{s_{jk}}{S_k} \quad \text{if the population of the town is } < 100,000$$

where s_{jk} is the total number of households in the j -th town (with a population $< 100,000$) in the k -th district, c_k is the number of towns selected in district k , and S_k is the total number of households in towns with less than 100,000 population in district k .

Let u_{ijk} denote the probability of selecting the i -th urban block from the j -th town and k -th district. Then u_{ijk} is obtained as

$$u_{ijk} = b_{jk} * \frac{x_{ijk}}{Y_{jk}} * t_{jk} * r_k$$

where b_{jk} is the number of urban blocks selected and Y_{jk} is the total number of households in the j -th town and k -th district, and x_{ijk} is the number of households in the i -th block, j -th town and k -th district.

The probability of selecting a household from the i -th urban block and the k -th district, denoted as v_{ijk} , is given as,

$$v_{ijk} = u_{ijk} * \frac{15}{x_{ijk}}$$

where 15 is the number of households drawn from the selected urban block.

The weights for urban blocks and households are then the inverse of their selection probabilities, i.e., $1/u_{ijk}$ and $1/v_{ijk}$, and are denoted as UW_{ijk} and HW_{ijk} respectively. Since the population-level estimates are based on individuals, all individuals in a selected household received the household weight. No selection procedure was used for eligible respondents within a household.

2.2.4 Adjustment for Household Questionnaire for Non-response and Over-sampling of Urban Blocks

The adjustment of the household weight for non-response is done under the assumption of random non-response within the village (or urban block) and is carried out as follows:

Let n_1 be the number of households selected and n_2 be the number of households where interviews are completed. Then the adjusted weight for households due to non-response is defined as

$$HW_{2ijk} = HW_{1ijk} * \frac{n_1}{n_2}$$

The final household weight also includes an adjustment of proportion of urban population in the district, where an over-sampling of urban blocks has occurred (districts with less than 20 percent of urban population).

Let n_3 be the actual proportion of urban population in a district and n_4 the proportion of urban population in the sample. Then the adjusted weight for households due to non-response and over-sampling of urban blocks is defined as

$$HW_{3ijk} = HW_{2ijk} * \frac{n_3}{n_4}$$

2.2.5 Selection of Service Delivery Points in Sample Districts

To obtain a probability sample of service delivery points, FSDPs and ISAs were selected in relation to the SSUs, i.e., the villages or urban blocks, as follows:

- 1) All private and public sector health institutions in selected rural and urban SSUs;
- 2) All sub-centres, primary health centres, community health centres, post-partum centers providing services to the population in the selected rural SSUs;

- 3) All private hospitals with 10 or more beds in the nearest town (with fewer than 100,000 population) within 30 kms of selected rural SSUs;
- 4) All municipal hospitals, district hospitals, and medical college hospitals;
- 5) All clinics and hospitals runs by voluntary agencies, the organized sector, and cooperatives; and
- 6) All ISAs in selected villages and urban blocks.

It is probably helpful first to describe the organized delivery of health care through the government sector. Residents of all villages are entitled to obtain health care from a government sub-centre (SC), a primary health centre (PHC), and a community health centre (CHC). Villages with 5,500 population or more often have an SC located within their boundaries. Approximately six SCs will report to one PHC, and PHCs in turn are linked to a CHC. At times the PHC is integrated with the CHC; as a result, our estimation must be of CHCs and PHCs combined, while SCs are estimated separately. (Population growth has led to the establishment of "additional PHCs" and redistricting of the original PHC catchment areas. These additional PHCs have been included in the estimation of the number of PHCs.) All SCs assigned to a sampled village were visited, as were their affiliated PHCs and CHCs.

At the time of listing and mapping households in each urban block and village, the FSDPs and ISAs were also listed and mapped. In addition, key informants in each SSU were interviewed regarding health outlets not visibly obvious. The selection of service delivery points – FSDPs and ISAs – within the SSU boundaries, or affiliated with the government's health subcentre, involved a full census. The one exception to this was for municipal hospitals, district hospitals and medical colleges, which were self-selected and thus had a weight of unity. The selection probabilities of the other FSDPs and ISAs are then a function of the probability of selecting the SSU, and the inverse of the latter serves as the weight of the FSDP or ISA unit. Weights for CHCs, PHCs, and SCs were calculated with the procedure below after determining some fieldwork "failure" in selecting these types of facilities correctly. (This failure is discussed later.)

Since CHCs and PHCs are associated with more than one SSU, we have assumed that one PHC exists per 30,000 population (which is approximately the actual average for Uttar Pradesh) and that one SC serves approximately 5,500 (actual district averages range from 4,000 to 6,500). Under this assumption, the CHC/PHC weight for each selected SSU is then

$$W_{CHC/PHC} = \frac{\text{Total population in selected SSU}}{30,000} * VW_{ijk} \text{ (or } UW_{ijk})$$

and the SC weight for each selected SSU is

$$W_{SC} = \frac{\text{Total population in selected SSU}}{5,500} * VW_{ijk} \text{ (or } UW_{ijk}).$$

All weights for FSDPs that were not self-selected had to be adjusted for multiplicity, *i.e.*, when an FSDP was selected into the sample on the basis of more than one SSU. For example, a CHC/PHC might be selected because of two sampled SSUs. In this case, the weight for the CHC/PHC was the sum of the weights of the two selected SSU, *i.e.*, $W_{CHC/PHC}$, associated with its selection.

2.3 Survey Implementation

Fieldwork for the PERFORM Survey was conducted from June to September 1995 in Uttar Pradesh. The survey was executed by four organizations contracted following a competitive procurement process. One organization that had tested the PERFORM survey design in one district a year earlier served as the nodal or coordinating organization. Master training to survey project coordinators and supervisors was provided, including a field pretest. The actual fieldwork for PERFORM was carried out in six-member teams composed of 1 male supervisor, 1 female editor, 1 male interviewer and 4 female interviewers. Each fieldwork organization on average engaged 3 teams to cover one district, or a total of 18 field staff for data collection per district (or 21 teams for a total of 126 field staff to cover 7 districts). Overall field supervision was the responsibility of a specially-appointed four-member team, one assigned to each consulting fieldwork organization. Following field editing, the questionnaires were transported to the home offices of the survey organizations for data entry and cleaning. One type of staff person, the auxiliary nurse-midwife who is stationed at a subcentre, was difficult to reach, even after the standard three attempts.

3. RESULTS

Table 1 gives the sample coverage for the PERFORM survey, in terms of the number of units selected of each type, the number successfully interviewed, and the completion rate. The completion rates are very high for ample units requiring personal contact – ranging from 94.3

for eligible women to 96.7 percent for households. Interview completion rates were 95 percent for facilities and agents. Only for fixed facility staff was the rate somewhat lower at 90 percent, a respectable although not an outstanding level. (One type of staff person, the auxiliary nurse-midwife who is stationed at a subcentre, was difficult to reach, even after the standard three attempts.)

3.1 Population Size and Characteristics

We compare first population-level measures on selected demographic indicators obtained from other sources with those from the PERFORM survey, as shown in Table 2. The figures indicate that PERFORM results compare favorably with census measures as well as these from the recent National Family Health Survey (NFHS) conducted in Uttar Pradesh in late 1992 and early 1993, with a sample size of 11,438 ever-married women aged 13 to 49. The enumerated population shows a growth of almost 10.5 million persons since the 1991 census, and the percentage of households in urban areas is close across all three sources. The ratio of women to men is slightly lower in PERFORM (891) than in the NFHS (917). The percentage of the population in the two age groups (0 to 14 and 65 and over) compares well, as does the percentage of households belonging to the scheduled castes. The percentage of households belonging to scheduled tribes is 3.1, higher than the 1.1 observed in the NFHS. This may reflect an actual growth in such households with increased in-migration to large towns and cities by scheduled tribe members. The proportions literate show small gains since the NFHS but compare well overall. The total fertility rate and the level of modern contraceptive use also are similar and change in a consistent direction between the dates of the two Uttar Pradesh surveys. Results in Table 2 suggest that PERFORM's sample design, based on traditional multistage cluster sample designs used for demographic surveys, was executed properly to produce state-level results comparable to the census and earlier NFHS survey. The standard error and design effect of the estimates were also given in the Table

Table 1
Coverage of Sample Units of PERFORM Survey: Uttar Pradesh, 1995

Sample Coverage	Sample Units						
	Villages	Urban Blocks	Households	Eligible Women	Fixed SDPs	FSDP Staff	Individual Agents
Number Sampled	1,539	738	42,006	48,009	2,549	7,026	23,364
Number Interviewed	1,539	738	40,633	45,277	2,428	6,320	22,335
Percent completed	100.00	100.00	96.7	94.3	95.3	89.9	95.6

Notes: Villages and urban blocks served as the primary sampling units; eligibility criteria for women were currently married and between ages 13 to 49 years; SDP = service delivery point.

Table 2
Basic Demographic Indicators for Uttar Pradesh, India

Index	Uttar Pradesh				
	Census (1991)	NFHS (1992-93)	PERFORM (1995)	Standard Error	Design Effect
Population	139,112,287	<i>u</i>	149,758,641	1,542,952	—
Percent urban	19.8	22.6 ^a	21.6 ^a	0.6553	12.6095
Sex ratio ^b	879	917	891	34.1010	0.9727
Percentage 0-14 years old	39.1	41.8	40.2	0.1306	1.9049
Percent 65+ years old	3.8	4.8	4.7	0.0513	1.5789
Percentage scheduled	21.0	18.0 ^a	20.0 ^a	0.3790	3.6536
Percentage scheduled tribe	0.2	1.1 ^a	3.1 ^a	0.1818	4.4694
Percent Literate ^c					
Male	55.7	65.3	67.6	0.3352	6.4634
Female	25.3	31.4	37.4	0.3824	8.6821
Total	41.6	49.9	53.3	0.3352	12.2385
Total fertility rate	5.1	4.8	4.5	—	—
Modern contraceptive	<i>u</i>	18.5 ^d	22.0 ^d	0.3499	3.4111

u = Unavailable.

^a Based on number of households

^b Number of females per thousand males

^c Based on population aged 7 and above for the census and population aged 6 and above for NFHS and PERFORM

^d Percentage of currently married women aged 15 to 49 using modern contraceptive method.

In Table 3 we compare the age and sex distributions for Uttar Pradesh obtained from the NFHS and PERFORM, as well as from the Sample Registration System, operated by the Office of the Registrar General. The sex ratios for the two surveys are also given. The age-sex distributions are again comparable across the three sources. However, there is a markedly lower sex ratio for the age group 30-49 years (820) in PERFORM and a slightly higher one for ages 50-64 (993) than those in the NFHS (941 and 960 respectively). We suspect some of this difference is due to a "push" of females out of the end of childbearing ages by field investigators of *one* survey organization to avoid completion of the pregnancy calendar and history portions of the questionnaire. (Upon further investigation, we found the sex ratios for women aged 50-64 to be uniformly higher in the seven districts under one organization's responsibility than those of others.) As a result, there are somewhat more women aged 50-64 enumerated in the PERFORM Survey than may actually be the case. This also may mean that births to women who were actually under age 50 were under-enumerated. Because this is not a high-fertility age group, the bias is not likely to be large.

3.2 Facility Size and Characteristics

By visiting and interviewing the facilities selected through the SSUs or cluster, we are able to generate an independent sample of health facilities and service providers. (These include those who currently, as well as potentially can, provide family planning services, *i.e.*, not all the estimated number of retail outlets (general merchant, kirana and pan shops) shown presently dispense contraceptives.) The weighted counts of these outlets is shown in Table 4. Our ability to validate the estimates of independent agents is weakened by the fact that many of them are not registered, particularly the "unqualified" (or quack) doctors. Narayana, Cross and Brown (1994: Table 8) report a 1991 total number of 112,568 villages in Uttar Pradesh, which would suggest almost one traditional birth attendant per village and 1 anganwadi worker for every 4.5 villages on average. These ratios appear reasonable given known circumstances regarding access to such types of care. The figures are quite close and provide evidence of the utility of the linked cluster sample design.

Table 3
Percent Distribution of the De Jure Population by Age and Sex, Based on SRS, NFHS, and PERFORM Sources for 1991-95

Age	SRS (1991)		NFHS (1992-93)			PERFORM (1995)		
	Male	Female	Male	Female	Sex Ratio	Male	Female	Sex Ratio
0-4	14.4	14.4	14.6	14.6	917	13.8	14	909
5-14	24.9	24.4	27.5	26.0	868	27.2	26.3	861
15-29	28.4	26.8	25.1	26.4	967	25.4	27.7	972
30-49	20.7	21.9	19.2	19.7	941	19.8	18.3	820
50-64	8.2	8.5	8.4	8.8	960	8.6	9.6	993
65+	3.6	4.0	5.2	4.4	718	5.2	4.1	702
Total	100.0	100.0	100.0	100.0		100.0	100.0	

Source for sample Registration System (SRS): Office of the Registrar India (1993a)

Source for NFHS: National Family Health Survey, Uttar Pradesh (1992-93)

Table 4
Total Number of Estimated Public and Private Sector Delivery Points by Type in Uttar Pradesh, India: 1995

Fixed service delivery points	Number	Individual service agents	Number
Total	31,400	Total	1,099,825
Hospitals		Physicians	
Government allopathic	968	Private resident allopathic	32,182
Government ISM	688	Private visiting allopathic	9,011
Municipal allopathic	57	Private resident (unqualified)	62,880
Municipal ISM	23	Private resident ISM	42,343
Private	5,212	Private visiting ISM	9,138
Private voluntary	130	Anganwadi workers	25,994
Private ISM	35	Village health workers	65,532
Industrial	61	Traditional birth attendants	110,546
Medical colleges	9	Medical shops	40,979
CHC/PHC/Additional PHC	3,948	General merchants	133,517
Subcentres	20,151	Kirana shops	376,679
Other	137	Pan shops	136,353
		Depot holders	5,818
		Other	48,855

3.3 Estimation Approaches

The estimated number of CHC/PHCs and SCs in Table 4 is based on the assumption that each such facility serves a fixed population size, *i.e.*, 30,000 and 5,500 respectively – the figures used by the government for planning health service delivery. The precision of the estimation would have been improved if the actual size of the local catchment population were known. In the absence of this information, we have used a constant population estimate for these two facility types.

Alternate estimation approaches were used prior to arriving at the above procedure. The first is illustrated in Table 5, which presents the actual and weighted counts of CHC/PHCs and SCs in each of the 28 survey districts. These figures are based on weighting the selected facilities by the SSU size only and without adjusting for multiplicity. The PERFORM sample selected in a total of 633 CHC/PHCs or 34.8 percent of the total (1818) and 1,267 subcenters or 13.3 percent to the total (9,491) in the 28 districts. These can be compared against the actual numbers

of CHC/PHCs and SCs in 1995 obtained from the Uttar Pradesh Department of Health and Family Welfare. It is evident that this weighting approach substantially overestimates the number of CHC/PHCs (3,472 compared to 1,818) but yields a nearly identical number of SCs (9,495 compared to 9,491). Using the villages and urban blocks as SSUs is reasonable as they are the public administration units (and population sizes) used to determine the location of subcenters.

They, however, do not offer an adequate stratification basis for the larger health facilities. Precision is lost because we weight with the inverse of the SSU's population and when CHC/PHCs are selected in for very small SSUs, the associated weight is disproportionately inflated. This results in a higher-than-actual count of such facilities, a situation most problematic in two districts – Allahabad and Sultanpur. If these two districts are eliminated, the over-estimation is 22.5 (± 0.8) percent instead of 91 percent. (Under-estimation of CHC/PHCs results where the reverse occurs, as in Bareilly district. Because of PPS, large stratum IV villages have small weights, and in fact most selected FSDPs in this district have been sampled in the SSUs of this size.)

A second estimation approach used was to calculate the expected number of CHC/PHCs and SCs based on *a priori* knowledge that such facilities were located in SSUs of minimum size 30,000 or 5,500, respectively. With 1991 census information on the SSU population, we reconstructed the distribution of each district's population by stratum size and divided each stratum by the CHC/PHC or SC catchment size (30,000 or 5,500 respectively). This provides the expected number of CHC/PHCs and SCs for each district. We can compare this with the observed number of such facilities, obtained at the time of fieldwork where local community informants were asked whether there was a CHC/PHC and/or SC located within the SSU. This comparison is shown in Table 6, which also includes a fieldwork organization code (I to IV) in the event any pattern of survey error is evident. This approach overestimates the number of subcenters by 19.6 percent and under-estimates the number of CHC/PHCs by 26.5 percent. Excluding the two districts with a high number of stratum I SSUs (Allahabad and Sultanpur) reduces the CHC/PHC underestimation to 10.2 percent. Tabulation of estimation bias by fieldwork organization shows no systematic bias.

The results from the two weighting approaches suggest that the SSU offers an appropriate measure of size (MoS) for the selection of subcenters, since its average population size may approximate the SC's catchment size of 5,500. A larger MoS may have served the selection of CHC/PHCs better, since this facility's catchment size covers those for five to six subcenters. Because SSU size is the basis for the weight for CHC/PHCs, when the selected SSU is small, the bias in estimated counts can be large. A future design to consider is to use a cluster of SSUs that are contiguous to the selected SSU and have an MoS similar to the catchment size of CHC/PHCs. The probability of such a facility being present within the boundaries of the SSU cluster will then be higher and the weight, constructed on the basis of the

total population in the SSU cluster, more reliable. In other words, our estimation is limited by not knowing how many SSUs are served by one CHC/PHC.

Table 5
Total Actual and Estimated Total Number of Community Health Centres, Primary Health Centers,^a and Subcentres by District in Uttar Pradesh, India: 1995

District	CHC/PHC		Sub-centre	
	Actual	Estimated	Actual	Estimated
Aligarh	77	69	399	369
Azamgarh	103	69	475	949
Almora	44	104	254	468
Allahabad	112	981	594	677
Ballia	73	93	357	485
Banda	89	101	322	302
Bareilly	71	42	355	162
Dehradun	24	41	139	60
Etawah	69	84	323	364
Fatehpur	57	73	309	327
Firozabad	33	34	234	236
Gonda	107	183	528	461
Gorakhpur	59	84	470	460
Jhansi	51	77	251	157
Kanpur Nagar	12	13	81	74
Maharajgang	30	39	195	180
Meerut	76	187	410	119
Mirzapur	64	69	309	302
Moradabad	92	81	485	248
Nainital	53	79	287	344
Rampur	37	19	170	139
Saharanpur	60	49	293	388
Shahjahanpur	52	59	301	298
Sultanpur	70	487	394	649
Tehri Garhwal	31	5	159	63
Unnao	63	162	344	106
Sitapur	87	44	437	450
Varanasi	122	144	616	658
Total	1818	3472(± 21)	9491	9495(± 15)
Total ^b	1636	2004(± 13)		

^a Includes additional primary health centres

^b Excludes Allahabad and Sultanpur districts

Source for 1995 actual figures from Government of Uttar Pradesh Department of Medical and Family Welfare.

Table 6
Observed and Expected Sampled Number of CHCs/PHC^a and Subcentres Within the
Rural Village (Urban Block) by District in Uttar Pradesh, India: 1995

District	CHC/PHC		Sub-Centre		Field Work Company
	Actual	Estimated	Actual	Estimated	
Aligarh	6	5	10	17	II
Azamgarh	3	5	24	15	III
Almora	5	2	14	9	I
Allahabad	19	4	17	18	III
Ballia	9	7	34	27	III
Banda	8	9	19	27	III
Bareilly	5	3	10	16	II
Dehradun	5	7	10	21	I
Etawah	8	7	17	20	II
Fatehpur	9	7	22	25	IV
Firozabad	6	6	28	30	II
Gonda	8	5	15	18	IV
Gorakhpur	5	4	16	20	IV
Jhansi	7	6	16	24	II
Kanpur Nagar	2	2	6	8	II
Maharajgang	4	4	9	13	IV
Meerut	12	8	12	34	II
Mirzapur	7	7	22	22	III
Moradabad	5	5	9	19	I
Nainital	6	4	19	19	I
Rampur	2	5	14	16	I
Saharanpur	6	6	25	21	I
Shahjahanpur	5	3	14	15	II
Sultanpur	16	6	21	15	IV
Tehri Garhwal	1	3	3	10	I
Unnao	3	6	17	17	IV
Sitapur	10	6	9	24	IV
Varanasi	6	5	18	18	III
Total	186	147	450	538	
Total ^b	151	137			

^a Includes additional primary health centres

^b Excludes Allahabad and Sultanpur districts.

4. DISCUSSION

The cluster-based sample design for generating independent samples of facilities and households, which can be analyzed individually or jointly, does warrant more extensive consideration in data collection efforts for health program research and evaluation in developing countries. Careful design and fieldwork sampling and execution can yield high-quality and acceptably precise survey estimates, as our results show. The weighted totals, rather than sample totals, themselves are numbers useful to program planners who decide the flow of personnel, material, and financial

resources to and among various facility sites and area locations. The linkage of facility to individual records offers further important analytic opportunities to assess the relative importance of personal background and service supply factors on health outcomes of interest (e.g., Boyd and Iversion 1979).

At the same time, our application of this design reveals several lessons. First there is an obvious need to monitor the survey fieldwork closely with increased on-site data entry so that the apparent "push" of eligible women out of the older age ranges can be prevented. This is difficult to detect through individual questionnaire spot checks but can

be observed in aggregate tabulations produced, say, weekly on completed questionnaires. Second, the excess count of CHCs/PHCs in two districts, where the survey fieldwork involved two *different* organizations suggests that stratum I villages might have been disproportionately selected or that some of the CHCs/PHCs reported to be within the SSU boundaries were in fact not. The former may have occurred as a sampling error since each fieldwork organization was provided with a list of sampled SSUs. Third, the listing and mapping of SSUs for facilities, individual health care providers, and households are an important stage of the fieldwork. Careful execution of this task allows the sampled units to be re-located for future follow-up. This will be an essential measurement effort for evaluating the IFPS project.

Certainly for a survey as complex as PERFORM, scaled to capture the levels of and differentials in the patterns of health service delivery and client use in an area as populous as Uttar Pradesh, the fact that the quality of the data meets most standards of precision evidences an important fieldwork achievement as well as design innovation.

ACKNOWLEDGMENTS

Partial support for this study has been provided by The EVALUATION Project, USAID Contract #DPE-3060-C-00-1054-00. The views contained herein are solely those of the authors and not the sponsoring agency. The authors

acknowledge with appreciation earlier assistance on the sample design from Daniel Horowitz and T.K. Roy. We thank Lynn Moody Igoe of Carolina Population Center for editing the paper. Authors are also thankful to the anonymous referees for their useful comments and suggestions.

REFERENCES

- ADAY, L.A. (1991). *Designing and Conducting Health Surveys: A Comprehensive*. San Francisco: Jossey-Bass Publishers.
- BOYD, L.H., Jr., and IVERSION, G.R. (1979). *Contextual Analysis: Concepts and Statistical Techniques*. Belmont, CA: Wadsworth.
- MACRO INTERNATIONAL, INC. (1996). *Demographic and Health Surveys Newsletter*, 8, 1-12.
- MILLER, R.A., NDHIOVU, L., GACHARA, M.M., and FISHER, A.A. (1991). The situation analysis study of the family planning program in Kenya. *Studies in Family Planning*, 22, 131-143.
- NARAYANA, G., CROSS, H.E., and BROWN, J.W. (1994). Family planning programs in Uttar Pradesh issues for strategy development: tables. Centre for Population and Development Studies, Hyderabad, India.
- ROSS, J.A., and McNAMARA, R. (Eds.) (1983). *Survey Analysis for the Guidance of Family Planning Programs*. Liege, Belgium: Ordina Editions.

Computer-assisted Interviewing in a Decentralised Environment: The Case of Household Surveys at Statistics Canada

J. DUFOUR, R. KAUSHAL and S. MICHAUD¹

ABSTRACT

In 1993, Statistics Canada implemented Computer-assisted Interviewing (CAI) for conducting interviews for some household surveys that were conducted in a decentralised environment. The technology has been successfully used for a number of years, and most household surveys have now been converted to this collection mode. This paper is a summary of the experience and the lessons that have been learned since the research started. It describes some of the tests that led to the implementation of the technology, and some of the new opportunities that have arisen with its implementation. It also discusses some challenges that were faced when CAI was implemented (some are on-going issues), and ends with a brief overview of where this may lead us in the future.

KEY WORDS: Household surveys; Data collection; Computer-assisted interviewing; Decentralised environment.

1. INTRODUCTION

The first systems of computer-assisted interviewing (CAI) were developed in the early 1970s (see Nicholls and Groves 1986). These systems were mainly developed by market research organisations in the United States and, a little later, independently by well-known university research centres. During the late 1970s and early 1980s, computer-assisted interviewing systems became much more sophisticated, and their use expanded greatly. By the late 1980s, a number of universities and survey research centres in the United States had a computerised collection system (see Lyberg, Biemer, Collins, de Leeuw, Dippo, Schwarz and Trewin 1997). Clark, Martin and Bates (1997) provide an overview of the development and implementation of such systems in four major government statistical agencies.

In 1987, Statistics Canada conducted its first experiment with computer-assisted interviewing for household surveys. At that time, the tests were done in a "centralised telephone collection environment". The series of tests with computer-assisted interviewing was extended into the early 1990s to try to adapt to the more general collection methodology.

At Statistics Canada most household surveys share a common sampling frame and data collection environment. The main user of this frame is the monthly Labour Force Survey (LFS). Data collection is decentralised with the initial interview in person at the selected dwelling and the subsequent five interviews by telephone from the interviewer's home. To accomplish this, almost a thousand interviewers have been equipped with portable computers. Interviewers are attached to one of the five regional offices located throughout Canada. A number of household surveys in the bureau follow a similar collection strategy by subsampling from the Labour Force Survey sample, by administering a series of supplementary questions after the Labour Force Survey interview or by contacting persons who have formerly participated in the survey. As a result,

not only is the Labour Force Survey sample shared with other surveys, but so is the collection infrastructure. All interviewers are required to work on the Labour Force Survey for a specified week each month, and for the rest of the time, they have been trained and equipped to collect data for other surveys. For further details on the Labour Force Survey methodology, see Statistics Canada (1998).

The 1990s saw testing of the implementation of the computer-assisted collection mode not only for the LFS but also for other surveys sharing that common infrastructure and having very different requirements. The results of the various tests led to the implementation of computer-assisted interviewing for the LFS in November 1993 (Dufour, Kaushal, Clark and Bench 1995) while its supplementary monthly surveys have been changed gradually. In January 1994, a new longitudinal survey, the Survey of Labour and Income Dynamics (SLID) was launched using computer-assisted interviewing (see Lavigne and Michaud 1995). Since then, the National Population Health Survey (NPHS) along with the National Longitudinal Survey of Children and Youth (NLSCY) introduced in August and November 1994 respectively, have also adopted this collection mode (see Tambay and Catlin 1995, Brodeur, Montigny and Bérard 1995). For further details on the structure and implementation of this computerised collection mode in longitudinal surveys, see Brown, Hale and Michaud (1997). Today most of Statistics Canada's household surveys are collected using a computerised mode and a common infrastructure.

This article focuses primarily on methodology aspects of decentralised computer-assisted interviewing for household surveys. We provide an overview of the implementation process for the statistical agency as a whole, a brief discussion of the challenges associated with the new collection vehicle and a list of references for more detailed information on specific topics. Despite "growing pains", Statistics Canada is continuing to experiment with and

¹ J. Dufour and R. Kaushal, Household Survey Methods Division; S. Michaud, Social Survey Methods Division, Statistics Canada, Ottawa, K1A 0T6.

implement this new technology in various surveys to render these surveys more cost efficient and to improve data quality and the survey monitoring process.

The article is divided into five sections. In the next section, aspects of implementation are discussed with reference to several surveys. Section 3 details new opportunities arising from computer-assisted interviewing. The ongoing challenges and new problems that surveys face as a result of using a decentralised computerised collection mode, as well as the changes that are taking place, are discussed in Section 4. The last section describes the future of CAI for household surveys at Statistics Canada.

2. FIRST YEARS OF IMPLEMENTATION

Adopting a computerised collection method for household surveys held the promise of several benefits: (i) a decrease in survey costs, (ii) better data quality, (iii) the possibility of using more complex questionnaires, (iv) data made available more quickly, (v) a tool for tracing operations, (vi) the possibility of using dependent interviews, and (vii) a generalised collection method for all of the agency's household surveys. However, these benefits were not realised overnight, or without effort. Ongoing evaluations and adjustments were required in the introduction and stabilisation phases.

Despite a number of tests being conducted before the implementation of CAI, unforeseeable problems occurred with the adoption of this method, but over time, they became less frequent and easier to solve. In addition, during this period, the series of quality indicators analysed carefully by different groups of Statistic Canada experts were somewhat disrupted. It took about one year to realise the anticipated benefits. This section describes the main points in the process of changing from the traditional paper approach to computer-assisted interviewing, where collection and capture are integrated.

2.1 Centralised Computer-assisted Telephone Interviewing

The traditional approach to interviewing used a paper questionnaire filled out in pencil to facilitate edits made by the interviewer. Often such an approach is referred to as Paper and Pencil Interviewing (PAPI). In this traditional mode, an interviewer edited the questionnaire to ensure that the information was correct and complete. Information abbreviated to shorten the interview was filled-in in detail after the interview and before the form was sent for data capture. The first change towards computerisation was the use of Computer-assisted Telephone Interviewing (CATI). This computerised collection mode was used for surveys that were conducted by telephone from a central location. CATI was the first instance of amalgamation of the collection and capture of information in household surveys. Given the state of technology at that point, the computers capable of handling the complexity associated with computer-assisted interviewing were fairly large. Hence,

CATI could replace PAPI only in centralised telephone surveys. In the 1990s, with the advent of more powerful portable computers decentralised CAI replaced PAPI. A decentralised collection mode is, in effect, what is used in most household surveys. In addition, data collection often required the ability to do either telephone interviews or personal visits. However, much of the know-how and experience of computer-assisted telephone interviewing could be applied to decentralised computer-assisted interviewing.

Since the 1980s, it was the Labour Force Survey (LFS) that served as the main research and testing vehicle for CATI technology. The first test, conducted in 1987, was a controlled study that compared CATI in a centralised environment to PAPI. It consisted of a research project carried out jointly between Statistics Canada and the US Bureau of the Census (see Catlin and Ingram 1988). The study showed that there were differences between the two collection methods in terms of data quality indicators, and those differences were in favour of CAI in terms of lower rejection rates on edit, reduction in path errors on the questionnaire and decrease in undercoverage in the LFS.

While CATI was never implemented for the LFS, the experience was used to set up a CATI facility for use in random digit dialling (RDD) in household surveys. As technology progressed, CATI was used to collect more complicated RDD surveys like the General Social Survey (GSS) and the Violence against Women Survey. Computer-assisted telephone interviewing continues to be used as an integral part of household collection at Statistics Canada complemented by the computer-assisted interviewing infrastructure.

2.2 Technological Testing

A new wave of testing began in the early 1990s as part of the decennial redesign of the LFS (Singh, Gambino and Laniel 1993; Drew, Gambino, Akyeampong and Williams 1991). The launching of three large scale longitudinal surveys by Statistics Canada made the investment for a CAI infrastructure possible by sharing the costs among a number of surveys. Consequently, in 1991, a second test was conducted using the LFS and SLID to study the feasibility of using new technologies (see Williams and Spaul 1992). Portable computers which require the use of a stylus rather than a keyboard for entering data were tested. The results showed that the technology was promising but that it needed further improvements for it to be used to handle the requirements of Statistics Canada's household surveys.

The following year, from July 1992 to January 1993, a third and a fourth test were conducted, this time using conventional portable computers. The results for the LFS are documented in Kaushal and Laniel (1995), while the results for SLID are reported in Michaud, Le Petit, and Lavigne (1993) and Michaud, Lavigne and Pottle (1993). For the LFS, the main objective of this third test was to determine if the transition to the new technology would disrupt the LFS data series. The secondary objective of the test was to determine whether the new technology affected

data quality and interview costs. Additional objectives of this test were the operational development and evaluation of the CAI approach. For the longitudinal surveys, the main concern was the length and complexity of the questionnaires and the addition of new functions, such as tracing. Consequently, the main criterion in assessing the application was the feasibility of developing various functions. The results showed that CAI had no major impact for the LFS on either the data series disseminated, the survey's main quality indicators, or interview costs. On the strength of general comparisons with outside sources and an analysis of missing variables, the new technology was adopted.

2.3 New Dimension of Nonresponse

With the adoption of CAI, there was an unintentional development of a new dimension of nonresponse that is due to "technical problems". Such nonresponse resulted from cases that were lost or not received before the end of the collection period. The PAPI version of this type of nonresponse was related to occasional postal problems. Conceptually, these situations do not refer to real nonrespondents; however, the information is not available in time to produce estimates.

These technical problems assume three different forms: (i) transmission problems, (ii) equipment problems, and (iii) unavoidable problems. Transmission problems are the most common. They arise, for example, when telephone lines are down, when there is a problem with the automatic downloading of data, when an attempt is made to download data while maintenance is being carried out on the mainframe computer, or simply because of a malfunction in the CAI system. The second type of problem, although less common, occurs when a hard drive crashes, the magnetic tape drive fails, there is insufficient memory or there are computer equipment problems at the regional offices. Finally, unavoidable problems, which are even less common, include specific problems implicitly created by the above two categories, for example when only one of the two components expected from a respondent is transmitted or if the initialisation parameters needed for the proper functioning of the programs are missing.

Nonresponse due to technical problems diminished over the initial months. This component of nonresponse was analysed quite carefully to explain an upward trend in nonresponse and to assess the performance of the CAI approach (see Simard, Dufour and Mayda 1995; Dufour, Simard and Mayda 1995). At the start of the conversion of the household surveys to CAI, technical problems represented on average 15% of total nonresponse and could alone explain up to 25% of nonresponse. It took almost a full year before any significant reduction was observed in this component of nonresponse. Today, in 1997, the nonresponse due to technical problems is practically non-existent.

In the first year, the bulk of the problems were due to a conflict over memory management in the notebook computer between two pieces of software used in case management. This was resolved by a re-write of a part of

the software, which eliminated the conflict and made the system more efficient. The more subtle issues of the transition were communication and experience. A communication strategy was developed to enable the different players (in particular technical personnel and interviewers) to better understand each other, disseminate information more quickly and adequately inform all persons concerned. When CAI was first introduced, it took technical support personnel more than a day to find a solution to some problems. Faster response procedures were established, and a 24-hour support service was set up at head office in Ottawa. With such a substantial change, a learning and adjustment period is required, and Statistics Canada was no exception.

2.4 Impact of CAI on Nonresponse

Are there grounds for believing that the use of CAI had an effect on nonresponse rates? The answer to such a question has to be yes in light of the technical problems encountered, primarily at the beginning of the conversion process. However, if this aspect of the nonresponse is discounted, there is no indication that CAI had any lasting effect on nonresponse rates. The LFS nonresponse fluctuated following the introduction of CAI, but these fluctuations may be explained by a number of other factors (the redesign of the sample, which is now more urbanised; hiring of new interviewers; *etc.*), since the LFS was undergoing a major overhaul. It took just under two years for overall nonresponse to return to levels similar to those recorded in the paper and pencil era.

In the LFS, the conversion took place over a period of five months during which time the CAI and PAPI nonresponse rates could be compared. These comparisons show that the nonresponse rates for CAI (excluding technical problems) and those for PAPI were in the same range and exhibited the same trends (see Simard and Dufour 1995). Moreover, all the main components of nonresponse, namely refusal to participate in the survey, household temporarily absent, no one at home and other reasons, exhibited similar annual patterns before and after the implementation. There were concerns that respondents would be more reluctant to answer due to the presence of a computer for personal interviews, resulting in an increase in refusals. However, no change in the refusal component was detected.

In early 1995, the three longitudinal surveys (SLID, NLSCY and NPHS), as well as the LFS, were conducted during similar collection periods. The current case management environment, as well as the sharing of the infrastructure among surveys, created extra pressure on interviewers in the field. Moreover, the survey collection periods were limited because there was a limited number of applications that could reside on the computers at the same time. Analysis was done to determine if response problems arose from conducting several surveys simultaneously, or in quick succession, in the field using CAI. For the quarterly collection of the NPHS, interviewers followed-up nonrespondents in previous collections. An analysis was

carried out to determine the possible conversion rate. The results showed that in the case where there were fewer CAI surveys in the field at the same time, a first wave of follow-ups of nonrespondents increased the response rate, but continuing the process for a second or third time brought few gains (an increase of 5.76% from the first to the second quarter, 0.97% from the second to the third, and 0.91% from the third to the fourth). However, a last follow-up was carried out in June 1995 when there were almost no surveys in the field. This procedure improved the overall response rate by approximately 5%, which was higher than expected. This led to the conclusion that CAI had to be able to give more flexibility in the length of the collection period and allow multiple applications to reside on the computer in order to maintain the response rates that would have been obtained in a paper and pencil environment.

3. NEW OPPORTUNITIES FOR HOUSEHOLD SURVEYS

The adoption of CAI collection has added new opportunities to household surveys. These new opportunities, which were either non-existent or operationally difficult in a paper and pencil mode, help to reduce non-sampling errors, to collect more specialised information, to facilitate the reconstruction of family units and to make contact with family units that break apart or merge. In fact, this collection method is better suited to adjust the collection process according to the changing needs of today's society.

3.1 Dependent Interviews

The introduction of the new technology served to resolve household survey problems that had proven intractable under the traditional paper and pencil interview approach. In particular, CAI helped to increase the information that could be provided by the interviewer to a respondent contacted for the second time for the reduction of (i) response error (coding, capture or recall error), in particular the seam problem and telescoping, and (ii) response burden by confirming the information instead of requesting it again (or by requesting only partial information).

The seam problem has been documented for longitudinal surveys in Murray, Michaud, Egan and Lemaître (1990), which notes that the problem arises in reconciling data from successive collection periods. If no reconciliation has been attempted between collections, an artificially large change in estimates is generally observed at each collection transition. This problem is generally explained by respondents' difficulty in pinpointing the date when a change occurs. As to telescoping, it results from a tendency to include certain events that occurred outside the reference period.

Under the traditional PAPI approach, the type of information that could be provided to interviewers was limited. Questionnaires could only be pre-printed with basic

information, as there were physical limits to the amount of information that could be pre-printed, especially for long questionnaires. In some cases, additional information was even printed on a separate questionnaire. This procedure also involved additional logistical problems for the interviewer. The use of information from earlier occasions in the process is known as feedback. With computer-assisted interviewing, feedback is made possible in two ways: proactively and reactively. A discussion of this is also provided in Brown *et al.* (1997).

Proactive use of feedback is used to reduce response error by helping the respondent to situate him/herself. For example, SLID gathers detailed information on a maximum of six jobs in the previous year. Without feedback, the name of the employer or the occupation might be written slightly differently, and a job that continued over a period of two years could be incorrectly classified as a change. Initially there was some concern that the respondent would perceive feedback negatively, but in fact, few negative comments have been received.

The confirmation rate is generally high – over 90% for data that are presented to the respondent (see Hale and Michaud 1995). The study of Hiemstra, Lavigne and Webber (1993) concerning the labour market suggests that while feedback generally serves to reduce the seam effect, the problem is only partially solved. For example, SLID confirms employment, job search or joblessness at the beginning of the previous calendar year over a one-year recall period. Micro-comparisons with a cross-sectional monthly survey, conducted over the first five months of the year, suggest that feedback greatly reduces the seam effect. However, consistency with cross-sectional data decreases over the months, which seems to suggest that response error, although eased by feedback, is still a problem.

The proactive use of feedback may, however, underestimate measures of change. For this reason, for sensitive information and for reasons of confidentiality, the technique is also used reactively. The reactive use of feedback can be used to detect unusual changes, or to confirm inconsistencies in the data. As an illustration, in the interview for the first wave of SLID, jobless spells are identified and for each spell the respondent is asked whether employment insurance benefits have been received. The second wave interview asks for detailed information on various sources of income and amounts received including employment insurance benefits. Comparisons with outside sources suggest that traditionally, the amounts of employment insurance reported in a survey represent approximately 80% of the contributions paid. In SLID, previous information was stored in memory. If an amount was not reported and there was an indicator flagging an inconsistency with the first-wave interview, an additional question was asked to determine whether the amount had been omitted. An analysis of the first wave of SLID suggests that reactive checking increased the number of reported cases by nearly 30%. However, 28% of these persons who had neglected to report an amount, confirmed that they had received an amount but were unwilling to

report that amount. There was thus confirmation of the source, but the amount had to be imputed and the problem was not totally solved. More details on this subject may be found in Dibbs, Hale, Loverock and Michaud (1995).

3.2 A More Efficient Tool

With an efficient collection tool like CAI, it is now possible to collect, to limit, to access and to transfer detailed information which would traditionally have been very difficult, or even not possible, to do with PAPI.

3.2.1 Matrix of Relationships Between the Various Members of a Household

Household surveys create different levels for analysis such as the economic family and the census family, by using the relationships between the various persons in the household with a single person often called the "family head". There are limitations to this method for example, in identifying the children of blended families or reconstructing families to three generations. In a longitudinal context, the concept of family head is a definition that can vary over time and so a number of longitudinal surveys have used a matrix of relationships for all members of the household. CAI can limit collection to the lower diagonal of the matrix. Provided that the composition of a household does not change between two collections, it is not necessary to re-ask it for the relationship matrix. Interactive edits (based on age, for example) serve to correct any relationships captured in reverse (e.g., a parent-child relationship). It took a number of attempts to develop an effective means of identifying relationships that would allow not only for the collection of the information but also for easy correction. With the improved version of the collection procedure, less than 1% of relationships required further correction after collection (as compared to 5.3% inconsistency before the interactive edits on the relationship matrix). Corrections in a CAI environment probably continue to be one of the areas in which research is still required.

3.2.2 Access to More Sophisticated Collection Instruments

CAI has also provided access to more sophisticated collection instruments. For example, the NLSCY obtains a variety of information on a cohort of children aged 0-11 years. One part of the interview is designed to measure the child's vocabulary level. The survey uses the Peabody Picture Vocabulary Test (PPVT) as one of its collection instruments. However, the PPVT is normally used in a more specialised environment, and persons administering it generally need several days of in-depth training since the test involves a series of images, and the child is asked to choose the image that corresponds to a given word. The starting level depends on the child's age. Questions are administered until the child gets a certain number of wrong answers. At this point, the interviewer must return to the starting level and re-administer the previous questions, until

the child gives a pre-determined number of wrong answers. The administration of the test calls for determining a threshold based on criteria, counting the number of wrong answers, skipping between questions depending on the number of wrong answers, and stopping the test. These procedures would have required a considerable amount of training if it had been necessary to administer the test on paper. CAI has greatly facilitated the process by allowing programming of the edit rules in advance. The data from the first collection suggest that the computer-assisted conditions of administration yield good-quality results when compared to external norms.

3.2.3 Establishing Longitudinal Links

In the case of longitudinal links, it may happen that all the members of an initial household may be part of the longitudinal sample, as in SLID for example. In subsequent collections, the longitudinal persons are interviewed along with all persons with whom they live. In the case of a household that splits, a new household must be created for the persons who left the original household. With the adoption of CAI, it became possible to create new unique household identifiers linked to the original identifiers, this made it easier to reconcile the dynamics of change in household composition. A particular problem that has been greatly lessened is the treatment of the real duplicates that occur as a result of changes in household composition. For example, an adolescent might belong to a given household at the time of the first collection, then leave his parent's household by the time of the second collection but return to the original household by the time of the third collection. In the second collection, the person is identified as belonging to a new household, and a new identifier is thus associated with him. In the third collection, when the parents' household is again contacted, the adolescent who has returned may be indicated as a new person in the household. If the interviewer is shown the list of persons who have formerly been part of the household, the need to reconcile duplicates is greatly reduced. A similar treatment has been carried out for jobs where a list of previous employers is used for longitudinal reconciliation of jobs.

3.2.4 Tracing of Individuals

With the conversion to CAI, certain procedures such as tracing were automated. Brown *et al.* (1997) gives specific examples. As noted above with respect to establishing longitudinal links, traced individuals may all be put into a new household with a unique identifier. Fewer paper manipulations are required, and it is now possible to obtain more management information. CAI has made it possible to set up a two-level tracing procedure. The interviewer first attempts the tracing. If this is not successful, all information on the case is transferred to a tracing unit in the regional office where more sources for tracing are available. Automation has eliminated many manipulations and transcriptions of records on paper. Formerly when a household split, a new identification sheet was usually created on paper with a link to the previous household. The

names of the persons who had moved were entered on it. If the person to be traced was not found, all the forms for all the persons who had been living together the previous year were transferred. These manipulations greatly increased the risk of error. Transfers of cases between tracing levels are also done more quickly. In addition, each call is recorded automatically along with its result. While there was a similar procedure with the paper and pencil approach, the information was seldom entered. It was also hard to analyse the information for determining the most useful tracing sources.

Tracing is a key factor in maintaining data quality. With current tracing procedures, cases requiring tracing can be kept in the field a little longer, but the collection window remains limited. It is possible that more effective procedures can be established if the efforts of the various longitudinal surveys are integrated. Increased functionality, combined with central tracing, is currently being examined. This would make it possible to combine the tracing efforts of the various surveys, and it might also make it possible to have batch entries to try to link cases requiring tracing to databases.

3.3 New Quality Indicators

The CAI approach adopted by Statistics Canada for its household surveys features a complex system capable of monitoring survey activities during the collection period to ensure their smooth operation. This system called the "case management system" (CMS), is a sophisticated system that manages all survey activities from the beginning to the end of the survey cycle. This system is flexible, since it can be adapted to the requirements of the different household surveys that use it. The CMS performs three main functions: (i) routing of cases, (ii) reporting of activities and (iii) assisting interviewers. The routing component directs the movements of cases during the survey, whether from an interviewer to the regional office, from the regional office to head office, *etc.* The second component of the CMS produces different reports for describing the status of the survey at a given point in time, evaluating the performance and progress of the survey, and describing the status of interviews. A whole range of information is generated by this second component of the CMS. Lastly, the third module enables interviewers to perform their tasks more effectively, by giving options for making appointments, recording notes and so on.

As a result, this system provides a mass of information on what is actually happening in the field during a survey; every action taken on a case is recorded by the CMS. The main challenge with such a system is to avoid getting lost in the great mass of information available. Work teams have been set up to master these information sources, develop new quality indicators using this information or combining it with information already available, find uses (*e.g.*, additional training, improvement of the collection instrument), and develop ways to present these indicators effectively.

A large number of quality indicators have been produced (see Simard *et al.* 1995; Allard, Brisebois, Dufour and Simard 1996) on a regular basis at different levels of interest (geographic, interviewers, administrative). These indicators may be grouped into two categories: informational and for monitoring purposes. Examples of informational indicators are: number of attempts before completing a case, distribution of interviews completed per day of collection, best day-hour combination for reaching a respondent, median duration of interviews, and number of edit rules triggered and ignored or triggered and acted upon (see Brisebois, Dufour, Lévesque 1997). Information indicators are used to improve or make changes to the collection strategy or process.

In terms of monitoring, a series of indicators are used to trace irregularities, technical or human, in the field. Among these are: calls and visits done after the date of transmission but before the survey week, calls and visits done after Sunday of survey week, working period too early, working period too late, interviews too short, *etc.* This information serves to show whether instructions issued by head office are followed, and whether some interviewers require additional training. However, all data need to be analysed with caution to determine the cause of the irregularity. For example, an interview conducted at 4:30 am may well be at the request of a respondent, like a farmer, or due to an incorrect time on the computer clock (see Brisebois *et al.* 1997).

CAI also offers interviewers the opportunity to include a comment for each question or to explain the reason for the code used. It is therefore possible to develop adequate training, to better understand the surveys and accordingly to adapt them to realities in the field. For example, this feature made it possible to conduct a special study on the reasons for refusal to participate in one of Statistics Canada's household surveys; to conduct such a study would have formerly required a great deal of effort (see Allard, Dufour, Simard and Bastien 1996).

4. ONGOING CHALLENGES OF CAI

This section describes long-term challenges in developing, implementing and understanding the use of CAI for survey applications. The powerful tools provided by CAI have led us to degrees of complexity in content, software and electronic communications that may not be widely appreciated. The conversion to CAI has implied a new dependence on informatics. This dependence is one of the major challenges that Statistics Canada has to face with CAI, since the technology is changing so quickly.

4.1 Workload of Interviewers

A common infrastructure requires the sharing of limited resources, such as trained interviewers equipped with portable computers, by different surveys. As a consequence, any increase in either the number of surveys or the amount of information collected must be carried out jointly with the

other surveys. It should be noted that the same interviewers tend to be used by a large number of surveys, which can result in fairly large workloads, exacerbated by a short collection period. While response rates have recovered since the introduction of CAI, a heavy workload for interviewers can lead to deterioration in data quality, owing to fewer follow-ups and higher nonresponse.

Given the nature of the CMS, an administrative structure for communication, based on the needs of a given survey (based on the response codes), must be put in place to provide for the routing of cases between the interviewers, their supervisors and the regional offices. Since CAI was first introduced, there have been great improvements in the communications process to ensure that all interviewers correctly receive their assignments, the latest version of the application or various changes; nevertheless, this process must be constantly monitored. For example, after the end of the collection period, cases must be transmitted and deleted from the interviewers' computers. Often, the cases that were not transmitted consist mainly of nonresponse cases. The fact that these cases are not transmitted to head office after the end of collection means that the reasons for nonresponse are sometimes lost. While many of these problems can be detected during testing, the fact remains that a few exceptional cases still remain.

4.2 Control Procedures for CAI

The CMS and survey applications have the potential to generate many databases. The quantity of data is often overwhelming, and the data are not currently being used to their maximum potential. In addition, the speed inherent in CAI sometimes does not allow for sufficient time and resources to analyse and control this mass of information. For the moment, this information is used after the fact, but it would be highly desirable to be able to use it while the survey is in the field.

This information should be made available to interviewers in an integrated format. However, a balance is needed to avoid excessive surveillance where interviewers focus more on the quality indicators than on the quality of the data. Ideally, analysis across several surveys could identify specific problems, which could then be dealt with in training kits that are brief and focused. In addition, response rates and coverage rates could be integrated for surveys. All this information could be used to achieve more efficient time management or to develop training in specific interview skills.

4.3 Editing During Collection

While CAI offers the possibility of including a great number of edit rules at the time of the interview, it is important here as well to maintain a balance between the rules programmed into the collection instrument and the rules applied during batch processing at head office. The rules programmed into the instrument prolong the interview, which results in an increase in both costs and response burden. Over time, and with rapid changes in technology, it should be possible to apply a larger number

of edits during the interview without interfering with its flow. On the other hand, clarifications at the time of the interview undeniably result in better quality data. The NPHS obtains better quality data in the second quarter by using information from the first quarter to feed the edit system. For example, clarifying with the respondent at the interview, led to the discovery that, for the arthritis variable, of the 7.0% of individuals who indicated a change in condition between the two quarters, 3.3% actually experienced a change while 3.5% represented errors. For further details, see Catlin, Roberts and Ingram (1996).

With CAI, it is also possible to store information to identify which edit rules have been triggered and what corrections were made. A study of the most frequently triggered edit rules would determine which rules most affect data quality, with the results of these studies serving not only as information but also as inputs, for changing overly strict edit rules and also for sustaining a dynamic correction system. Another aspect that is just as important is the ease with which the interviewer can make the necessary corrections. If the corrections can be made to the actual response or the preceding response to a question, the interviewer can easily identify the changes to be made. If the correction involves editing between several answers, then the need to determine which one requires correction, and to move between the various answers in which there may be an error, sometimes makes the process too complex for the edit to be carried out during the interview.

Apart from technical problems, there are methodological problems associated with the effect of edit rules on data quality. At what stage are the different edit rules the most effective? The rules that affect the flow of the questionnaire and those that determine which persons are outside the scope of the survey, are critical edit rules. The key variables used for poststratification and key estimates are best resolved at the time of the interview. The quantity of edit rules that can be incorporated into the CAI system must be balanced with the speed of the portable computer. In addition, when some edit rules are being developed for the instrument and others for central processing, care must be taken to ensure that the two types of rules are not contradictory.

4.5 Data Confidentiality

Maintaining data confidentiality, as stipulated by the *Statistics Act*, is one of the fundamental requirements of the use of CAI and the systems that support it. To meet such a requirement, a number of procedures have been developed including a computing environment with two communication networks, one external and the other internal. The data are transferred physically, by tape, from the external network to the confidential internal network since there is no link between these two networks. It is impossible to access the internal network using a public modem. Confidentiality is also ensured by encryption of data whenever they must be transmitted over telephone lines. In addition, an access control system is incorporated into all portable computers, enabling only the interviewer to access

the information. The data are also encrypted while residing on the notebook.

The challenges relating to confidentiality in a CAI environment are quite different from those encountered with PAPI. Dependent interviews offer such a challenge for SLID. Information available from the preceding wave family unit may become sensitive in the case of, say, a family break-up. Thus, while the new technology offers the benefits of dependent interviews, these are accompanied by drawbacks that must be analysed for the specific situation.

With the arrival of audio-CASI (known by the acronym CASI-A), sensitive subjects may be handled more easily. With this interview technique, respondents are linked to the computer with earphones, and the questions are read by a digitised voice. Since the question is heard via the headset, the respondent can choose whether or not to display the question on the screen. With these features, the respondent can complete the questionnaire in total anonymity. The NLSCY is planning to begin using this collection instrument by the year 2000.

4.6 Re-Interview Programs

CAI offers some enhancements over PAPI-based re-interview programs. Firstly, the rapid electronic transmission of data reduces discrepancies due to recall and memory problems since re-interview can be conducted quicker after the initial interview. Strict adherence to reconciliation procedures built into the software provides more accurate estimates of measurement error. This would eradicate the problem of interviewers peeking at the questionnaire before starting the re-interview. As well, reconciliation can be done after a subset of questions, a section or at the end of the questionnaire and as many times as desired. Re-interview cases are easily automated and integrated into a quality control process based on characteristics of the interviewer or the interview (e.g., specific cases related to training issues, cases belonging to a specific group, *etc.*). The quality of the data is better since a great number of edit rules, identical to the ones used during the interview, are programmed for the re-interview. The features available from the CMS are also an asset for the re-interview program: progress of the re-interview program, performance and progress of the re-interview, easy transfer of cases, *etc.*

4.7 Interviewer Training

With the adoption of CAI, interviewers had to cope with a major change in their work method. Training was therefore an essential stage in enabling them to adapt effectively to the computerised collection method. They became familiar with new work tools, including the keyboard, the portable computer and all the computer procedures, such as saving data, charging batteries and transmitting by modem. They also had to adapt their interview style to the requirements of CAI. New interviewers, for their part, had to familiarise themselves with survey concepts, interview techniques and the

collection instrument. To meet this challenge, Statistics Canada developed a training strategy based on the experience acquired during the previous testing, as well as on the experience of British and American colleagues.

Interviewer training will always be one of the key factors in the success of Statistics Canada surveys, and the agency is continually innovating in this field. For example, one of the initiatives for the LFS is a training strategy to enable senior interviewers to regularly receive a small CAI assignment (approximately 15 cases), just so they can practice collection by this method and thereby stay abreast of changes in the CAI application. In addition to the regular practice cases that are always available on the computer, the CAI system will provide interviewers with modules integrated into the collection system, dealing with such complex subjects as coverage and multiple dwellings, to enable them to always be updated or to review various difficult concepts.

5. FUTURE OF CAI AT STATISTICS CANADA

In the new environment of limited resources and high response burden, collection is becoming increasingly customised. While business surveys have been doing it for some time, mixed collection is beginning to be in demand for household surveys. Centralised collection outside the collection window for a limited number of respondents can be used to improve response rates (to focus on tracing for example). The environment necessary for this type of collection more closely resembles a CATI environment in which shared database functions for a small sample are available, with call planning functions.

A complete redesign of the CAI application and the case management system is expected to be completed by the turn of the century. In this redesign, work teams must take account not only of computer capacity but also of the human factor. The latter factor is important since data collection and data quality depend on it. Interviewers must read the screen and enter the responses, tasks that call for perceptual and motor skills different from those required for pencil and paper interviews. The wording of questions is also harder to read on the screen, and interviewers mention that it is now harder to visualise the overall structure of a questionnaire. Hence special attention must be paid to screen design, the choice of colours, the amount of text displayed, the key functions pre-programmed and the ease of moving between screens. Since interviewers are also asked to work on several surveys, an effort should be made to standardise screen formats as much as possible.

As regards the hardware and software components, work teams are currently concentrating on choosing the best combination. At present, different softwares are used for different components of some surveys. In order to standardise the applications available as much as possible, there are plans to use a uniform platform for all surveys in a Windows environment. The Windows environment should give both interviewers and programmers greater

flexibility. The security systems must also be redesigned to conform to the technology adopted and to satisfy the requirements of Statistics Canada. Harmonisation of questions among surveys should be attempted, which would allow CAI programming to become more modularised. Respondent burden would also be reduced.

The new system will have to be able to take account of both past and present requirements. For example, system features are re-examined in the light of the progress reports provided to operational staff in order to determine which areas need improvement. As noted in Section 4, a number of other possibilities are being considered such as, interactive training of interviewers, special training modules, the possibility of conducting re-interviews and better tracing tools. These procedures should make it possible to make better use of the flexibility resulting from the automation of the process.

The case management system is also being redeveloped. One major consideration here is to obtain a robust communications system, in which changes can be sent out uniformly with a replication capability. While we still hope to develop a computer system that will be used for many years, the current reality seems to suggest that CAI is likely to continue to evolve rapidly. One challenge, then, since the technology is changing quickly (one need only think of the Internet), is to develop a new system that is flexible, so as to allow for adaptations without requiring a complete overhaul.

ACKNOWLEDGEMENTS

The authors would like to thank the many people of Household Survey Methods, Social Survey Methods, Household Surveys and Survey Operations Divisions who have contributed to the development of CAI at Statistics Canada over the years. It is their work that has made this paper possible. They would also like to thank Ann Brown, Brian Williams, Jean-Louis Tambay and Frank Mayda for their valuable comments that helped improve the quality of the paper.

REFERENCES

- ALLARD, B., BRISEBOIS, F., DUFOUR, J., and SIMARD, M. (1996). How Do Interviewers Do Their job? A Look at New Data Quality Measures for the Canadian Labour Force Survey. Presented at the International Conference on Computer-assisted Survey Information Collection.
- ALLARD, B., DUFOUR, J., SIMARD, M., and BASTIEN, J.-F. (1996). Pourquoi refuse-t-on de participer aux enquêtes? Le cas de l'Enquête sur la population active. Methodology Branch Working Paper, HSMD, 96-003F. Statistics Canada.
- BRISEBOIS, F., DUFOUR, J., and LÉVESQUE, I. (1997). New LFS quality measures. Methodology Branch Working Paper, to be published. Statistics Canada.
- BRODEUR, M., MONTIGNY, G., and BÉRARD, H. (1995). Challenge in developing the National Longitudinal Survey of Children. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 21-28.
- BROWN, A., HALE, A., and MICHAUD, S. (1997). Use of Computer-assisted Interviewing in Longitudinal Surveys. Presented at the International Conference on Computer-assisted Survey Information Collection.
- CATLIN, G., and INGRAM, S. (1988). The effects of CATI on cost and data quality. In *Telephone Survey Methodology*, edited by R.M. Groves *et al.*, New York: John Wiley and Sons.
- CATLIN, G., ROBERTS, K. and INGRAM, S. (1996). The validity of self-reported chronic conditions in the National Population Health Survey. Presented at Symposium 96, Nonsampling Errors, Statistics Canada.
- CLARK, C., MARTIN, J., and BATES, N. (1997). Development and Implementation of CASIC in Government Statistical Agencies. Presented at the International Conference on Computer-assisted Survey Information Collection.
- DIBBS, R., HALE, A., LOVEROCK, R., and MICHAUD, S. (1995). Some Effects of Computer-assisted Interviewing on the Data Quality of the Survey of Labour and Income Dynamics. Survey of Labour and Income Dynamics Research Paper, 95-07. Statistics Canada.
- DREW, D., GAMBINO, J., AKYEAMPONG, E., and WILLIAMS, B. (1991). Plans for the 1991 redesign of the Canadian Labour Force Survey. *Proceedings of the Section on Survey Research Methods, American Statistical Association*.
- DUFOUR, J., KAUSHAL, R., CLARK, C., and BENCH, J. (1995). Converting the Labour Force Survey to Computer-assisted Interviewing. Methodology Branch Working Paper, HSMD, 95-009E. Statistics Canada.
- DUFOUR, J., SIMARD, M., and MAYDA, F. (1995). The First Year of Computer-assisted Interviewing for the Canadian Labour Force Survey: An Update. Methodology Branch Working Paper, HSMD, 95-011E. Statistics Canada.
- HALE, A., and MICHAUD, S. (1995). Dependent Interviewing: Impact on Recall and on Labour Market Transitions. Survey of Labour and Income Dynamics Research Paper, 95-06. Statistics Canada.
- HIEMSTRA, D., LAVIGNE, M., and WEBBER, M. (1993). Labour Force Classification in SLID: Evaluation of Test 3A Results. Survey of Labour and Income Dynamics Research Paper, 93-14. Statistics Canada.
- KAUSHAL, R., and LANIEL, N. (1995). Computer-assisted interviewing data quality test. *Proceedings of the 1993 Annual Research Conference*. U.S. Bureau of the Census, 513-524.
- LAVIGNE, M., and MICHAUD, S. (1995). Aspects généraux de l'Enquête sur la dynamique du travail et du revenu. *Recueil des textes des présentations du colloque sur les applications de la statistique*. L'association canadienne française pour l'avancement des sciences.
- LYBERG, L., BIEMER, P., COLLINS, M., de LEEUW, E., DIPPO, C., SCHWARZ, N., and TREWIN, D. (1997). *Survey Measurement and Process Quality*. New York: John Wiley and Sons.

- MICHAUD, S., LE PETIT, C., and LAVIGNE, M. (1993). Qualitative Aspects of SLID Test 3A Data Collection. Survey of Labour and Income Dynamics Research Papers, 93-07. Statistics Canada
- MICHAUD, S., LAVIGNE, M., and POTTLE, J. (1993). Qualitative Aspects of SLID Test 3B Data Collection. Survey of Labour and Income Dynamics Research Papers, 93-11. Statistics Canada.
- MURRAY T.S., MICHAUD, S., EGAN, M., and LEMAÎTRE, G. (1990). Invisible seams? The experience with the Canadian Labour Market Activity Survey. *Proceedings of the 1990 Annual Research Conference*. U.S. Bureau of the Census.
- NICHOLLS II, W.L., and GROVES, R.M. (1986). The status of computer-assisted telephone interviewing: Part I. *Journal of Official Statistics*, 2, 93-115.
- SIMARD, M., and DUFOUR, J. (1995). Impact of The Introduction of Computer-assisted Interviewing as the New Labour Force Survey Data Collection Method. Technical Report, Household Survey Methods Division, Statistics Canada.
- SIMARD, M., DUFOUR, J., and MAYDA, F. (1995). The first year of computer-assisted interviewing as the Canadian Labour Force Survey data collection method. *Proceedings of Section on Survey Research Methods, American Statistical Association*, 533-538.
- SINGH, M.P., GAMBINO, J., and LANIEL, N. (1993). Research studies for the Labour Force Survey sample redesign. *Proceedings of the Section on Survey Research Methods, American Statistical Association*.
- STATISTICS CANADA (1998). *Methodology of the Canadian Labour Force Survey*. Catalogue 71-526. To appear.
- TAMBAY, J.-L., and CATLIN, G. (1995). Sample Design of the National Population Health Survey. Health Reports. Catalogue: 82-003, Statistics Canada. 7, 29-38.
- WILLIAMS, B., and SPAULL, M. (1992). Computer-assisted Personal Interviewing LFS Datellite Test 0691-1191. Internal report. ISS Managers Conference, Statistics Canada.

Regression Analysis of Data Files that are Computer Matched - Part II

FRITZ SCHEUREN and WILLIAM E. WINKLER¹

ABSTRACT

Many policy decisions are best made when there is supporting statistical evidence based on analyses of appropriate microdata. Sometimes all the needed data exist but reside in multiple files for which common identifiers (e.g., SIN's, EIN's, or SSN's) are unavailable. This paper demonstrates a methodology for analyzing two such files: (1) when there is common nonunique information subject to significant error and (2) when each source file contains noncommon quantitative data that can be connected with appropriate models. Such a situation might arise with files of businesses only having difficult-to-use name and address information in common, one file with the energy products consumed by the companies, and the other file containing the types and amounts of goods they produce. Another situation might arise with files on individuals in which one file has earnings data, another information about health-related expenses, and a third information about receipts of supplemental payments. The goal of the methodology presented is to produce valid statistical analyses; appropriate microdata files may or may not be produced.

KEY WORDS: Edit; Imputation; Record linkage; Regression analysis.

1. INTRODUCTION

1.1 Application Setting

To model the energy economy properly, an economist might need company-specific microdata on the fuel and feedstocks used by companies that are only available from Agency A and corresponding microdata on the goods produced for companies that is only available from Agency B. To model the health of individuals in society, a demographer or health science policy worker might need individual-specific information on those receiving social benefits from Agencies B1, B2, and B3, corresponding income information from Agency I, and information on health services from Agencies H1 and H2. Such modeling is possible if analysts have access to the microdata and if unique, common identifiers are available (e.g., Oh and Scheuren 1975; Jabine and Scheuren 1986). If the only common identifiers are error-prone or nonunique or both, then probabilistic matching techniques (e.g., Newcombe, Kennedy, Axford and James 1959, Fellegi and Sunter 1969) are needed.

1.2 Relation to Earlier Work

In earlier work (Scheuren and Winkler 1993), we provided theory showing that elementary regression analyses could be accurately adjusted for matching error, employing knowledge of the quality of the matching. In that work we relied heavily on an error-rate estimation procedure of Belin and Rubin (1995). In later research e.g., (Winkler and Scheuren 1995, 1996), we showed that we could make further improvements by using noncommon quantitative data from the two files to improve matching

and adjust statistical analyses for matching error. The main requirement – even in heretofore seemingly impossible situations – was that there exist a reasonable model for the relationships among the noncommon quantitative data. In the empirical example of this paper, we use data for which a very small subset of pairs can be accurately matched using name and address information only and for which the noncommon quantitative data is at least moderately correlated. In other situations, researchers might have a small microdata set that accurately represents relationships of noncommon data across a set of large administrative files or they might just have a reasonable guess at what the relationships among the noncommon data are. We are not sure, but conjecture that, with a reasonable starting point, the methods discussed here will succeed often enough to be of general value.

1.3 Basic Approach

The intuitive underpinnings of our methods are based on now well-known probabilistic record linkage (RL) and edit/imputation (EI) technologies. The ideas of modern RL were introduced by Newcombe (Newcombe *et al.* 1959) and mathematically formalized by Fellegi and Sunter (1969). Recent methods are described in Winkler (1994, 1995). EI has traditionally been used to clean up erroneous data in files. The most pertinent methods are based on the EI model of Fellegi and Holt (1976).

To adjust a statistical analysis for matching error, we employ a four-step recursive approach that is very powerful. We begin with an enhanced RL approach (e.g., Winkler 1994, Belin and Rubin 1995) to delineate a subset of pairs of records in which the matching error rate is estimated to be very low. We perform a regression analysis, RA, on the

¹ Fritz Scheuren, Ernst and Young, 1225 Connecticut Avenue, N.W., Washington, DC 20036, U.S.A., Scheuren@aol.com; William E. Winkler, U.S. Bureau of the Census, Washington, DC 20023, U.S.A.

low-error-rate linked records and partially adjust the regression model on the remainder of the pairs by applying previous methods (Scheuren and Winkler 1993). Then, we refine the EI model using traditional outlier-detection methods to edit and impute outliers in the remainder of the linked pairs. Another regression analysis (RA) is done and this time the results are fed back into the linkage step so that the RL step can be improved (and so on). The cycle continues until the analytic results desired cease to change. Schematically, these *analytic linking* methods take the form



1.4 Structure of What Follows

Beginning with this introduction, the paper is divided into five sections. In the second section, we undertake a short review of Edit/Imputation (EI) and Record Linkage (RL) methods. Our purpose is not to describe them in detail but simply to set the stage for the present application. Because Regression Analysis (RA) is so well known, our treatment of it is covered only in the particular simulated application (Section 3). The intent of these simulations is to use matching scenarios that are more difficult than what most linkers typically encounter. Simultaneously, we employ quantitative data that is both easy to understand but hard to use in matching. In the fourth section, we present results. The final section consists of some conclusions and areas for future study.

2. EI AND RL METHODS REVIEWED

2.1 Edit/Imputation

Methods of editing microdata have traditionally dealt with logical inconsistencies in data bases. Software consisted of if-then-else rules that were data-base-specific and very difficult to maintain or modify, so as to keep current. Imputation methods were part of the set of if-then-else rules and could yield revised records that still failed edits. In a major theoretical advance that broke with prior statistical methods, Fellegi and Holt (1976) introduced operations-research-based methods that both provided a means of checking the logical consistency of an edit system and assured that an edit-failing record could always be updated with imputed values, so that the revised record satisfies all edits. An additional advantage of Fellegi and Holt (1976) systems is that their edit methods tie directly with current methods of imputing microdata (e.g., Little and Rubin 1987).

Although we will only consider continuous data in this paper, EI techniques also hold for discrete data and combinations of discrete and continuous data. In any event, suppose we have continuous data. In this case a collection of edits might consist of rules for each record of the form

$$c_1X < Y < c_2X$$

In words,

Y can be expected to be greater than c_1X and less than c_2X ; hence, if Y less than c_1X and greater than c_2X , then the data record should be reviewed (with resource and other practical considerations determining the actual bounds used).

Here Y may be total wages, X the number of employees, and c_1 and c_2 constants such that $c_1 < c_2$. When an (X, Y) pair associated with a record fails an edit, we may replace, say, Y with an estimate (or prediction).

2.2 Record Linkage

A record linkage process attempts to classify pairs in a product space $A \times B$ from two files A and B into M , the set of true links, and U , the set of true nonlinks. Making rigorous concepts introduced by Newcombe (e.g., Newcombe *et al.* 1959; Newcombe, Fair and Lalonde 1992), Fellegi and Sunter (1969) considered ratios R of probabilities of the form

$$R = \Pr((\gamma \in \Gamma \mid M) / \Pr((\gamma \in \Gamma \mid U))$$

where γ is an arbitrary agreement pattern in a comparison space Γ . For instance, Γ might consist of eight patterns representing simple agreement or not on surname, first name, and age. Alternatively, each $\gamma \in \Gamma$ might additionally account for the relative frequency with which specific surnames, such as Scheuren or Winkler, occur. The fields compared (surname, first name, age) are called *matching variables*. The decision rule is given by

If $R > \text{Upper}$, then designate pair as a link.

If $\text{Lower} \leq R \leq \text{Upper}$, then designate pair as a possible link and hold for clerical review.

If $R < \text{Lower}$, then designate pair as a nonlink.

Fellegi and Sunter (1969) showed that this decision rule is optimal in the sense that for any pair of fixed bounds on R , the middle region is minimized over all decision rules on the same comparison space Γ . The cutoff thresholds, *Upper* and *Lower*, are determined by the error bounds. We call the ratio R or any monotonely increasing transformation of it (typically a logarithm) a *matching weight* or *total agreement weight*.

With the availability of inexpensive computing power, there has been an outpouring of new work on record linkage techniques (e.g., Jaro 1989, Newcombe, *et al.* 1992, Winkler 1994, 1995). The new computer-intensive methods reduce, or even sometimes eliminate, the need for clerical review when name, address, and other information used in matching is of reasonable quality. The proceedings from a recently concluded international conference on record linkage showcase these ideas and might be the best single reference (Alvey and Jamerson 1997).

3. SIMULATION SETTING

3.1 Matching Scenarios

For our simulations, we considered a scenario in which matches are virtually indistinguishable from nonmatches. In our earlier work (Scheuren and Winkler 1993), we considered three matching scenarios in which matches are more easily distinguished from nonmatches than in the scenario of the present paper.

In both papers, the basic idea is to generate data having known distributional properties, adjoin the data to two files that would be matched, and then to evaluate the effect of increasing amounts of matching error on analyses. Because the methods of this paper work better than what we did earlier, we only consider a matching scenario that we label "Second Poor," because it is more difficult than the poor (most difficult) scenario we considered previously.

We started here with two population files (sizes 12,000 and 15,000), each having good matching information and for which true match status was known. Three settings were examined: high, medium and low – depending on the extent to which the smaller file had cases also included in the larger file. In the high file inclusion situation, about 10,000 cases are on both files for a file inclusion or intersection rate on the smaller or base file of about 83%. In the medium file intersection situation, we took a sample of one file so that the intersection of the two files being matched was approximately 25%. In the low file intersection situation, we took samples of both files so that the intersection of the files being matched was approximately 5%. The number of intersecting cases, obviously, bounds the number of true matches that can be found.

We then generated quantitative data with known distributional properties and adjoined the data to the files. These variations are described below and displayed in Figure 1 where we show the poor scenario (labeled "first poor") of our previous 1993 paper and the "second poor" scenario used in this paper. In the figure, the match weight, the logarithm of R , is plotted on the horizontal axis with the frequency, also expressed in logs, plotted on the vertical axis. Matches (or true links) appear as asterisks (*), while nonmatches (or true nonlinks) appear as small circles (o).

3.2 "First Poor Scenario" (Figure 1a)

The first poor matching scenario consisted of using last name, first name, one address variation, and age. Minor typographical errors were introduced independently into one fifth of the last names and one third of the first names in one of the files. Moderately severe typographical errors were made independently in one fourth of the addresses of the same file. Matching probabilities were chosen that deviated substantially from optimal. The intent was for the links to be made in a manner that a practitioner might choose after gaining only a little experience. The situation is analogous to that of using administrative lists of individuals where information used in matching is of poor quality. The true mismatch rate here was 10.1%.

3.3 "Second Poor" Scenario (Figure 1b)

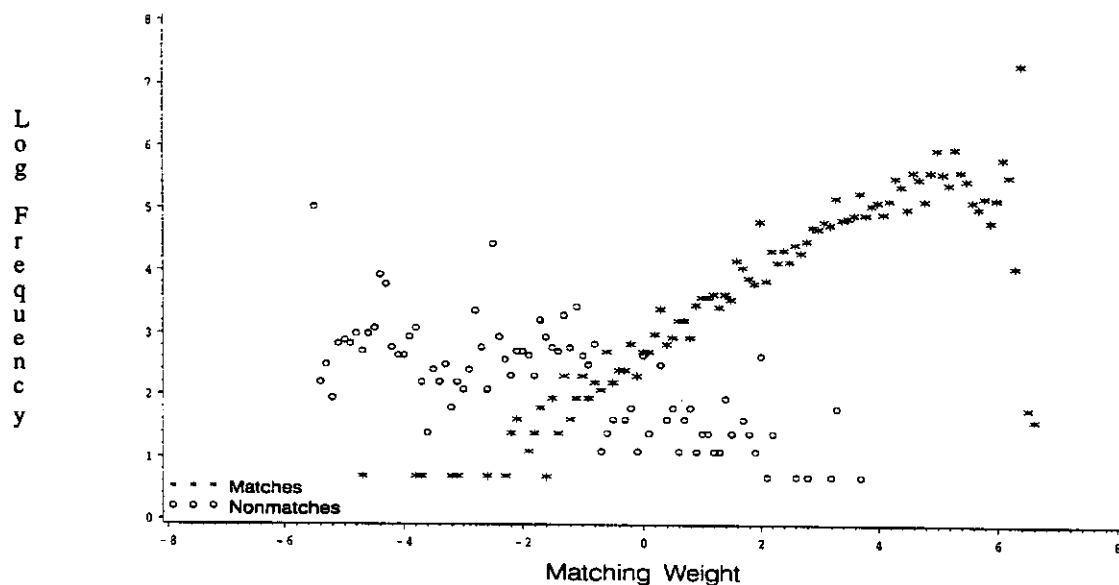
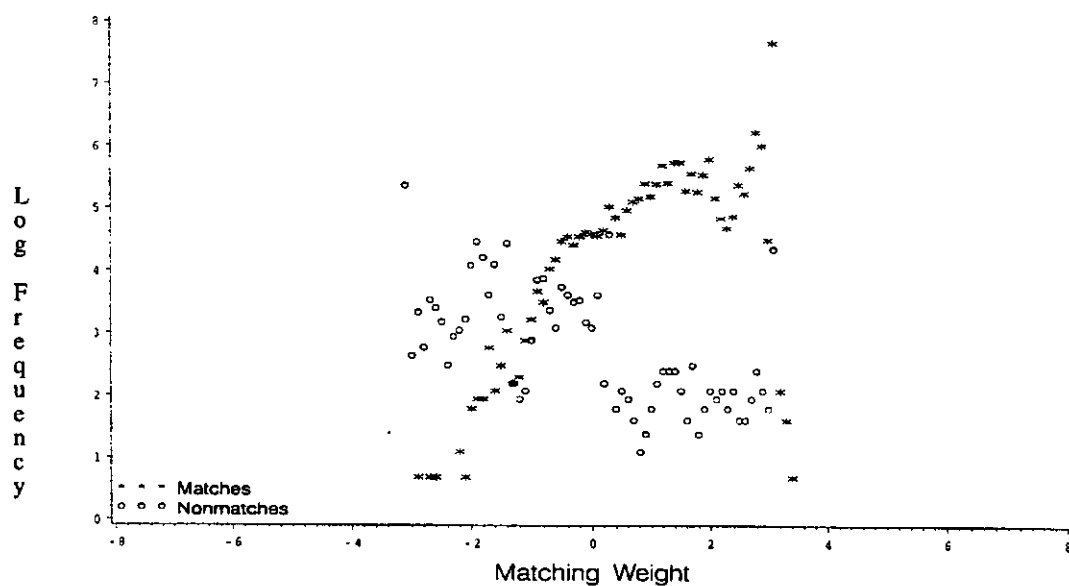
The second poor matching scenario consisted of using last name, first name, and one address variation. Minor typographical errors were introduced independently into one third of the last names and one third of the first names in one of the files. Severe typographical errors were made in one fourth of the addresses in the same file. Matching probabilities were chosen that deviated substantially from optimal. The intent was to represent situations that often occur with lists of businesses in which the linker has little control over the quality of the lists. Name information – a key identifying characteristic – is often very difficult to compare effectively with business lists. The true mismatch rate was 14.6%.

3.4 Summary of Matching Scenarios

Clearly, depending on the scenario, our ability to distinguish between true links and true nonlinks differs significantly. With the first poor scenario, the overlap, shown visually between the log-frequency-versus-weight curves, is substantial (Figure 1a); and, with the second poor scheme, the overlap of the log-frequency-versus-weight curves is almost total (Figure 1b). In the earlier work, we showed that our theoretical adjustment procedure worked well using the known true match rates in our data sets. For situations where the curves of true links and true nonlinks were reasonably well separated, we accurately estimated error rates via a procedure of Belin and Rubin (1995) and our procedure could be used in practice. In the poor matching scenario of that paper (first poor scenario of this paper), the Belin-Rubin procedure was unable to provide accurate estimates of error rates but our theoretical adjustment procedure still worked well. This indicated that we either had to find an enhancement to the Belin-Rubin procedures or to develop methods that used more of the available data. (That conclusion, incidentally, from our earlier work, after some false starts, to the present approach.)

3.5 Quantitative Scenarios

Having specified the above linkage situations, we used SAS to generate ordinary least squares data under the model $Y = 6X + \epsilon$. The X values were chosen to be uniformly distributed between 1 and 101. The error terms, are normal and homoscedastic with variances 13,000, 36,000, and 125,000, respectively. The resulting regressions of Y on X have R^2 values in the true matched population of 70%, 47%, and 20%, respectively. Matching with quantitative data is difficult because, for each record in one file, there are hundreds of records having quantitative values that are close to the record that is a true match. To make modeling and analysis even more difficult in the high file overlap scenario, we used all false matches and only 5% of the true matches; in the medium file overlap scenario, we used all false matches and only 25% of true matches. (Note: Here to heighten the visual effect, we have introduced another random sampling step, so the reader can "see"

Figure 1a. 1st Poor Matching ScenarioFigure 1b. 2nd Poor Matching Scenario

better in the figures the effect of bad matching. This sample depends on the match status of the case and is confined only to those cases that were matched, whether correctly or falsely.)

A crucial practical assumption for the work of this paper is that analysts are able to produce a reasonable model (guesstimate) for the relationships between the noncommon quantitative items. For the initial modeling in the empirical example of this paper, we use the subset of pairs for which matching weight is high and the error-rate is low. Thus, the number of false matches in the subset is kept to a minimum. Although neither the procedure of Belin and Rubin (1995) nor an alternative procedure of Winkler (1994), that requires an *ad hoc* intervention, could be used to estimate error rates, we believe it is possible for an experienced matcher to pick out a low-error-rate set of pairs even in the second poor scenario.

4. SIMULATION RESULTS

Most of this Section is devoted to presenting graphs and results of the overall process for the second poor scenario, where the R^2 value is moderate, and the intersection between the two files is high. These results best illustrate the procedures of this paper. At the end of the Section (in subsection 4.8), we summarize results over all R^2 situations and all overlaps. To make the modeling more difficult and show the power of the analytic linking methods, we use all false matches and a random sample of only 5% of the true matches. We only consider pairs having matching weight above a lower bound that we determine based on analytic considerations and experience. For the pairs of our analysis, the restriction causes the number of false matches to significantly exceed the number of true matches. (Again, this is done to heighten the visual effect of matching failures and to make the problem even more difficult.)

To illustrate the data situation and the modeling approach, we provide triples of plots. The first plot in the triple shows the true data situation as if each record in one file was linked with its true corresponding record in the other file. The quantitative data pairs correspond to the truth. In the second plot, we show the observed data. Where many of the pairs are in error because they correspond to false matches. To get to the third plot in the triple, we model using a small number of pairs (approximately 100) and then replace outliers with pairs in which the observed Y -value is replaced with a predicted Y -value.

4.1 Initial True Regression Relationship

In Figure 2a, the actual true regression relationship and related scatterplot are shown, for one of our simulations, as they would appear if there were no matching errors. In this figure and the remaining ones, the true regression line is always given for reference. Finally, the true population slope or β coefficient (at 5.85) and the R^2 value (at 43%) are provided for the data (sample of pairs) being displayed.

4.2 Regression After Initial RL-RA Step

In Figure 2b, we are looking at the regression on the actual observed links – not what should have happened in a perfect world but what did happen in a very imperfect one. Unsurprisingly, we see only a weak regression relationship between Y and X . The observed slope or β coefficient differs greatly from its true value (2.47 v. 5.85). The fit measure is similarly affected – falling to 7% from 43%.

4.3 Regression After First Combined RL-RA-EI-RA Step

Figure 2c completes our display of the first cycle of the iterative process we are employing. Here we have edited the data in the plot displayed as follows. First, using just the 99 cases with a match weight of 3.00 or larger, an attempt was made to improve the poor results given in Figure 2b. Using this provisional fit, predicted values were obtained for all the matched cases; then outliers with residuals of 460 or more were removed and the regression refit on the remaining pairs. This new equation, used in Figure 2c, was essentially $Y = 4.78X + \epsilon$, with a variance of 40,000. Using our earlier approach (Scheuren and Winkler 1993), a further adjustment was made in the estimated β coefficient from 4.78 to 5.4. If a pair of matched records yielded an outlier, then predicted values (not shown) using the equation $Y = 5.4X$ were imputed. If a pair does not yield an outlier, then the observed value was used as the predicted value.

4.4 Second True Reference Regression

Figure 3a displays a scatterplot of X and Y as they would appear if they could be true matches based on a second RL step. Note here that we have a somewhat different set of linked pairs this time from earlier, because we have used the regression results to help in the linkage. In particular, the second RL step employed the predicted Y values as determined above; hence it had more information on which to base a linkage. This meant that a different group of linked records was available after the second RL step. Since a considerably better link was obtained, there were fewer false matches; hence our sample of all false matches and 5% of the true matches dropped from 1,104 in Figures 2a through 2c to 650 for Figures 3a through 3c. In this second iteration, the true slope or β coefficient and the R^2 values remained, though, virtually identical for the estimated slope (5.85 v. 5.91) and fit (43% v. 48%).

4.5 Regression After Second RL-RA Step

In Figure 3b, we see a considerable improvement in the relationship between Y and X using the actual observed links after the second RL step. The estimated slope has risen from 2.47 initially to 4.75 here. Still too small but much improved. The fit has been similarly affected, rising from 7% to 33%.

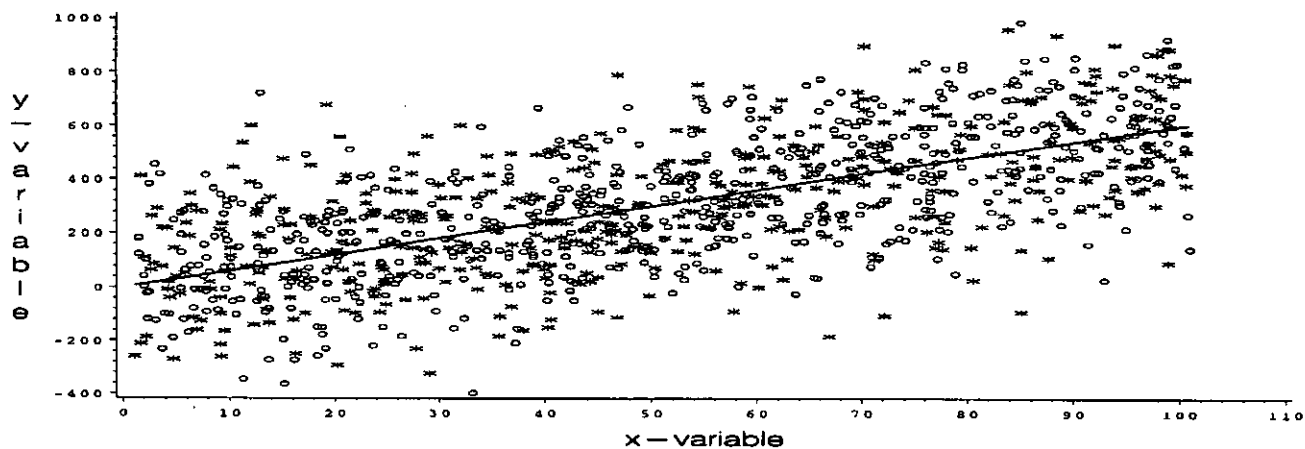


Figure 2a. 2nd Poor Scenario, 1st Pass
All False & 5 % True Matches, True Data, HighOverlap,
1104 Points, $\beta = 5.85$, $R\text{-square} = 0.43$

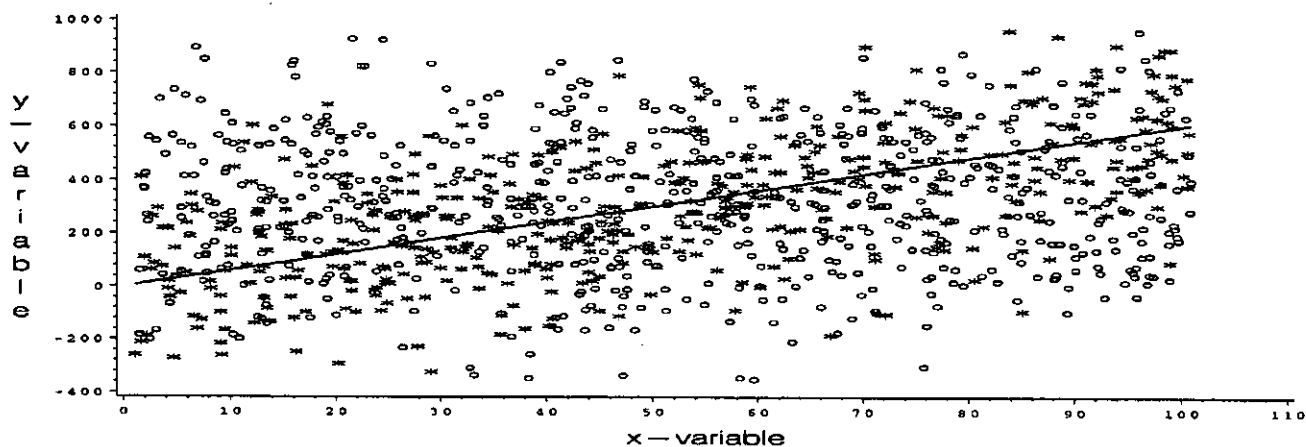


Figure 2b. 2nd Poor Scenario, 1st Pass
All False & 5 % True Matches, Observed Data, HighOverlap,
1104 Points, $\beta = 2.47$, $R\text{-square} = 0.07$

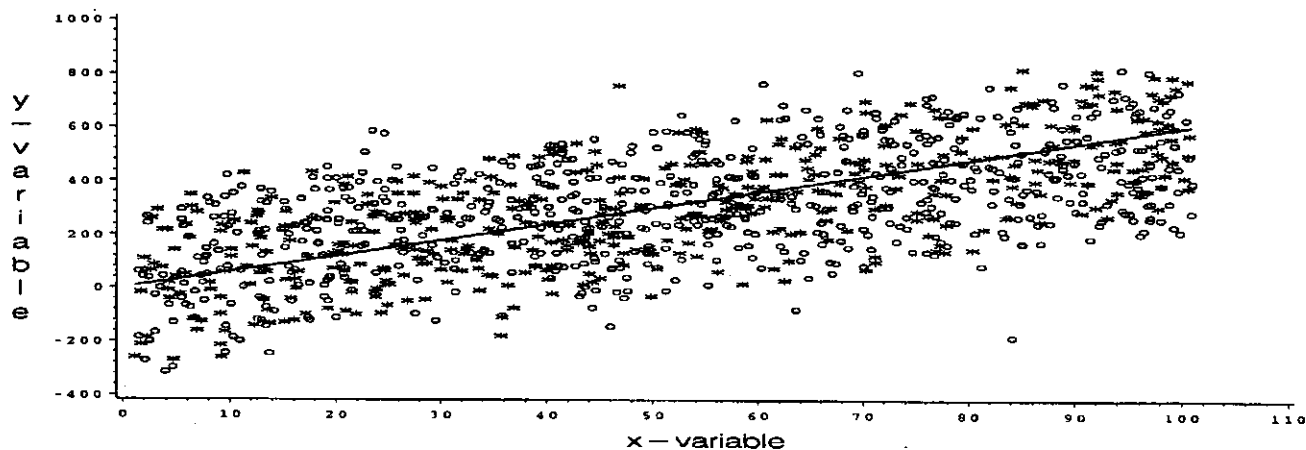


Figure 2c. 2nd Poor Scenario, 1st Pass
All False & 5 % True Matches, Outlier - Adjusted Data
1104 Points, $\beta = 4.78$, $R\text{-square} = 0.40$

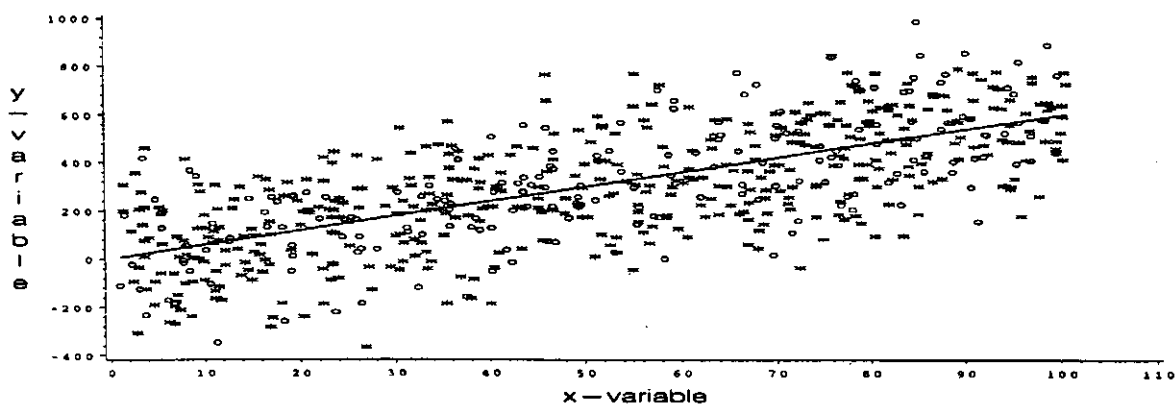


Figure 3a. 2nd Poor Scenario, 2nd Pass
All False & 5 % True Matches, True Data, HighOverlap,
650 Points, $\beta = 5.91$ R - square = 0.48

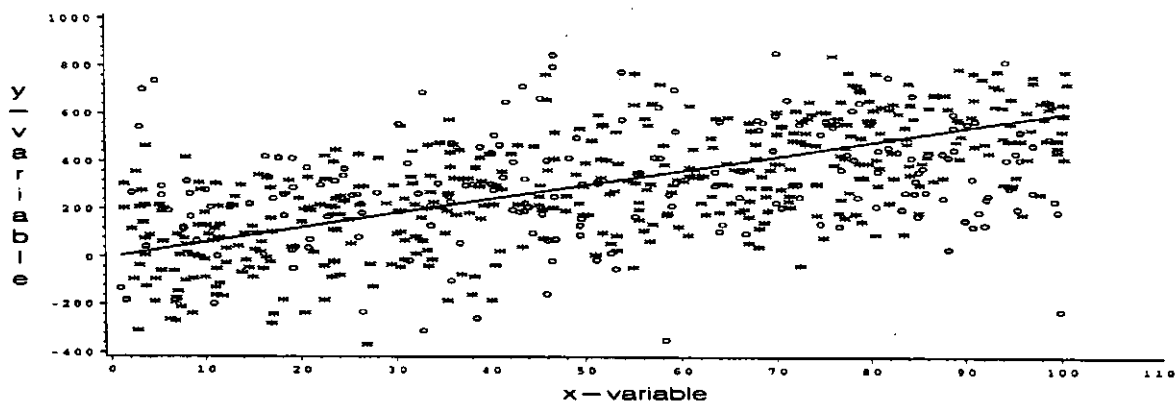


Figure 3b. 2nd Poor Scenario, 2nd Pass
All False & 5 % True Matches, Observed Data, HighOverlap
650 Points, $\beta = 4.75$, R - square = 0.33

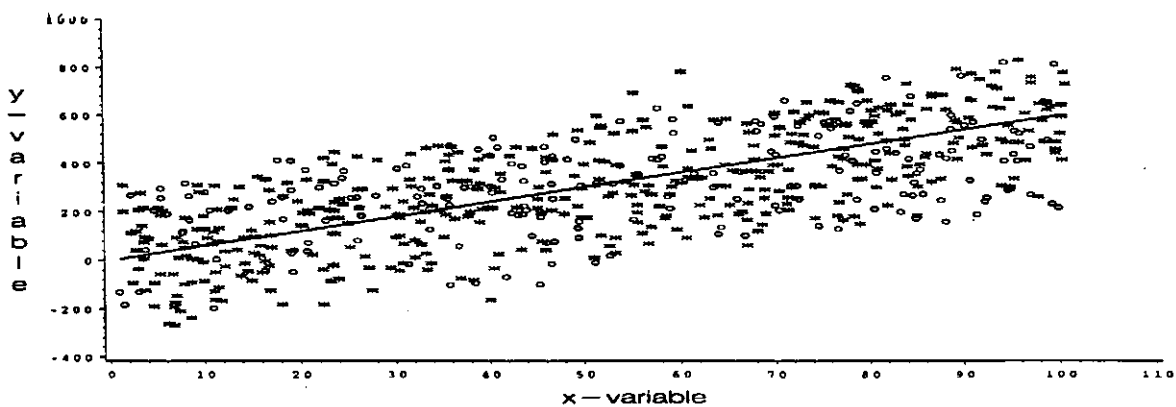


Figure 3c. 2nd Poor Scenario, 2nd Pass
All False & 5 % True Matches, Outlier - Adjusted Data
650 Points, $\beta = 5.26$, R - square = 0.47

4.6 Regression After Second Combined RL-RA-EI-RA Step

Figure 3c completes the display of the second cycle of our iterative process. Here we have edited the data as follows. Using the fit (from subsection 4.5), another set of predicted values was obtained for all the matched cases (as in subsection 4.3). This new equation was essentially $Y = 5.26X + \epsilon$, with a variance of about 35,000. If a pair of matched records yields an outlier, then predicted values using the equation $Y = 5.3X$ were imputed. If a pair does not yield an outlier, then the observed value was used as the predicted value.

4.7 Additional Iterations

While we did not show it in this paper, we did iterate through a third matching pass. The *beta* coefficient, after adjustment, did not change much. We do not conclude from this that asymptotic unbiasedness exists; rather that the method, as it has evolved so far, has a positive benefit and that this benefit may be quickly reached.

4.8 Further Results

Our further results are of two kinds. We looked first at what happened in the medium R^2 scenario (*i.e.*, R^2 equal to .47) for the medium- and low- file intersection situations. We further looked at the cases when R^2 was higher (at .70) or lower (at .20). For the medium R^2 scenario and low intersection case the matching was somewhat easier. This occurs because there were significantly fewer false-match candidates and we could more easily separate true matches from false matches. For the high R^2 scenarios, the modeling and matching were also more straightforward than they were for the medium R^2 scenario. Hence, there were no new issues there either.

On the other hand, for the low R^2 scenario, no matter what degree of file intersection existed, we were unable to distinguish true matches from false matches, even with the improved methods we are using. The reason for this, we believe, is that there are many outliers associated with the true matches. We can no longer assume, therefore, that a moderately higher percentage of the outliers in the regression model are due to false matches. In fact, with each true match that is associated with an outlier Y -value, there may be many false matches that have Y -values that are closer to the predicted Y -value than the true match.

5. COMMENTS AND FUTURE STUDY

5.1 Overall Summary

In this paper, we have looked at a very restricted analysis setting: a simple regression of one quantitative dependent variable from one file matched to a single quantitative independent variable from another file. This standard analysis was, however, approached in a very nonstandard setting. The matching scenarios, in fact, were quite

challenging. Indeed, just a few years ago, we might have said that the "second poor" matching scenario appeared hopeless.

On the other hand, as discussed below, there are many loose ends. Hence, the demonstration given here can be considered, quite rightly in our view, as a limited accomplishment. But make no mistake about it, we are doing something entirely new. In past record linkage applications, there was a clear separation between the identifying data and the analysis data. Here, we have used a regression analysis to improve the linkage and the improved linkage to improve the analysis and so on.

Earlier, in our 1993 paper, we advocated that there be a unified approach between the linkage and the analysis. At that point, though, we were only ready to propose that the linkage probabilities be used in the analysis to correct for the failures to complete the matching step satisfactorily. This paper is the first to propose a completely unified methodology and to demonstrate how it might be carried out.

5.2 Planned Application

We expect that the first applications of our new methods will be with large business data bases. In such situations, noncommon quantitative data are often moderately or highly correlated and the quantitative variables (both predicted and observed) can have great distinguishing power for linkage, especially when combined with name information and geographic information, such as a postal (*e.g.*, ZIP) code.

A second observation is also worth making about our results. The work done here points strongly to the need to improve some of the now routine practices for protecting public use files from reidentification. In fact, it turns out that in some settings – even after quantitative data have been confidentiality protected (by conventional methods) and without any directly identifying variables present – the methods in this paper can be successful in reidentifying a substantial fraction of records thought to be reasonably secure from this risk (as predicted in Scheuren 1995). For examples, see Winkler (1997).

5.3 Expected Extensions

What happens when our results are generalized to the multiple regression case? We are working on this now and results are starting to emerge which have given us insight into where further research is required. We speculate that the degree of underlying association R^2 will continue to be the dominant element in whether a usable analysis is possible.

There is also the case of multivariate regression. This problem is harder and will be more of a challenge. Simple multivariate extensions of the univariate comparison of Y values in this paper have not worked as well as we would like. For this setting, perhaps, variants and extensions of Little and Rubin (1987, Chapters 6 and 8) will prove to be a good starting point

5.4 "Limited Accomplishment"

Until now an analysis based on the second poor scenario would not have been even remotely sensible. For this reason alone we should be happy with our results. A closer examination, though, shows a number of places where the approach demonstrated is weaker than it needs to be or simply unfinished. For those who want theorems proven, this may be a particularly strong sentiment. For example, a convergence proof is among the important loose ends to be dealt with, even in the simple regression setting. A practical demonstration of our approach with more than two matched files also is necessary, albeit this appears to be more straightforward.

5.5 Guiding Practice

We have no ready advice for those who may attempt what we have done. Our own experience, at this point, is insufficient for us to offer ideas on how to guide practice, except the usual extra caution that goes with any new application. Maybe, after our own efforts and those of others have matured, we can offer more.

REFERENCES

- ALVEY, W., and JAMERSON, B. (Eds.) (1997). *Record Linkage Techniques - 1997*. Proceedings of An International Record Linkage Workshop and Exposition, March 20-21, 1997, Arlington, VA.
- BELIN, T.R., and RUBIN, D.B. (1995). A method for calibrating false-match rates in record linkage. *Journal of the American Statistical Association*, 90, 694-707.
- FELLEGI, I., and HOLT, T. (1976). A systematic approach to automatic edit and imputation. *Journal of the American Statistical Association*, 71, 17-35.
- FELLEGI, I., and SUNTER, A. (1969). A theory of record linkage. *Journal of the American Statistical Association*, 64, 1183-1210.
- JABINE, T.B., and SCHEUREN, F. (1986). Record linkages for statistical purposes: Methodological issues. *Journal of Official Statistics*, 2, 255-277.
- JARO, M.A. (1989). Advances in record-linkage methodology as applied to matching the 1985 census of Tampa, Florida. *Journal of the American Statistical Association*, 89, 414-420.
- LITTLE, R.J.A., and RUBIN, D.B. (1987). *Statistical Analysis With Missing Data*. New York: John Wiley.
- NEWCOMBE, H.B., KENNEDY, J.M., AXFORD, S.J., and JAMES, A.P. (1959). Automatic linkage of vital records. *Science*, 130, 954-959.
- NEWCOMBE, H., FAIR, M., and LALONDE, P. (1992). The use of names for linking personal records. *Journal of the American Statistical Association*, 87, 1193-1208.
- OH, H.L., and SCHEUREN, F. (1975). Fiddling around with mismatches and nonmatches. *Proceedings of the Social Statistics Section, American Statistical Association*.
- SCHEUREN, F. (1995). Review of private lives and public policies: Confidentiality and accessibility of government services. *Journal of the American Statistical Association*, 90, 386-387.
- SCHEUREN, F., and WINKLER, W.E. (1993). Regression analysis of data files that are computer matched. *Survey Methodology*, 19, 39-58.
- WINKLER, W.E. (1994). Advanced methods of record linkage. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 467-472.
- WINKLER, W.E. (1995). Matching and record linkage. *Business Survey Methods*, (Eds. B.G. Cox et al.). New York: John Wiley, 355-384.
- WINKLER, W.E., and SCHEUREN, F. (1995). Linking data to create information. *Proceedings: Symposium 95, From Data to Information-Methods and Systems*, Statistics Canada, 29-37.
- WINKLER, W.E., and SCHEUREN, F. (1996). Recursive analysis of linked data files. *Proceedings of the 1996 Annual Research Conference*. U.S. Bureau of the Census.
- WINKLER, W.E. (1997). Producing Public-Use Microdata That are Analytically Valid and Confidential. Presented at the 1997 Joint Statistical Meetings, Anaheim, CA.

ACKNOWLEDGEMENTS

Survey Methodology wishes to thank the following persons who have served as referees during 1997. An asterisk indicates that the person served more than once.

- J.C. Arnold, *Virginia Polytechnic Institute*
M. Bankier, *Statistics Canada*
* D.R. Bellhouse, *University of Western Ontario*
* T.R. Belin, *University of California - Los Angeles*
* D.A. Binder, *Statistics Canada*
G.J. Brackstone, *Statistics Canada*
F.J. Breidt, *Iowa State University*
A. Brinkley, *U.S. Bureau of the Census*
L. Cahoon, *U.S. Bureau of the Census*
N. Caron, *Institut national de la statistique et des études économiques*
R. Caspar, *Research Triangle Institute*
R. Chambers, *University of Southampton*
S.X. Chen, *New York University*
G.H. Choudhry, *Statistics Canada*
W. Davis, *Klemm Analysis Group*
* J. Denis, *Statistics Canada*
J.-C. Deville, *Institut national de la statistique et des études économiques*
* P. Dick, *Statistics Canada*
J.D. Drew, *Statistics Canada*
D.F. Findlay, *U.S. Bureau of the Census*
B. Forsyth, *Westat, Inc.*
L.A. Franklin, *Indiana State University*
W.A. Fuller, *Iowa State University*
J. Gambino, *Statistics Canada*
G. Gates, *U.S. Bureau of the Census*
B.V. Greenberg, *U.S. Bureau of the Census*
* R.M. Groves, *University of Maryland*
J.-P. Gwet, *Westat, Inc.*
* M.A. Hidirolou, *Statistics Canada*
D. Holt, *Central Statistical Office, U.K.*
C. Julien, *Statistics Canada*
* G. Kalton, *Westat, Inc.*
S. Kaufman, *National Center for Education Statistics*
D. Kerr, *Statistics Canada*
J.J. Kim, *U.S. Bureau of the Census*
P. Kokic, *University of Southampton*
M. Kovacevic, *Statistics Canada*
R. Lachapelle, *Statistics Canada*
M. Latouche, *Statistics Canada*
* P. Lavallée, *Statistics Canada*
J. Ledent, *Université de Québec*
S. Linacre, *Australian Bureau of Statistics*
R. Little, *University of Michigan*
D. Malec, *National Center for Health Statistics*
* H. Mantel, *Statistics Canada*
N. Mathiowetz, *University of Maryland*
C. Moriarity, *National Center for Health Statistics*
* B. Nandram, *Worcester Polytechnic Institute*
G. Nathan, *Central Bureau of Statistics, Israel*
D. Pfeffermann, *Hebrew University*
* B. Quenneville, *Statistics Canada*
T.E. Raghunathan, *University of Michigan*
E. Rancourt, *Statistics Canada*
* J.N.K. Rao, *Carleton University*
* L.-P. Rivest, *Université Laval*
G. Roberts, *Statistics Canada*
* I. Sande, *Bell Communications Research, U.S.A.*
G. Sande, *Sande & Assoc.*
F.J. Scheuren, *George Washington University*
* J. Sedransk, *Case Western Reserve University*
J. Shao, *University of Wisconsin - Madison*
* A.C. Singh, *Statistics Canada*
* M.P. Singh, *Statistics Canada*
B.K. Sinha, *University of Maryland*
* R. Sitter, *Simon Fraser University*
C.J. Skinner, *University of Southampton*
G. Smith, *Statistics Canada*
P. Steel, *U.S. Bureau of the Census*
* D. Stukel, *Statistics Canada*
W. Sun, *Statistics Canada*
J.-L. Tambay, *Statistics Canada*
A. Théberge, *Statistics Canada*
* R. Thomas, *Carleton University*
M. Thompson, *University of Waterloo*
I. Thomsen, *Statistics Norway*
Y. Tillé, *École nationale de statistique et de l'analyse de l'information*
R. Valliant, *U.S. Bureau of Labor Statistics*
V.K. Verma, *University of Essex*
P.J. Waite, *U.S. Bureau of the Census*
J. Waksberg, *Westat, Inc.*
K.M. Wolter, *National Opinion Research Center*
F. Yu, *Australian Bureau of Statistics*
M. Yu, *Statistics Canada*
* A. Zaslavsky, *Harvard University*

Acknowledgements are also due to those who assisted during the production of the 1997 issues: S. Beauchamp and L. Durocher (Composition Unit) and L. Perreault (Official Languages and Translation Division). Finally we wish to acknowledge D. Blair, S. DiLoreto, C. Larabie and D. Lemire of Household Survey Methods Division, for their support with coordination, typing and copy editing.

CONTENTS

TABLE DES MATIÈRES

Volume 25, No. 4, December/décembre 1997

Christian GENEST

Statistics on statistics: measuring research productivity by journal publications between 1985 and 1995

Debajyoti SINHA

Time-discrete beta process model for interval-censored survival data

Lynn KUO and Bani MALLICK

Bayesian semiparametric inference for the accelerated failure time model

Stephen G. WALKER and Bani K. MALLICK

A note on the scale parameter of the Dirichlet process

Nancy HECKMAN and John RICE

Line transects of two dimensional random fields: Estimation and design

Fulvio DE SANTIS and Fulvio SPEZZAFERRI

Alternative Bayes factors for model selection

Gemai CHEN and Richard A. LOCKHART

Box-Cox transformed linear models: A parameter based asymptotic approach

Holger DETTE

E-optimal designs for regression models with quantitative factors - a reasonable choice?

Jeesen CHEN

A general lower bound of minimax risk for absolute error loss

Yodit SEIFU and N. REID

Applications of bivariate and univariate local Lyapunov exponents

Robert TIBSHIRANI and Donald A. REDELMEIER

Cellular telephones and motor vehicle collisions: some variations on matched pairs analysis

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 13, Number 4, 1997

A Sampling Scheme With Partial Replacement <i>J.L. Sánchez-Crespo</i>	327
Sources of Error in a Survey on Sexual Behavior <i>R. Tourangeau, K. Rasinski, J.B. Jobe, T.W. Smith, and W.F. Pratt</i>	341
Developing an Estimation Strategy for a Pesticide Data Program <i>Phillip S. Kott and D. Andrew Carr</i>	367
Estimating Interpolated Percentiles from Grouped Data with Large Samples <i>Edward L. Korn, Douglas Midthune, and Barry I. Graubard</i>	385
Ratio Estimation of Hardcore Drug Use <i>Doug Wright, Joe Gfroerer, and Joan Epstein</i>	401
Statistical Disclosure Control and Sampling Weights <i>A.G. de Waal and L.C.R.J. Willenborg</i>	417
Book Reviews	435
Editorial Collaborators	447
Index to Volume 13, 1997	449

All inquiries about submissions and subscriptions should be directed to the Chief Editor:
Lars Lyberg, R&D Department, Statistics Sweden, Box 24 300, S - 104 51 Stockholm, Sweden.

DIRECTIVES CONCERNANT LA PRÉSENTATION DES TEXTES

Avant de dactylographier votre texte pour le soumettre, prière d'examiner un numéro récent de *Techniques d'enquête* (à partir du vol. 19, n° 1) et de noter les points suivants:

1. Présentation

- 1.1 Les textes doivent être dactylographiés sur un papier blanc de format standard (8½ par 11 pouces), sur une face seulement, à double interligne partout et avec des marges d'au moins 1½ pouce tout autour.
- 1.2 Les textes doivent être divisés en sections numérotées portant des titres appropriés.
- 1.3 Le nom et l'adresse de chaque auteur doivent figurer dans une note au bas de la première page du texte.
- 1.4 Les remerciements doivent paraître à la fin du texte.
- 1.5 Toute annexe doit suivre les remerciements mais précéder la bibliographie.

2. Résumé

Le texte doit commencer par un résumé composé d'un paragraphe suivi de trois à six mots clés. Éviter les expressions mathématiques dans le résumé.

3. Rédaction

- 3.1 Éviter les notes au bas des pages, les abréviations et les sigles.
- 3.2 Les symboles mathématiques seront imprimés en italique à moins d'une indication contraire, sauf pour les symboles fonctionnels comme $\exp(\cdot)$ et $\log(\cdot)$ etc.
- 3.3 Les formules courtes doivent figurer dans le texte principal, mais tous les caractères dans le texte doivent correspondre à un espace simple. Les équations longues et importantes doivent être séparées du texte principal et numérotées en ordre consécutif par un chiffre arabe à la droite si l'auteur y fait référence plus loin.
- 3.4 Écrire les fractions dans le texte à l'aide d'une barre oblique.
- 3.5 Distinguer clairement les caractères ambigus (comme w , ω ; o , O , 0 ; l , 1).
- 3.6 Les caractères italiques sont utilisés pour faire ressortir des mots. Indiquer ce qui doit être imprimé en italique en le soulignant dans le texte.

4. Figures et tableaux

- 4.1 Les figures et les tableaux doivent tous être numérotés en ordre consécutif avec des chiffres arabes et porter un titre aussi explicatif que possible (au bas des figures et en haut des tableaux).
- 4.2 Ils doivent paraître sur des pages séparées et porter une indication de l'endroit où ils doivent figurer dans le texte. (Normalement, ils doivent être insérés près du passage qui y fait référence pour la première fois).

5. Bibliographie

- 5.1 Les références à d'autres travaux faites dans le texte doivent préciser le nom des auteurs et la date de publication. Si une partie d'un document est citée, indiquer laquelle après la référence.
Exemple: Cochran (1977, p. 164).
- 5.2 La bibliographie à la fin d'un texte doit être en ordre alphabétique et les titres d'un même auteur doivent être en ordre chronologique. Distinguer les publications d'un même auteur et d'une même année en ajoutant les lettres a, b, c, etc. à l'année de publication. Les titres de revues doivent être écrits au long. Suivre le modèle utilisé dans les numéros récents.

