# SURVEY METHODOLOGY 



JUNE 2001
-
VOLUME 27
$\bullet$
NUMBER 1

# SURVEY METHODOLOGY 

## A JOURNAL PUBLISHED BY STATISTICS CANADA

JUNE 2001 • VOLUME 27 • NUMBER 1

Published by authority of the Minister
responsible for Statistics Canada
© Minister of Industry, 2001
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.
August 2001
Catalogue no. 12-001-XPB
Frequency: Semi-annual
ISSN 0714-0045
Ottawa

## SURVEY METHODOLOGY

## A Journal Published by Statistics Canada

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

## MANAGEMENT BOARD

Chairman G.J. Brackstone

| Members | D.A. Binder R. Platek (Past Chairman) <br>  G.J.C. Hole <br>  E. Rancourt (Production Manager) | D. Roy <br>  <br>  <br> C. Patrick |
| :--- | :--- | :--- |
| M.P. Singh |  |  |

EDITORIAL BOARD
Editor M.P. Singh, Statistics Canada

## Associate Editors

D.R. Bellhouse, University of Western Ontario<br>G. Nathan, Hebrew University, Israel<br>D.A. Binder, Statistics Canada<br>D. Norris, Statistics Canada<br>C. Clark, U.S. Bureau of the Census<br>J.-C. Deville, INSEE<br>J. Eltinge, U.S. Bureau of Labor Statistics<br>W.A. Fuller, Iowa State University<br>J. Gambino, Statistics Canada<br>M.A. Hidiroglou, Statistics Canada<br>D. Holt, University of Southampton, U.K.<br>G. Kalton, Westat, Inc.<br>P. Kott, National Agricultural Statistics Service<br>P. Lahiri, University of Nebraska-Lincoln<br>S. Linacre, Australian Bureau of Statistics<br>D. Pfeffermann, Hebrew University<br>J.N.K. Rao, Carleton University<br>L.-P. Rivest, Université Laval<br>F.J. Scheuren, The Urban Institute<br>R. Sitter, Simon Fraser University<br>C.J. Skinner, University of Southampton<br>E. Stasny, Ohio State University<br>R. Valliant, Westat, Inc.<br>J. Waksberg, Westat, Inc.<br>K.M. Wolter, National Opinion Research Center<br>A. Zaslavsky, Harvard University

Assistant Editors J.-F. Beaumont, P. Dick, H. Mantel and W. Yung, 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 CDN $\$ 47$ per year. The price excludes Canadian sales taxes. Additional shipping charges apply for delivery outside Canada: United States, CDN $\$ 12$ ( $\$ 6 \times 2$ issues); Other Countries, CDN $\$ 20$ ( $\$ 10 \times 2$ issues). Subscription order should be sent to Statistics Canada, Dissemination Division, Circulation Management, 120 Parkdale Avenue, Ottawa, Ontario, Canada K1A 0T6 or by dialling $1800700-1033$, by fax $1800889-9734$ or by E-mail: order@statcan.ca. A reduced price is available to members of the American Statistical Association, the International Association of Survey Statisticians, the American Association for Public Opinion Research, the Statistical Society of Canada and l'Association des statisticiennes et staticiens du Québec.

## SURVEY METHODOLOGY

## A Journal Published by Statistics Canada <br> Volume 27, Number 1, June 2001

## CONTENTS

In This Issue ..... 1
Waksberg Invited Paper Series
G. NATHAN
Telesurvey Methodologies for Household Surveys - A Review and Some Thoughts for the Future? ..... 7
Special Section on Composite Estimation
A.C. SINGH, B. KENNEDY and S. WU
Regression Composite Estimation for the Canadian Labour Force Survey with a Rotating Panel Design ..... 33
W.A. FULLER and J.N.K. RAO
A Regression Composite Estimator with Application to the Canadian Labour Force Survey ..... 45
P. BELL
Comparison of Alternative Labour Force Survey Estimators ..... 53
J. GAMBINO, B. KENNEDY and M.P. SINGH
Regression Composite Estimation for the Canadian Labour Force Survey: Evaluation and Implementation ..... 65
Regular Papers
J.-K. KIM
Variance Estimation After Imputation ..... 75
T.E. RAGHUNATHAN, J.M. LEPKOWSKI, J. VAN HOEWYK and P. SOLENBERGER
A Multivariate Technique for Multiply Imputing Missing Values Using a Sequence of Regression Models ..... 85
J. DUFOUR, F. GAGNON, Y. MORIN, M. RENAUD and C.-E. SÄRNDAL
A Better Understanding of Weight Transformation Through a Measure of Change ..... 97
P. ARDILLY and D. LE BLANC
Sampling and Weighting a Survey of Homeless Persons: A French Example ..... 109

## In This Issue

This issue of Survey Methodology contains the first in an annual invited paper series in honour of Joseph Waksberg. A brief description of the newly instituted series and a biography of Joseph Waksberg are given before the paper itself. I would like to thank Danny Levine for writing the biography of Joseph Waksberg. I would also like to thank David Binder, Paul Biemer, Graham Kalton, and Chris Skinner, the current members of the Committee for choosing a very prominent survey researcher to author the first paper of the Waksberg Invited Paper Series. My special thanks are due to Graham Kalton who, as the founding Chairman of the Committee, took the lead, negotiated the necessary arrangements with Westat, the American Statistical Association and Survey Methodology to set the wheel in motion and worked hard to meet the deadline set by the journal for publication of the June Issue.

The author of the Waksberg Invited Paper for 2001 is Gad Nathan. His paper, "Telesurvey Methodologies for Household Surveys - A Review and Some Thoughts for the Future", presents a methodological history of telephone surveys from the 1930s up to the present day. Topics covered include sampling designs, sampling frames, coverage, nonresponse and weighting. He finishes the paper by describing some of the challenges and opportunities posed by more recent developments such as email, the intemet, cell phones, and other emerging technological and social changes.

This issue of Survey Methodology also includes a special section on composite estimation with four papers. The first of these papers, by A.C. Singh, Kennedy and Wu, describes the method of regression composite estimation developed by Singh and colleagues over the past few years. They compare the new approach to previous methods of composite estimation, most notably the $K$ composite and the $A K$-composite estimators. The paper also includes a heuristic description and motivation of the new approach. Advantages of the new approach are that it yields a single set of estimation weights, leading to internal consistency of estimates, while improving on the efficiency of conventional regression estimators.

Fuller and Rao give an analytical evaluation of the properties of regression composite estimation. They first describe two earlier variants of regression composite estimation called modified regression estimators (MR1 and MR2), and analyse the efficiency and behaviour of the estimates over time using a simple time series model for the survey panel estimates. They conclude that a modification which can be viewed as a compromise between $M R 1$ and $M R 2$ would have the best properties overall.

In his paper, Bell compares a range of alternative estmators for use in the Australian Labour Force Survey. Estimators considered include the $A K$-composite estimator, the early variant of regression composite estimation called MR2, Fuller and Rao's variant of regression composite estimation, and a BLUE estimator chosen as an "optimal" linear combination of panel estimates. An improved BLUE, obtained by calibrating the BLUE estimator to some population benchmarks, is also proposed. These estimators are compared in terms of their differences from the conventional regression estimator, their standard errors, and their usefulness for seasonal adjustment and trend estimation.

The final paper of the special section, by Gambino, Kennedy and M.P. Singh, describes the regression composite estimator that was implemented for the Canadian Labour Force Survey. This estimator is based on the work of A.C. Singh and colleagues and the compromise suggested by Fuller and Rao. The new estimators are compared to the previously used regression type estimators for a number of series. They find that the new estimators are usually more efficient and stable, and more often allow succesful seasonal adjustment of the estimate series.

Kim proposes a new method for variance estimation that accounts for random imputation based on a linear regression imputation model. The method is based on creating a set of pseudo-values for $y$, such that a conventional variance estimator based on these pseudo-values also accounts for the imputation. Calculation of the pseudo-values is described first for simple random sampling and then for complex designs. The approach is shown to be asymptotically equivalent to the adjusted jackknife of Rao and Sitter, and properties are investigated in a simulation study.

Raghunathan, Lepkowski, Van Hoewyk and Solenberger in "A Multivariate Technique for Multiply Imputing Missing Values Using a Sequence of Regression Models" address the important issue of imputing into a complex data structure where explicit full multivariate models cannot be easily constructed. They adopt the approach of imputing on a variable by variable basis conditioned on all the observed variables. This implies that the imputations are created through a sequence of multiple regressions that vary depending on the type of variable being imputed.

In their article, Dufour, Gagnon, Morin, Renaud and Särndal propose a measurement of distance which can be used to measure the relative incidence of the nonresponse adjustment, calibration and the interaction between these two procedures. This measurement enables them to study and measure the change (from the initial to the final weight) resulting from the weight modification procedure. They use this measurement as a tool to compare the effectiveness of various non-response adjustment methods through a simulation study applied to the data from the Survey of Labour and Income Dynamics. The measurement is also applied to data from the National Longitudinal Survey of Children and Youth.

In recent years there has been an increasing number of attempts to survey homeless people in major cities. The difficulty of constructing a reliable and efficient survey frame and sampling method, and the fluidity of the population over time make surveying of this population particularly difficult. The final paper of this issue, by Ardilly and Le Blanc, describes sampling and estimation for a current survey of homelessness in France. Problems and challenges particular to this type of survey are also described. The proposed survey will sample homeless individuals indirectly by sampling the services such as shelters and meal services which they may use. The weight-share method is shown to be an effective way to obtain unbiased weights for different periods of time such as an average day or an average week.

Finally, I would like to take this opportunity to express my sincere thanks to Frank Mayda, Production Manager of Survey Methodology, who recently retired. His involvment with Survey Methodology since 1987 has been invaluable. I would also like to announce that Eric Rancourt has replaced Frank Mayda as Production Manager.

M.P. Singh

## Waksberg Invited Paper Series


#### Abstract

Survey Methodology has established an annual invited paper series in honor of Joseph Waksberg, who has made many important contributions to survey methodology. Each year, a prominent survey researcher will be chosen to author a paper that will review the development and current state of a significant topic in the field of survey methodology. The author receives a cash award, made possible through a grant from Westat in recognition of Joe Waksberg's contributions during his many years of association with Westat. The grant is administered financially and managed by the American Statistical Association. The author of the paper is selected by a four-person committee appointed by Survey Methodology and the American Statistical Association.




JOSEPH WAKSBERG

Joseph Waksberg (known universally as "Joe") currently is Chair of the Board of Directors of Westat, a statistical research firm located in Rockville, MD. Throughout a career that now spans more than 60 years, he has made important contributions to sampling theory, developed innovative applications of the theory, and conducted research in a broad array of survey methodology issues. He is author or co-author of numerous papers on sampling methods, including random digit dialing, sampling for rare populations, sampling for panel and rotating design surveys, and the role of sampling in population censuses. Additional contributions have ranged from methodological research on labor force measurement, evaluation of the quality of U.S. censuses, the effects of telescoping and other problems of recall on survey results, research on the effects of cash incentives on response rates and survey costs, small area estimation, and the development of models to estimate
election night results. His goal has been to improve both survey theory and practice. Last, but not least, he has been teacher and mentor to generations of statisticians.

Born in Kielce, Poland in September 1915, Joe immigrated with his family to the United States in 1921. Shortly after graduating from the City University of New York (CUNY) in 1936 with a degree in mathematics, he moved to the Washington D.C. area and, after a brief stint with the Navy Department, joined the Census Bureau in 1940 as a clerk. He remained at the Census Bureau for 33 years, retiring in 1973 as Associate Director for Statistical Methods, Research, and Standards. In the early 1960's, Waksberg, in association with Neter, initiated a classic study on the magnitude of various types of memory recall problems. This landmark effort led to procedures for reducing the effects of recall problems through both an innovative sampling and data collection approach (Neter
and Waksberg 1964; Neter and Waksberg 1965). Joe's interest in this area has continued; for example, he helped design and analyze results from an experiment to measure the direction and magnitude of possible biases from a one year recall survey for the U.S. Fish and Wildlife Service (Chu, Eisenhower, Hay, Morganstein, Neter and Waksberg 1992). The results of that experiment had a substantial effect on the redesign of the survey. More importantly, the work also added significantly to knowledge about respondent bias when respondents are asked to recall the frequency of activities under varying recall periods, and indicated methods of minimizing the mean square errors in the design of such surveys.

The current stature of the U.S. Current Population Survey (CPS) as a model of statistical efficiency fully reflects his influence and contributions while in charge of sampling, statistical standards, and research for the Census Bureau's household survey program. Notable among the changes introduced during his tenure which bear his imprint are the improved methods of sample selection and estimation, including the use of list sampling, replication variances, determination of appropriate cluster size, treatment of rare events, and composite estimation. At the same time, he played a major role in the experimental research carried out on alternative rotation and estimation patterns, on the use of a single household respondent, and on the effects of variable recall periods on labor force measurement.

No discussion of Joe's stay at the Census Bureau is complete without some reference to his many contributions to the decennial census programs. A good example is the evaluation program for the 1970 Census, which Waksberg developed, designed, and directed. Consisting of a series of 25 separate projects, it was considered at that time as "radical"; today that program stands as the model for ongoing programs of decennial census research. When early field returns in the 1970 Census showed a serious overstatement in the reporting of "vacant" units, Waksberg designed, developed, and implemented, under great time constraints, an innovative sample survey program which revisited a sample of vacant units to estimate the proportion occupied. An adjustment procedure was then developed and applied, at the small area level, to the universe of vacant units identified in the census (Waksberg 1998). Subsequently, with the introduction of Revenue Sharing legislation in 1972, with its requirement that the Bureau produce annual estimates of population and per capita income for all 39,000 governmental units in the U.S., Waksberg proposed using administrative records in concert with survey data to provide the required local area estimates of population and per capita income. He initiated research on matching IRS records for adjacent years in order to obtain small-area (county) estimates of gross and net migration and changes in income levels, research that led to the development and implementation of a small area estimation program that is basically still in use today.

Waksberg's years at Westat, which began in 1973, first as Senior Statistician and Vice President, and recently as in-house consultant and Chair of the Board, have shown the same dedication to innovation, experimentation, and quality in meeting the needs of its clients and in developing samples and carrying out survey research. In assisting the National Center for Health Statistics in designing samples for both the National Health Interview Survey and the National Health and Nutrition Examination Survey, he made major contributions to innovative methods for efficient oversampling of minority populations, by following up work he had done earlier on this subject (Wasksberg 1973). His work with Judkins and Massey provides important information on residential concentrations by race and ethnic origin, essential to assessing the usefulness of oversampling geographical areas for minority populations, and persons in poverty, another subpopulation for which oversampling is often required (Waksberg, Judkins and Massey 1997). He was a co-developer of the MitofskyWaksberg method of two-stage sampling of telephone households (Waksberg 1978), which became the standard approach for RDD sampling in the United States. Waksberg continued to explore ways of improving RDD sampling by examining the bias from list-assisted samples (Waksberg 1983; Brick and Waksberg 1991), which have resulted in modifications and improved efficiencies of the method and, subsequently, to a completely different method of RDD sampling (Brick, Waksberg, Kulp and Starer 1995). More recently, he participated in an examination of alternative ways of adjusting for households lacking telephones (Brick, Waksberg and Keeter 1996). His work in RDD sampling clearly demonstrates his life-long desire to constantly reexamine statistical approaches and find new methods to improve upon or even replace the standards, including those he helped establish.

Mr. Waksberg has shared his knowledge and expertise in a wide range of venues outside his office. For many years, he taught at the Graduate School of the U.S. Department of Agriculture, and was a regular lecturer at the University of Michigan summer program in sampling methods. He also has been a frequent consultant on sampling and survey techniques to governmental statistical organizations throughout the world, through the sponsorship of the U.S. Agency for International Development and the United Nations, as well at the request of individual countries, and has provided advice to the statistical offices of China, Argentina, Brazil, Cuba, Venezuela, Turkey, and South Vietnam. He has also represented the United States at international statistical meetings, served as technical expert under UN auspices, and been a member of a team sent to South America by the American Statistical Association to coordinate activities of their national statistical societies.

He is a member of the American Statistical Association, of which he has been elected Fellow, the International Association of Survey Statisticians, and the International

Statistical Institute, and has served as a member of various panels of the National Academy of Sciences to evaluate specific Federal Statistical programs. He was the first recipient of the Roger Herriot Award, awarded by the Washington Statistical Society and the ASA Sections on Government Statistics and on Social Statistics for "innovation in federal statistics", and is a recipient of the Gold Medal Award of the U.S. Commerce Department. Finally, his greatest impact may be through the large number of colleagues who were inspired in their own efforts by his personal example, by his teaching, by his leadership, and by his kindness, thoughtfulness, and understanding.

## REFERENCES

BRICK, J. M., and WAKSBERG, J. (1991). Avoiding sequential sampling with random digit dialing. Survey Methodology, 17, 27-41.

BRICK, J. M., WAKSBERG, J. and KEETER, S. (1996). Using data on interruptions in telephone service as coverage adjustments. Survey Methodology, 22, 185-197.

BRICK, J. M, WAKSBERG, J., KULP, D. and STARER, A. (1995). Bias in list-assisted telephone samples. Public Opinion Quarterly, 59, 218-235.

CHU, A., EISENHOWER, D., HAY, M. MORGANSTEIN, D., NETER, J. and WAKSBERG, J. (1992). Measuring the recall error in self-reported fishing and hunting activities. Journal of Official Statistics, 8, 19-39.
NETER, J., and WAKSBERG, J. (1964). A study of response errors in expenditure data from household interviews. Journal of the American Statistical Association, 59, 18-55.
NETER, J., and WAKSBERG, J. (1965). Response Errors in Collection of Expenditure Data from Household Interviews: An Experimental Study. (Bureau of the Census Technical Paper No. 11). Washington, DC: U.S. Government Printing Office.

WAKSBERG, J. (1973). The effect of stratification with differential sampling rates on attributes of subsets of the population. Proceedings of the Social Statistics Section, American Statistical Association, 1973, 429-434.
WAKSBERG, J. (1978). Sampling methods for random digit dialing. Journal of the American Statistical Association, 73, 40-46.
WAKSBERG, J. (1983). A note on locating a special population using random digit dialing. Public Opinion Quarterly, 47, 576-579.
WAKSBERG, J. (1998). The Hansen Era: Statistical research and its implementation at the U.S. Census Bureau, 1940-1970. (With Discussion). Journal of Official Statistics, 14, 119-147.

WAKSBERG, J., JUDKINS, D. and MASSEY J. (1997). Geographic-based oversampling in demographic surveys of the United States. Survey Methodology, 23, 61-71.

# 2001 WAKSBERG INVITED PAPER 

Author: Gad Nathan

Gad Nathan is Professor of Statistics at the Hebrew University of Jerusalem and has long been associated with the Israel Central Bureau of Statistics, most recently as Chief Scientist. He received his Ph.D. from Case Institute of Technology, Cleveland OH and has published numerous papers in leading statistical joumals, including Journal of the American Statistical Association, Journal of the Royal Statistical Society, Survey Methodology, Journal of Official Statistics and Sankhya. His main research areas are sampling methodology, inference from complex samples, computer assisted interviewing and telesurveys. He has held visiting and consulting positions at several academic institutions and statistical agencies in North America and in Europe and has served as Vice-President of the International Statistical Institute and of the International Association of Survey Statisticians, as well as President of the Israel Statistical Association and Chairman of the Israel Public Council of Statistics.

# Telesurvey Methodologies for Household Surveys A Review and Some Thoughts for the Future 

GAD NATHAN ${ }^{1}$


#### Abstract

We consider 'telesurveys' as surveys in which the predominant or unique mode of collection is based on some means of electronic telecommunications - including both the telephone and other more advanced technological devices such as e-mail, Internet, videophone or fax. We review, briefly, the early history of telephone surveys, and, in more detail, recent developments in the areas of sample design and estimation, coverage and nonresponse and evaluation of data quality. All these methodological developments have led the telephone survey to become the major mode of collection in the sample survey field in the past quarter of a century. Other modes of advanced telecommunication are fast becoming important supplements and even competitors to the fixed line telephone and are already being used in various ways for sample surveys. We examine their potential for survey work and the possible impact of current and future technological developments of the communications industry on survey practice and their methodological implications.


KEY WORDS: Telephone surveys; Internet surveys; Sample design; Nonresponse; Coverage.

## 1. INTRODUCTION

Electronic telecommunications have become a predominant factor in practically all aspects of modern life at the beginning of the new millennium. Sample surveys are no exception and the widespread use of the telephone as a prime mode of communication for at least the past quarter of a century has had an important influence on survey practice. In fact, the telephone survey has become the major mode of collection in the sample survey field, especially in North America and Western Europe, both for surveys of households and individuals and for surveys of establishments. Other modes of advanced telecommunication, such as e-mail, Internet, videophone, fax and mobile phones are fast becoming important supplements and even competitors to the fixed line telephone. They are already being used in various ways for sample surveys and in this review paper we intend to examine their potential for survey work and the methodological implications of their use. We therefore wish to use the term 'telesurvey' for any survey in which the predominant or unique mode of collection is based on some means of electronic telecommunications - including both the telephone and other more advanced technological devices. Conventional surveys based on face-to-face interviews in the home or (snail-)mail surveys are not included, unless a substantial component of the survey is based on some telecommunications instrument. Although this paper focuses on surveys of individuals and households, much of it is relevant to establishment surveys too. We refer to telesurvey 'methodologies' in the plural, since it seems obvious that no single methodology will be suitable for use with the plethora of possible communication devices available in the future and their combinations.

This paper has been prepared in recognition of Joe Waksberg's unique contributions to survey methodology,
in general, and to telephone survey methodology in particular. It is well recognized today that his groundbreaking paper, Waksberg (1978), paved the way for the widespread efficient use of random digit dialing for telephone surveys and serves as a threshold point in the development of telesurvey methodology. Together with many of his subsequent papers, his work has had a profound influence on the theory and practice of telephone survey methodology, some of which will be examined in this paper.

We shall deal primarily with the statistical aspects of telesurvey methodology but recognize that these are not independent of non-statistical aspects, such as the cognitive features of telesurvey interviewing, survey administration and ethical considerations. In the following section we briefly review the early history of telephone surveys, through 1978. Section three reviews in some detail more recent developments in the areas of sample design and estimation, coverage and nonresponse and evaluation of data quality. Finally in section four we consider the possible impact of current and future technological developments of the communications industry on survey practice and their methodological implications.

## 2. THE EARLY HISTORY OF TELEPHONE SURVEYS

In the following we review briefly the overall early development of the use of telephones for survey work, as background for the developments in telesurvey methodologies to be described later. More detailed and comprehensive coverage is provided in several books and survey papers, e.g., Blankenship (1977a), Groves, Biemer, Lyberg, Massey, Nicholls and Waksberg (1988), Frey (1989),

[^0]Lavrakas (1993), Casady and Lepkowski $(1998,1999)$ and Dillman (1978, 2000).

Telephones have been used for survey work since the thirties, though generally as a supplementary mode of collection. Some have erroneously blamed the disastrous failure of the Literary Digest survey's prediction of a landslide victory of Landon over Roosevelt in 1936, at least partially, on telephone undercoverage (Katz and Cantril 1937; Payne 1956; and Perry 1968). In fact the survey was based on mail questionnaires and although telephone lists were used as a sampling frame (in combination with lists of automobile registrations), it seems that the failure was due more to nonresponse than to frame undercoverage (Bryson 1976; Squire 1988; and Cahalan 1989).

Most of the earliest reports on the use of the telephone in survey work were in the areas of public health or in market research applications. Many of them used some combination of telephone interviewing with other modes of collection and in some cases they included empirical comparisons of response rates or outcomes in order to assess mode effects. For instance, Cunningham, Westerman and Fischoff (1956) and Bennet (1961) report on telephone surveys for follow-up studies of patient treatment and Fry and McNaire (1958) on a national follow-up to a mail questionnaire to obtain opinions of hospital staff - all with high response rates. Mitchell and Rogers (1958) used telephone interviewing for a survey of telephone households on the consumption of dairy products and compare the results with those obtained from a control sample of non-telephone households. Cahalan (1960) compares results from telephone interviews with those from personal interviews in measuring newspaper readership with favourable results. Eastlack (1964) in a comparative telephone study of advertising recall and product usage shows that a rigorous call-back protocol provides more accurate results than a method without call-backs. Coombs and Freedman (1964) report on high telephone response (92\%) in a longitudinal fertility survey, supplemented by personal interviews. Sudman (1966) describes several supplementary uses of the telephone for survey work, which include making of advance appointments and screening for rare populations, with positive results for cooperation rates and cost reductions.

In the late sixties telephone surveys really came of age, as a result of several different developments. First of all the rapid increase in telephone coverage in Western Europe and North America implied that telephone interviewing could be used as a primary mode of collection. In the US household telephone coverage reached a level of $88 \%$. in 1970 (Massey and Botman 1988) and this level was reached somewhat later in most Westem European countries, in Australia and in New Zealand (Trewin and Lee 1988). In parallel to the rapid increase in telephone penetration in many countries a serious decline in response rates and difficulties in contacting respondents by door-to-door collection were experienced in the late sixties. This led to
serious consideration of telephone surveys both to reduce costs and to achieve higher cooperation rates. The use of telephone interviewing advanced most rapidly in commercial and academic survey organizations and less so in official government statistics. For instance the Federal Committee on Statistical Methodology (1984) reports that only about 11 percent of US Federal surveys in 1981 involved telephone interview in any form, in most cases in addition to other modes.

At first telephone interviewing was viewed with apprehension, even when used only as a supplementary mode of collection, due to fears of high nonresponse rates and response biases considered inherent when interviewing was not carried out face-to-face. Results of some of the earlier telephone surveys seemed to reinforce these fears. For instance, a study of leaflet receipt by Larson (1952) raises serious doubts on the validity of telephone responses on the basis of a face-to-face interview follow up. Similarly Oakes (1954) reports on suspiciously lower response on improvements to a consumer service via the telephone than obtained in face-to-face interviews. Schmiedeskamp (1962) in an attitude survey on consumer finances finds greater avoidance of taking strong positions when telephone interviewing was used. Wiseman (1972) in a comparison of mail questionnaire, telephone and face-to-face personal interviewing finds mode effects for sensitive issues (abortion and birth control). The main differences, however, are between responses to mail questionnaires and to personal interview (telephone or face-to face).

Many of these fears were allayed at an early stage by the results of a number of more rigorous empirical studies. Thus Hochstim (1967) in a well-designed controlled experiment compares collection by mail, telephone and personal interview as the primary mode of collection. The results demonstrate convincingly that the three strategies of data collection prove to be practically interchangeable when compared with respect to rate of return, completeness of return, comparability of findings and validity of responses. The major difference between modes is with respect to cost, with a clear preference for the mail or telephone strategy. Similarly a small test carried out by Colombotos (1965) on samples of a population of physicians shows no significance differences between responses obtained by telephone and by in-person interviews. Janofsky (1971) reports similarity in willingness to express feelings on health issues between telephone respondents and face-to-face interview respondents. A well designed validation study by Locander, Sudman and Bradburn (1976) of the effects of question threat and mode of collection found no meaningful differences in response bias between telephone and face-to-face interviews. Finally, in a small carefully controlled field experiment, Rogers (1976) tested the effects of altemative interviewing strategies on the quality of responses and on field performance in a survey on a variety of complex attitudinal, knowledge and personal items. The results again indicate that the quality of data obtained by telephone is
comparable to that obtained by interviews in person. A major national study comparing telephone and face-to-face interviewing was conducted by Groves and Kahn (1979). It was based on an intensive analysis of the large omnibus surveys carried out under the two modes by the University of Michigan Survey Research Center. It provided important information on data quality which did not indicate any substantial mode effects. These and other early studies, which foreshadowed several systematic studies of mode effects carried out in the eighties and nineties (to be discussed later) contributed to the legitimacy of telephone surveys as a standard mode of collection.

The initial use of telephones for sample surveys was usually based on samples selected from general frameworks, such as telephone directories, or from specific frameworks for small sub-populations. Towards the end of the sixties there was increased awareness of high rates of unlisted telephone numbers and of substantial differences between households with listed and non-listed numbers (see details in section 3.1.1). An important development that overcame this problem was the sampling method of Random Digit Dialing (RDD), first introduced by Cooper (1964) and further improved and developed by Eastlack and Assael (1966) and by Glasser and Metzger (1972). An inherent inefficiency of these basic element RDD methods was the large amount of numbers to be called that did not yield an interview (non working and non residential numbers). A two-stage RDD sampling method was first proposed to deal with this problem by Mitofsky (1970) and subsequently elaborated and put on a firm theoretical basis by Waksberg (1978). The introduction of what was to become known as the Mitofsky-Waksberg scheme contributed greatly to the widespread use of telephone surveys in the eighties and nineties.

Finally the technological advances in telecommunications and automation in the sixties and seventies contributed to the advantages of telephone surveying. Universal direct long distance dialing enhanced the possibilities of carrying out national surveys from a single center or from a small number of interviewing centers with all the advantages of central control and administration. However the greatest impact on the expansion of telephone surveys has undoubtedly been the introduction of Computer Assisted Telephone Interviewing (CATI) in the seventies. This is due both to the simplicity of CATI for conducting telephone interviews and to the possibilities it offers for the use of automation in many important non-interviewing tasks, (e.g., dialing, recall schedules etc.).

One of the first uses of the computer for on-line questioning was in the form of a multi-station computer-based laboratory experiment designed to elicit subjective information - Shure and Meeker (1970). A good account of the early history of CATI can be found in the special issue of Sociological Methods and Research (Freeman and Shanks 1983), following the Berkeley Conference on ComputerAssisted Survey Technology held in Spring 1981. Market
research organizations were the first to introduce CATI systems for their current operations. Chilton Research Services developed and used the Survey Response Processor on a current basis already in 1972 -Fink (1983). Other commercial survey organizations, applying different systems, realized early on the advantages of CATI - for instance the A\&S/CATI ${ }^{\text {TM }}$ system (Dutka and Frankel 1980). Academic survey research organizations were quick to follow with the earliest systems developed at UCLA and Berkeley for the large scale CATI-based California Disability Survey - Shanks, Nicholls and Freeman (1981) and Shanks (1983). Another early development of a CATI system at an academic survey organization, using a different approach, based on microcomputers, was that of the University of Wisconsin (Palit 1980; Palit and Sharp 1983). In Europe the first survey research organizations to use CATI were Social and Community Planning Research (SCPR - now the National Centre for Social Research) in the UK (Sykes and Collins 1987) and the State University of Utrecht, Netherlands (Dekker and Dorn 1984). The introduction of CATI systems into official statistics was slower. In the US it started in 1982 at the Census Bureau (Nicholls 1983) and at the National Agricultural Statistics Service (Tortora 1985) and at the same time at Statistics Netherlands (1987). By 1987 practically all organizations surveyed in a (non-probability) sample of 27 survey organizations (eighteen in the US and nine elsewhere) were using CATI for some or all of their telephone surveys - Berry and O'Rourke (1988). A report of the Federal Committee on Statistical Methodology (1990) indicated that the number of CATI installations worldwide at the end of the eighties was estimated to exceed 1,000 and that in 1988, the U.S Government had 51 cooperating CATI centers. It should be noted that the development of CATI quickly became part of a wider movement toward computer assisted interviewing (CAI) or computer assisted information collection (CASIC), which includes also CAPI (Computer Assisted Personal Interviewing) and CASI (Computer Assisted Self Interviewing) - Nicholls (1988). A more complete history of the development of CATI and of CASIC, in general, can be found in Couper and Nicholls (1998).

## 3. RECENT DEVELOPMENTS IN TELEPHONE SURVEYS

In the last quarter of a century telephone surveying has definitely come of age. Lyberg and Kasprzyk (1991) claim that it has become the dominant mode of collection in countries with extensive telephone coverage.

Hundreds of scientific papers have been published during this period on a wide range of different aspects of telephone surveys. Several general books on the subject have appeared - Blankenship (1977a), Groves and Kahn (1979), Frey (1989) and Lavrakas (1993). A number of conferences have been devoted to telephone survey
methodology or have dealt with specific aspects of the topic. The results have appeared in monographs or special issues of scientific journals. A major conference on telephone survey methodology was held in November 1987 in Charlotte, NC, with the resulting volume edited by Groves, Biemer, Lyberg, Massey, Nicholls and Waksberg (1988) and the special issue of the Joumal of Official Statistics, edited by Groves and Lyberg (1988b). The Berkeley Conference on Computer-Assisted Survey Technology held in Spring 1981 (Freeman and Shanks 1983) dealt primarily with telephone surveys and CATI was a major topic at the InterCASIC '96 International Conference on Computer Assisted Survey Information Collection, held in San Antonio, TX in December 1996 (Couper, Bethlehem, Baker, Clark, Martin, Nicholls and O'Reilly 1998) and at the ASC $3^{\text {rd }}$ International Conference at Edinburgh in September 1999 (Banks, Christie, Currall, Francis, Harris, Lee, Martin, Payne and Westlake 1999).

Extensive bibliographies with several hundred entries can be found in the above sources, as well as in Khurshid and Sahai (1995), which covers the period through 1991, and in Survey Research Center (2000), which updates previous bibliographies with respect to sample design for household telephone surveys through 2000.

In the following we review the development of telephone survey methodology for household surveys during the past 25 years in the areas of sample design and estimation, coverage and nonresponse and evaluation of data quality.

### 3.1 Sample Design and Estimation

Sampling methodology for telephone surveys is based on the general principles of sampling. It is primarily adapted to the special situation of telephone surveys with respect to the sampling framework used. Thus we adopt the classification proposed by Lepkowski (1988) for telephone sampling methods, according to the underlying sampling framework - directory and commercial lists, telephone numbers (RDD) and combined methods (list-assisted and dual frame).

### 3.1.1 List-based Sampling Procedures

As mentioned above, the earliest telephone surveys were all based on samples selected from lists. In many cases they were mixed-mode surveys where telephone interviewing was used to supplement for non-response in face-to-face interviews or for follow-up. Thus so-called'warm telephone interviewing' schemes have been used in the US Current Population Survey and in the Canadian Labour Force Survey - Drew, Choudhry and Hunter (1988). In these cases sampling is based on a general list framework to which information on telephone numbers is added and no special features of the use of the telephone are involved in the sample design. The same goes for 'pure' telephone surveys of special populations, such as physicians, for which a complete list of the population is available with telephone numbers and can be used as a sample framework

- see, for example, Gunn and Rhodes (1981). Another example is where telephone interviewing is used in follow-up waves of a panel survey with the first contact carried out by a face-to-face interview. For instance in the Israel Labor Force Survey the first contact is by a home visit and the second and third waves are carried out by telephone for households who are willing to respond by telephone Nathan and Eliav (1988). A related approach, used recently in a pilot study for the US National Study of Health and Activity (Maffeo, Frey and Kalton 2000), is to take an area sample, find telephone numbers where possible, for telephone interviewing, and use face-to-face interviewing for other households and for telephone nonrespondents.

The most easily obtained and low-cost directory that can be used as a framework for a telephone surveys is, of course, the telephone directory itself, or some modification of it. Originally the paper version of the directory was used, while nowadays an electronic version would usually be available. The major deficiencies of the telephone directory as a sampling framework are well documented. They are undercoverage, overcoverage, duplication and lack of auxiliary information. Undercoverage is by far the most serious deficiency and includes both non-telephone households and households with telephones unlisted by choice or those not yet included in the directory. The biases due to non-telephone households are, of course, irrespective of the framework used and will be dealt with in section 3.3.

The extent of unlisted telephones varies considerably by country and type of location, as well as by other household variables. Sykes and Collins (1987) report on an unlisted rate of $4 \%$ in the Netherlands and $12 \%$ in the UK. Frejean, Panzani and Tassi (1990) estimate the unlisted rate in France as $14 \%$ and national US estimates in the seventies were of over 17-19\% (Blankenship 1977b and Glasser and Metzger 1975). Rich (1977) reports on increasing rates of nonpublished telephones (excluding those involuntarily unlisted) in the Pacific Telephone's California serving area from $9 \%$ in 1964 to $\mathbf{2 8 \%}$ in 1977. In addition some $5 \%$ of home telephones in California were estimated to be involuntarily unlisted (assigned after publication of the directory). More recent studies show substantially higher unlisted rates. Thus Genesys (1996) reports unlisted rates of $40 \%$ in 1993 and of $37 \%$ in 1995, based on national samples of more than 100,000 RDD telephone interviews and Survey Sampling Inc. (1998) estimates the US national unlisted rate for 1997 at $30 \%$. Results of a small-scale study of the Jerusalem area (Nathan and Aframian 1996) indicate an unlisted rate of $27 \%$.

Many studies have shown substantial differences between listed and unlisted telephone household characteristics, 'indicating disturbing potential coverage biases for directory-based samples. In the US these differences were demonstrated, for instance, in a study by Brunner and Brunner (1971), who found highly significant differences between listed and unlisted telephone households with respect to a wide range of demographic and socio-economic
variables. Leuthold and Scheele (1971) found higher rates of nonlisting among blacks, city dwellers, young people, apartment dwellers, divorced and separated and among service workers. Similarly, Roslow and Roslow (1972) found significant differences in audience shares between listed and unlisted telephone households. Glasser and Metzger (1975) showed that nonlisted rates were higher in the West, in major metropolitan areas, among non-whites and the young. Blankenship (1977b) and Rich (1977) found highly significant differences between listed and unlisted households with respect to sex and age of household head, occupation, household size and income. In the UK Sykes and Collins (1987) found more unlisted numbers among the young, the poorest and those living in London. The results of Nathan and Aframian (1996) for the Jerusalem area showed lower rates of TV ownership and of TV viewing (of those with TV) in an RDD sample as compared with a directory listing sample.

Besides the undercoverage resulting from unlisted numbers, as indicated above, directory listings also suffer from problems of overcoverage, duplication and lack of updated auxiliary information. Overcoverage occurs when a unit outside the population is included in the framework. This may be due to the fact that disconnected numbers often remain in the directory, commercial numbers are not always clearly designated as such or other cases of unrecognized ineligibility. Duplication occurs when the same unit is represented in the frame more than once and the duplication is not recognized. Duplication can usually be discovered during sampling if the entries for the same household are listed consecutively but not if they appear separately (e.g., under different surnames). If duplication is ascertained during the interview (i.e., by obtaining information on the number of connected lines available to the household or the number of directory listings) it can be dealt with by appropriate weighting. Although these problems are surmountable, at a cost, that of undercoverage is not and this indicates the need for more representative sample frameworks than provided by directories. A popular altemative to the traditional telephone directory (in general prepared by the company providing telephone service to the area) has been the lists prepared by commercial firms, usually for purposes of marketing. These may be city directories, obtained from municipal address listings with telephone numbers obtained from directories or other sources, subscriber lists of telephone companies or national master address lists, such as that provided by Donelley Marketing, Inc. in the US - Lepkowski (1988). These lists provide important auxiliary data, such as geographic information, from the Census of Population and Housing and from other sources. They do not, in general, overcome the bias due to unlisted numbers and their cost may be high. They can result in some gain in sampling variance, due to the possibility of basing an efficient design on the auxiliary information. Potentially, lists used by emergency services to determine the physical location of callers could be used as
frameworks, although access to these lists would be difficult for non-government survey organizations.

### 3.1.2 Random Digit Dialing - The Mitofsky -Waksberg Scheme

In order to overcome many of the inherent problems of directories and commercial lists, Random Digit Dialing (RDD) methods have become a popular choice for telephone surveys, primarily in the US. These are based on the frame of all possible telephone numbers. The method was originally proposed by Cooper (1964), who added random four digit suffixes to known prefixes in a local survey. This basic element sampling method was further improved and developed by Eastlack and Assael (1966) and by Glasser and Metzger (1972), on a national level, by identifying 'working banks' of numbers from telephone company information.

The use of RDD has until recently been confined, by and large, to the US and Canada. Thus Sykes and Collins (1987) report that telephone surveys were still rare in the UK at the end of the eighties, primarily due to low telephone coverage. In particular RDD surveys were rarely used - one of the reasons being the lack of uniformity in the length of telephone numbers at the time. However recently, with the increase of telephone coverage in the UK to some $96 \%$ at the end of the nineties and the standardization of telephone numbers to ten digits, RDD surveys have become more popular - see e.g., Collins (1999) and Nicolaas, Lynn and Lound (2000). Similarly, Gabler and Haeder (2000) report that an RDD method, modified in order to deal with varying telephone number lengths (from 6 to 11 digits!), is now standard procedure for telephone surveys in Germany.

Mitofsky (1970) first proposed a two-stage RDD sampling method to deal with the problem of the inherent inefficiency of these basic element RDD methods due to the large amount of numbers to be called that did not yield an interview (non working and non residential numbers). This was subsequently elaborated and put on a firm theoretical basis by Waksberg (1978) and the method became known as the Mitofsky-Waksberg scheme. This scheme or variations of it have become the predominant sampling method for telephone surveys, at least in the US.

The method is based on the fact that household telephone numbers are, in general, clustered in series of consecutive numbers or within banks of numbers with the same first $r$ digits. For the US $r$ is usually set at eight (for ten digit telephone numbers, including area code), so that the banks or clusters (PSU's) are of size $N=100$ each. It is assumed that the telephone company can provide a list of all operating prefixes (area code + first three digits), i.e., those to whom residential numbers have been assigned. To the six digit numbers in this list all possible choices of two digits are added, resulting in a sampling frame of eight digit numbers that represent the M PSU's in the population. Sample PSU's are selected from this frame at random (with replacement) consecutively and for each PSU selected two
final digits are selected at random. The resulting ten digit number is dialed and if the number is not that of a residence (according to the survey definition), the PSU is dropped from the sample. If it is a residence a simple random sample (without replacement) of $k$ additional residential numbers is selected by contacting numbers selected at random (without replacement) from the PSU, until $k$ additional residential numbers are obtained. The procedure of PSU selection continues until a fixed number of PSU's, $m$, has been selected. It is easily seen that, assuming that the number of residential numbers in each selected PSU, $P_{i}$, is at least $k$, the total sample size of residential telephone households is $m(k+1)$ and that the final sample is an equal probability sample from the population of all residential telephone households.

Waksberg (1978) shows that if we designate by: $\pi=\left(\sum_{i=1}^{M} P_{i}\right) /(N M)$ the proportion of residential numbers in the population and by $t$ the proportion of PSU's with no residential numbers (i.e., for which $P_{i}=0$ ), then the expected number of total calls is given by: $m[1+(1-t) k] / \pi$, assuming that $P_{i} \geq k+1$ for all PSU's with at least one residential number. The last assumption can be dropped if PSU's are grouped so that the restriction holds in each group or if unequal weighting is used. Optimal values of the design parameters are obtained under a simple cost function and the method is extended to deal with repeated surveys. The main advantage of the method is the reduction in the expected number of calls which have to be made in order to attain a given effective sample size, especially if $t$, the proportion of PSU's with no residential numbers, is larger than 0.5 . Groves (1977) provides data for a national study indicating a value of $t$ of about 0.65 . This advantage has to be weighed against the increase in variance due to the effect of clustering. However, taking costs into account, illustrative calculations for typical values of the parameters show that reductions in costs run between 20 and $40 \%$.

The major operational drawback of the method is in its sequential nature. This makes it unwieldy to carry out manually. However the sequential operation poses no problem when the process of selection is fully automated. The method as described above has some additional problems, most of which can be overcome by simple modifications. Assuming that prior information on the number of telephone households is not available, selection probabilities are not known, although the value of $p$ can be estimated from the sample. The practical necessity to introduce a stopping rule for the number of calls to numbers which do not answer or to refusals to answer, even whether the number is a residential one, implies that the method cannot be strictly applied as designed, resulting in possible bias. The problem of households with multiple telephone numbers can be overcome if correct information on the number of different lines is obtained but the required re-weighting impinges on the simplicity of equal weighting. In some cases names and addresses can be obtained for RDD
numbers by matching with address lists so that advance notice can be sent to at least part of the potential respondents. However this is a complex procedure and the difficulties in sending advance notice to respondents (common to all RDD procedures) has made the procedure difficult to consider for some official statistical agencies.

### 3.1.3 Modifications of the Mitofsky-Waksberg and Other RDD Methods

Some of the drawbacks of the basic method are overcome by the generalization due to Potthoff (1987a, 1987 b ). The method is based on the definition of a set of auspicious telephone numbers. This could consist of only residential numbers, as in the Mitofsky-Waksberg method, or a wider set which includes all residential numbers - for instance the set of all numbers which ring (including engaged, recorded messages and operators). The first stage of selection is by simple random sampling of a fixed number, $m$, of PSU's. From each selected PSU a fixed number of calls, $c$, are made and for each of them it is determined whether the number is auspicious or not. A PSU is discarded if all $c$ numbers selected are inauspicious. Retained PSU's are defined as Type I if only one number is auspicious and as type II if two or more are auspicious. The second stage consists of selecting and dialing $k c$ numbers from Type I PSU's and $k(c-1)$ numbers from Type II PSU's, where $k$ is an integer. At all dialed numbers the unit is determined as residential or out-of-scope and an interview is attempted for all residential units. A supplementary sequential segment for Type I PSU's selects additional telephone numbers that are dialed until a total of $k$ auspicious numbers are obtained. An interview is attempted at each auspicious numbers dialed in the sequential segment. Potthoff (1987a) shows that, under certain conditions, all residential telephone numbers have the same probability of selection and develops unbiased and ratio estimates and their variances. Cost comparisons and some modifications to overcome practical problems are also given. The method reduces the problem of ambiguity on the status of dialed numbers from which no response is obtained and also the problem of exhaustion of the residential numbers in a PSU.

A large number of additional generalizations and modifications to the basic Mitofsky-Waksberg method have been proposed. Many of these attempt to reduce the burden of interviewing screening and to improve control over the initial contact sample size. Thus Hogue and Chapman (1984) propose determining cutoff points on the basis of an estimation of the probability that a PSU is 'sparse', i.e., has a small proportion of residential numbers, and propose to determine an optimal cutoff procedure on the basis of cost and variance considerations. Alexander (1988) considers two types of cutoff rules to limit interviewing screening for prefixes with low residential densities. An 'increasing rule' stops as soon as a predetermined number of calls, $c_{i}$, has been made and less than $i$ residences have been found,
where $\left\{c_{i}\right\}$ is an increasing series in $i$. A 'decreasing rule' stops when $i$ residences have been found if at least $c_{i}$ calls have been made, where $\left\{c_{i}\right\}$ is a decreasing series in $i$. The costs for these rules are evaluated under a simple model.

Lepkowski and Groves (1986a) propose a two phase design based on matching prefixes selected in the first stage of the Mitofsky-Waksberg scheme to a commercial directory to obtain counts of listed telephones for each prefix selected. Prefixes are allocated to two strata - a low density stratum where there are no listed telephone numbers, or only a small number of them, and a high density stratum. The Mitofsky-Waksberg design is applied to the lowdensity stratum and telephone numbers are selected with probability proportional to the number of listed telephone numbers in the high-density stratum.

Brick and Waksberg (1991) propose using a fixed number of telephone numbers in the second stage so as to avoid sequential sampling altogether with a resulting simplicity of operation. The design, originally proposed by Waksberg (1984), is not, however, self-weighting and involves a slight bias and increased variance. Brick and Waksberg (1991) suggest considerations for the choice between the original and modified Mitofsky-Waksberg designs. For an early application of the modified MitofskyWaksberg method to the collection of health attitude information, apparently in an erroneous attempt to implement the original method - see Cummings (1979). Smith and Frazier (1993) compare the original and modified schemes, using data collected in the California Behavioral Risk Factor Surveillance System. The results indicate that the modified scheme speeds up the data collection, resulting in a larger sample size for the same cost. This compensates for larger design effects of the modified scheme.

Another alternative to the basic Mitofsky-Waksberg method is the use of stratification and disproportionate allocation to improve 'hit rates', proposed by Palit (1983). An evaluation of alternative treatments of unanswered telephone numbers for the Mitofsky-Waksberg design is carried out by Palit and Blair (1986). The optimal determination of parameters for the Mitofsky-Waksberg method is dealt with by Burke, Morganstein and Schwartz (1981) and the optimal allocation for the stratified version of the method by Casady and Lepkowski $(1991,1993)$ and by Tucker, Casady and Lepkowski (1992). Further problems relating to minimal cost allocation are treated by Palit (1983) and by Mason and Immerman (1988).

### 3.1.4 List-Assisted Methods

Although RDD methods overcome the undercoverage of directories due to unlisted numbers, they all still suffer from the basic problem of undercoverage due to non-telephone households (see further detail in section 3.3). In addition the lack of auxiliary information (such as geographical information), which is often available in list frames, leads to inefficiencies, even in the more sophisticated modifications
of the basic methods, mentioned above. Thus alternative methods have been sought to combine RDD samples with samples based on list and directory frames. One of the earliest attempts in this direction was that proposed by Stock (1962) and elaborated by Sudman (1973), based on replacing the last two digits of telephone numbers, selected from a directory listing, by random digits. The method was applied by Hauck and Cox (1974) to a methodological study of mode effects in screening for a special subpopulation. A simpler version, popularly known as the 'Plus One' method, replaces each telephone number sampled from a directory by the number plus one (or some other digit - known as the 'plus digit method). This supposedly overcomes the bias due to unlisted numbers. Due to its simplicity, the method has gained popularity among market researchers. However several studies - e.g., Landon and Banks (1977); and Mullet (1982) - have indicated that it is not, in fact, bias-free and also suffers from low efficiency.

Forsman and Danielsson (1997) propose a model-based approach for plus digit sampling, based on the assumption of randomly mixed listed and unlisted numbers within prefix. The model, which is tested empirically, provides model unbiased estimates. Ghosh (1984) has proposed an improved method that continues adding one to the last telephone number dialed as long as a household is not reached and stopping once a household is reached. Although still biased, the bias is reduced as compared with the simple 'plus one' method. Other list-assisted methods with RDD components, are discussed by Potter, McNeill, Williams and Waitman (1991), who stratify prefixes according to counts of published telephone numbers, while ensuring inclusion of blocks without any published numbers.

Brick, Waksberg, Kulp and Starer (1995) propose a list-assisted method that overcomes the troublesome problem of the sequential nature of the second stage sampling inherent in the Mitofsky-Waksberg scheme. The method is based on dividing the file of exchanges ( 100 -banks) into two strata. The first consists of all exchanges with at least one listed residential phone and the second those that have none. Sampling only from the first stratum drastically reduces the proportion of nonresidential numbers which have to be dialed, but results in coverage bias. They investigate the bias and conclude that such truncated sampling methods are efficient and have operational advantages, while the resulting coverage bias (about 4\%) is not too important. The method has been widely applied to replace the classical Mitofsky-Waksberg method. Similarly Statistics Canada has used the method for their General Social Survey since 1991 for the whole sample, with simple random sampling within banks of numbers identified as having at least one residential number (Norris and Paton 1991). Modifications of this design include a complete stratification of number banks on the basis of list information and using simple RDD for strata with small proportions of banks with no listing and the

Mitofsky-Waksberg method in the remaining strata. A comparison of this design with other stratified designs based on a cost model is carried out by Casady and Lepkowski (1993). Their results show that for low cost ratios (of productive selections to unproductive selections) two and three stratum RDD designs are as efficient as the Mitofsky-Waksberg scheme and that for high cost ratios they are superior.

### 3.1.5 Multiple Frame Designs

In an attempt to overcome some of the inherent biases of telephone surveys due to directory and telephone undercoverage, the use of dual frame mixed mode surveys, combining telephone with face-to-face interviewing, has received increasing attention. These combine conventional samples for personal interview with RDD or directory samples for telephone interviewing. Biemer (1983) investigated the optimal mix for such designs, via a simulation study, and McCarthy and Bateman (1988) propose the use of mathematical programming for attaining optimal allocation of sample units for a dual frame design, which allows posterior analysis of the effects of variations in design and cost parameters on the optimization. Choudhry (1989) proposes a cost-variable optimization for estimating proportions and Brick (1990) proposes the use of multiplicity sampling for this purpose. In a series of papers, Groves and Lepkowski (1985, 1986); Lepkowski and Groves (1984, 1986b); and Traugott, Groves and Lepkowski (1987) develop error models for these dual frame survey designs. They also report on results of experiments to compare response rates and potential biases of RDD and list samples and of several interviewing methods. The results were applied to the large scale US National Crime Survey.

Whitmore, Mason and Hartwell (1985) report on applications of dual frame dual mode methods in a US Environment Protection Agency sponsored study of personal exposure to carbon monoxide in two metropolitan areas and in a state-wide study of social service needs. In both cases commercially available directory lists were used in association with area household sampling. On the basis of an analysis of their results, they recommend the use of such dual designs in order to both benefit from the relative efficiency of telephone interviewing and to overcome the biases inherent in the use of directories as sampling frames. A combination of RDD and area sampling is reported by Waksberg, Brick, Shapiro, Flores-Cervantes and Bell (1997) for the US National Survey of America's Families in with there was particular focus on the low-income population. The nontelephone households identified in the area screening were given cellular phones for responding to telephone interviewers, thereby avoiding the need to train the area screener interviewers in a non-telephone questionnaire (Cunningham, Berlin, Meader, Molloy, Moore and Pajunen 1997).

### 3.2 Other Sampling Issues

### 3.2.1 Sampling for Special Populations

The relative low costs of telephone interviewing have made this survey mode a prime candidate for use in screening large samples in order to locate small special populations. Thus Sudman (1978) discusses the conditions under which the use of a telephone sample for screening a subgroup, to be finally interviewed face-to-face, is more efficient than face-to-face screening. By analyzing cost functions, telephone screening is found to be efficient, unless within-cluster homogeneity is small, interview densities are low and/or location and screening costs are low, relative to interview costs. Blair and Czaja (1982) propose a modification of the Mitofsky-Waksberg procedure to locate special populations that cluster geographically and describe an application to the Black population. As pointed out however by Waksberg (1983), their method requires reweighting when clusters are exhausted, which may result in reduced efficiency. This implies that the method may be efficient for the Black population but not necessarily for other minorities. Another telephone sample design targeting the US black population is proposed by Inglis, Groves and Heeringa (1987). Mohadjer (1988) proposes the stratification of prefix areas in an RDD design for sampling rare populations. The use of the MitofskyWaksberg method for selection of households combined with a stratified sample of individuals within household is used for the selection of a population-based control group in four epidemiological studies reported by Hartge, Brinton, Rosenthal, Cahill, Hoover and Waksberg (1984). The effectiveness of the method is studied by Pemeger, Myers, Klag and Whelton (1993), on the basis of a simulation of simple random sampling, and found to be effective.

Local area surveys are another example of special populations that can be dealt with efficiently by a telephone survey. Although, in general, telephone exchanges do not define geographical areas exactly, there is a high degree of correspondence and, with some screening for those in the defined area, telephone interviewing can reduce costs considerably. For instance Banks and Hagan (1984) report on the reduction of interviewer screening by a combination of list sampling and RDD for a survey to assess the effectiveness of health programs in specific service areas. Similarly, Campbell and Palit (1988) tested a combination of list sampling and TDD - total digit dialing, using a frame of all numbers in exchanges corresponding to a given census area. They found that this resulted in a substantial saving in enumeration costs, versus face-to-face interviewing.

### 3.2.2 Sampling Individuals Within Households

Almost all household surveys include questions relating to individuals in the household. In some cases all individuals belonging to the household are included in the sample,
but in many cases, for various reasons, a sample of one or more individuals is selected within the household for individual questions. The classic Kish procedure (Kish 1949), predominantly used in face-to-face interview surveys raises particular problems for telephone surveys, because it requires obtaining complete household listings over the telephone. This is more difficult to obtain over the phone than in a face-to-face interview, where some of the persons may be physically present. It should be pointed out however that in many cases the information on household composition is required in any case. In addition the manipulation of the selection rules by the interviewer (e.g., to accomplish high response rates), which has long been suspected in face-to face interviewing is almost impossible in CATI surveys (where selection is invisible to the interviewer).

Troldahl and Carter (1964) proposed a method whereby only the number of persons of each sex is required. Probabilistic rules (e.g., 'oldest man') are then applied to determine the individual selected, ensuring known selection probabilities for each person. However a positive probability of selection for each individual is not ensured (e.g., in households with three males the one of intermediate age is never selected). The method (known as the 'Troldahl-Carter method') has been modified by Bryant (1975), in order to take into account the possibility of households with more than two individuals of the same sex. An altemative method proposed by Salmon and Nichols (1983) and by O'Rourke and Blair (1983) is to select the person with the next (or last) birthday (the 'next-birthday' or 'last-birthday' method), which ensures equal probability of selection for each household member, under the assumption that the date of interview is random. This is of course a reasonable assumption only for surveys carried out over a twelvemonth period but not for surveys with shorter interview periods. This and other factors may lead to selection probabilities that are correlated with the individual characteristics. Another selection method proposed by Hagan and Meier (1983), which does not require any preliminary information on household composition, selects a predefined person (e. g., 'eldest man'). The method again fails to ensure a positive probability of selection for each household member.

Several empirical comparisons of the above methods have been carried out. Czaja, Blair and Sebestik (1982) found no significant differences in response rates or in demographic profiles between two versions of the TroldahlCarter method and the Kish method. Hagan and Meier (1983) compare their method, described above, with the Troldahl-Carter method and find that the method they propose has a significantly lower refusal rate, with no significant differences in demographic profiles. Salmon and Nichols (1983) compare four procedures for selecting respondents within a household unit - Troldahl-Carter, male/female alternation, next-birthday and no-selection methods - in a small telephone survey. They reach the conclusion that the next-birthday method is a relatively
efficient procedure for selecting a sample that is representative of all household members. Oldendick, Bishop, Sorenson and Tuchfarber (1988) find no significant differences between the Kish method and the last-birthday method. In a study using the last birthday method, Romuald and Haggard (1994) find that informants self-select to participate at a higher rate than expected. They investigate the effect of using memory cues on respondent selfselection and reach the conclusion that there is no significant effect. Lavrakas, Bauman and Merkle (1993) evaluate the effect of the use of the last-birthday method on within-unit coverage in a national survey and report evidence to suggest that the method leads to incorrect selection in many cases. Forsman (1993) reviews experiences of within-household sampling for 18 private opinion research companies and report on a test to compare the Kish, next/last birthday and the Toldahl-Carter methods. They conclude that the Troldahl-Carter method is somewhat better than the Kish method and that both are superior to the birthday methods. Similarly, Binson, Canchola and Catania (2000) report on a three-way comparison in a national telephone survey between the Kish, next-birthday, and last-birthday methods, and find significant differences between the three methods in the dropout rate, during the initial stages of the screening process. The Kish method had the highest dropout rates and the 'next-birthday' had the lowest rate. They conjecture that interviewers, rather than respondents, are a primary source of the higher rate of refusals when using the Kish method, due to the fact that a full household roster is required.

### 3.3 Coverage and Nonresponse

### 3.3.1 Telephone Coverage

The problem of telephone noncoverage was until very recently a major drawback of telephone surveys. Even in the US overall person undercoverage (in nontelephone households) remained at $7.2 \%$ by the end of 1986 Thomberry and Massey (1988). By the mid-eighties household telephone undercoverage was less than $10 \%$ in most Western countries, with the highest coverage (99\%) in Sweden. But some countries still had high rates of telephone undercoverage, for instance: UK $\mathbf{2 5 \%}$, Italy $\mathbf{2 9 \%}$ Ireland $50 \%$, Israel $30 \%$ - Trewin and Lee (1988). The situation changed dramatically towards the end of the century, with most Western countries reaching virtual saturation. Telephone coverage reached $94.4 \%$ in the US in 1999 (NTIA 2000); 96.6\% in Australia in 1996 (St. Clair and Muir 1997); $97.0 \%$ in the UK (OFTEL 1999); 97.3\% in Israel (Central Bureau of Statistics 2000); $97.9 \%$ in Finland (Kuusela and Vikki 1999); $98.2 \%$ in Canada (Statistics Canada 1999); and $99 \%$ in Germany (Federal Republic of Germany 1999).

Obviously the major problem of telephone undercoverage lies primarily in differential undercoverage rather than in its overall rate and the fact that telephone under-
coverage is highly correlated with a wide range of demographic, economic and health variables. This has been demonstrated extensively in a large number of empirical studies in the US and elsewhere - see for instance Groves and Kahn (1979), Collins (1983, 1999), Thornberry and Massey $(1983,1988)$, Trewin and Lee (1988) and Botman and Allen (1990). The rapid increase in overall telephone coverage over the last decade has not caused any radical change in this situation. Thus in Finland, with an overall telephone undercoverage of $2.1 \%$ in 1999 , low income households (less than 675 Euros per month) had an undercoverage of $11.3 \%$ (vs. $0 \%$ for high income groups) and those living in rented accommodation $4.9 \%$ (Kuusela and Vikki 1999). In Israel telephone undercoverage was $17.9 \%$ for the lowest income decile as against $0.8 \%$ for the two highest deciles and $24.9 \%$ for single adult households with three or more children as against $2.4 \%$ for childless households with three or more adults (Central Bureau of Statistics 2000). Similarly in the US large geographical variations are still found and telephone undercoverage is found to correlate with housing deficiencies, race, education income and mobility (Shapiro, Battaglia, Hoaglin, Buckley and Massey 1996; Giesbrecht, Kulp and Starer 1996; Fox and Riley 1996; NTIA 2000). Health- related characteristics were found to differ somewhat between persons in telephone and non-telephone households in the National Health Interview Survey by Anderson, Nelson and Wilson (1998) and in the National Health and Nutrition Examination Survey by Ford (1998). However telephone coverage effects were considered to be minor in both studies.

However the main problem of telephone coverage foreseen for the near future relates to the introduction and rapid proliferation of mobile telephones. In the late nineties the proportion of households with access to at least one mobile telephone reached $76 \%$ in Finland, 59\% in Denmark, $35 \%$ in Italy (Rouquette 2000) and $52 \%$ in Israel (Central Bureau of Statistics 2000). If all these mobile telephones were additional to fixed line telephones no problem would arise. However there are already strong indications of a tendency in several countries to consider the mobile telephone as an alternative to a fixed line telephone, rather than a supplement. Kuusela and Vikki (1999) report that $20 \%$ of Finnish households now have exclusively one or more mobile telephones and no fixed line and predict that within a year the number of mobile phones will exceed the number of fixed lines. Similar figures for the UK are 3\% (OFTEL 2000) and for Israel 2.9\% (Central Bureau of Statistics 2000). This implies that fixed line telephone coverage is down to $77 \%$ in Finland and to $94 \%$ in the UK and in Israel. In Germany it is estimated that the percentage of households with fixed line telephones will decrease to $\mathbf{9 2 \%}$ by 2004 (Gabler and Haeder 2000). Furthermore the characteristics of persons with only mobile telephones are quite different from those with fixed telephone lines. In Finland, according to Kussela and Vikki (1999), they tend
to be young, often living alone in rented apartments in urban areas. It should be noted that the transfer from fixed phone lines to mobile telephones is apparently not occurring to any large extent in North America, due to differences in pricing strategies.

Theoretically RDD sampling could be extended to mobile telephones. In practice; this may be quite difficult due to the fact that mobile telephones are by nature a personal appliance, rather than a household one. Sampling persons within a household, via a mobile telephone contact with one of the members, is well nigh impossible. Interviewing via a mobile telephone of individuals who may be anywhere is also extremely difficult. Even the determination of the total number of telephone numbers (mobile and fixed line) available to a household (required for weighting) may be daunting. We consider some possible approaches to these and other problems of the move to mobile telephones in section four.

Undercoverage of persons within covered households relates primarily to the method of selection for individuals within the household - see section 3.2.2 - and to the undercoverage due to the failure to obtain complete listings of individuals in the households. The latter effect is investigated by Maklan and Waksberg (1988), by comparing data on individuals obtained from an RDD survey with those obtained from the US Current Population Survey and from the population census. They find that while mean household sizes are comparable, the RDD results are skewed towards two-person households and away from one-person households. Some of the difference could be attributed to different residence rules, but the results do not indicate undercoverage of persons in the RDD survey. They also report on an experiment in which more detailed questions were asked on household composition and found practically no improvement in accuracy of reporting. In a similar experiment, carried out by Bercini and Massey (1979), the effects of the use of names in the household roster and the position of the question on the household roster (before or after the first interview) were tested in a survey on smoking. They found that both the use of names and the position of the household roster had an effect on response and that obtaining the roster after the interview without names is optimal.

### 3.3.2 Nonresponse

The problem of nonresponse and the biases associated with nonresponse is basic to all survey research, but there are some specific issues of nonresponse associated with telephone interviewing. One of the major problems is the ambiguity of the results of many attempts at dialing - e.g., continually engaged or no reply, numbers connected to fax machines, computer modems or answering machines. Recently automated screening devices have been developed to identify telephone numbers connected to recordings indicating whether they are not in service (Casady and Lepkowski 1999). Thus proprietary hardware and software
have been developed to detect "tri-tone" recording which indicates "not-in-service" and these numbers when dialed can be removed from the sample. Prior removal of many business phones can be carried out by matching with "Yellow Page" files. These and other methods reduce the costs of screening and the ambiguity of calls that continually receive no reply.

Technological advances, such as "call forwarding" and caller identification enhance the possibilities for nonresponse. In addition refusals are easier over the phone than in face-to-face interviews and breaking off the interview in its midst is also easier. These and other problems of nonresponse for 'cold' telephone interviewing and the US experience in dealing with them are reviewed extensively by Groves and Lyberg (1988a). In particular they follow CASRO (1982) and White (1983) in recommending a definition of nonresponse rates which includes in the denominator an estimate of the number of unanswered numbers that are working numbers in addition to the complete and incomplete interviews, refused eligible numbers and other noninterviewed units. The estimate of the proportion of unanswered numbers that are eligible is obtained as the proportion of answered numbers that are eligible. However this may be a biased estimator. For instance the intensive use of answering technology by businesses implies that practically all businesses will respond and can be identified as businesses. Also, as pointed out by Massey (1995), this measure has to be modified in the case of screening by defining a household screening response rate as the estimated proportion of eligible households identified as such by the screening, rather than the proportion of all households screened for eligibility. Cunningham, Brick and Meader (2000) present several detailed measures of response rates and eligibility rates for each stage of a survey with screening, as well as overall rates, in reporting on the methodology of the National Survey of America's Families.

Telephone nonresponse rates are, in general, higher than those obtained from face-to-face interviews, due to the reasons mentioned above - see Hochstim (1967), Groves and Kahn (1979), Fitti (1979), Groves and Lyberg (1988a) for US experience; Wilson, Blackshaw and Norris (1988), and Collins, Sykes, Wilson and Blackshaw (1988) for experience in UK surveys; and Drew, Choudry and Hunter (1988) for the experience of Canadian government surveys. The latter includes also comparisons of 'cold' and 'warm' telephone interviews, which show only small differences in nonresponse rates. More recently an analysis of the experience in 39 US telephone surveys carried out in the nineties (Massey, O'Connor and Krotki 1997) indicates a slight further reduction in response rates to an average of $62 \%$ and a range from $42 \%$ to $79 \%$ (though it seems that Canadian response rates have not decreased over recent years). Among the factors to which this increase in nonresponse can be attributed are the increase in the use of technological devices (answering machines, call
forwarding, multi -purpose telephone lines) and the increased prevalence of telephone solicitation, already identified as a potential problem for telephone surveys by Biel (1967). The American Statistical Assocation (1999) considers the effect of near saturation calling conducted by telemarketers on lowering survey cooperation rates as a serious challenge not fully addressed by survey researchers. It concludes that unless the trend can be reversed, "telephone surveys, as we know them, could disappear within the next five years". A similar view is expressed by Kalton (2000).

As is the case for telephone noncoverage, the effect of nonresponse on biases in survey estimates is made more severe by the correlation between nonresponse and many socio-economic characteristics. Groves and Lyberg (1988a) on the basis of a review of previous work identify the main correlates of telephone nonresponse. They are age (elderly persons have higher refusal rates - see also Collins et al. 1988) and education (higher nonresponse among lower education groups - see, e.g., Cannel, Groves, Magilavy, Mathiowetz, Miller and Thornberry 1987). On the other hand, there is evidence showing that urban-rural differences in nonresponse are diminished in telephone surveys, as compared with face-to-face surveys - Groves and Kahn (1979). More recent papers on the effects of nonresponse concentrate on specific issues. Thus Diehr, Koepsell, Cheadle and Psaty (1992) investigate the relationship of response rate and other summary variables at the prefix and at the person level. They find relationships between nonresponse and age, race and family size and type. Merkle, Bauman and Lavrakas (1993) in an investigation of the impact of callbacks on the quality of survey estimates show that age and employment status are the major correlates with the number of callbacks required. Kalsbeek and Durham (1994) investigate the effect of nonresponse in a follow-up telephone survey on breastfeeding among low-income women and find that the main correlates with nonresponse are age and degree of urbanization. Finally, multilevel modeling is applied to an extensive meta-analysis of reports on inter-mode comparisons of nonresponse by Hox, DeLeeuw and Kreft (1991). The results, based on the analysis by multi-level modeling of a total of 45 studies ( 35 of which included a telephone component), indicate significantly lower response for telephone studies than for face-to-face studies when models with fixed slopes are used. However when random-slope models are used the difference is no longer significant.

In attempts to reduce nonresponse in telephone surveys the effect of survey operational variables on nonresponse has been investigated. Thus Sebold (1988) finds that doubling the survey period (from two to four weeks) increased the response rate by 3 percentage points in an experiment for the US National Crime Survey. Brick and Collins (1997) investigated the effect of advance letters and screening questions on response in the US National Household Education Survey. They found that a screen-out
question approach increased response rates considerably but that the advance letter did not add to the effect of screening. Other survey variables that have been found to affect response rates are interview length (Collins, et al. 1988) and interviewer vocal characteristics (Oksenberg and Cannel 1988). The effect of the method of selection of sample individuals on nonresponse (in particular the requirement for household rosters) has already been mentioned in section 3.2.2.

Finally, in recent years there has been a significant increase in the use of answering machines and caller D devices for screening unwanted calls, with obvious increased potential for nonresponse. For instance, the proportion of households with answering machines in France increased from $21 \%$ in 1995 to $40 \%$ in 1999 (Rouquette 2000), the same as in Germany (Federal Republic of Germany 1999), while in the US the proportion increased from about $25 \%$ in 1988 (Tuckel and Feinberg 1991) to over 73\% by 1997 (Decision Analyst 1997). However, based on a nationwide telephone survey, Tuckel and Feinberg (1991) reach the conclusion that, in comparison to other initial non-contact groups (e.g., 'no answer' or 'busy'), those with answering machines are more likely to respond and less likely to refuse, resulting in a contact rate which is definitely not smaller than that of other non-contacts. In fact, it seems, according to a study by Oldendick and Link (1994), that the use of answering machines to screen out survey calls is limited to some 2-3 percent. However screeners tend to be in higher income groups, urban and with higher education. Similarly, Piazza (1993) finds on the basis of extensive data from the California Disability Survey, a telephone survey with a high number of callbacks, that although answering machine owners are more difficult to contact initially, once contacted they are at least as likely to respond as those without answering machines. They point out also that reaching an answering machine ensures that a household has been reached and that its residents do not want to miss important calls. In a study by Xu, Bates and Schweitzer (1993), designed to investigate the effect of leaving messages on answering machines, households with answering machines were found to be more likely to be contacted and to complete the interview than those without answering machines. Furthermore leaving a message on the answering machine led to a significant increase in response rate and reduction in refusals. Similarly, Harlow, Crea, East, Oleson, Fraer and Cramer (1993), based on results of a controlled experiment, found that leaving a message on the answering machine led to an increase of $15 \%$ in response, after adjusting for age, interviewer and town of residence. Koepsell, McGuire, Longstreth, Nelson and van Belle (1996) carried out a randomized trial of leaving messages on answering machines and found an overall increase of $20 \%$ in response rate. Although in a similar study Tuckel and Shukers (1997) found no significant effect, the overall findings in a range of studies indicate that the increase in
the use of answering machines has a beneficial effect on survey response, probably due to their providing the possibility of leaving positive messages and thereby enabling the screening out of telemarketing calls.

Tuckel and ONeill (1996) estimate that the percentage of US households with caller ID increased from $3 \%$ in 1992 to $10 \%$ in 1996. Based on a national study, in which the profiles of both caller ID subscribers and answering machine owners are analyzed, they reach the conclusion that these technological devices do not yet present major obstacles for telephone survey research, since their owners tend to use the screening devices primarily to screen out recognized undesirable numbers of acquaintances rather than unrecognized numbers. However, they point out that the possibility of screening will probably lead to increases in answering machine response to repeated callbacks.

### 3.3.3 Weighting and adjustment

Telephone surveys often require special attention to weighting and adjustment. Although sampling designs are usually based on equal probabilities of selection, in practice these are not always achieved. For instance RDD sample designs are theoretically self-weighting but in fact unequal selection probabilities may result due to the multiplicity of telephone lines (numbers) for the same household. In this case, if information is collected on the number of telephone lines to which the household is connected, the required adjustment is straightforward. Similarly reweighting is required to take into account PSU's for which the number of in-scope numbers is less that the required cluster sample size. An additional problem arises due to the fact that it is often difficult to determine whether a telephone, from which no answer can be obtained after repeated attempts, is indeed a case of in-scope nonresponse or is, in fact, out-of-scope. Other problems requiring reweighting are nonresponse, the inherent undercoverage due to nontelephone households and the obvious necessity to use some form of multiplicity estimator for multiple-frame sample designs, based on information on the frames on which the unit is represented.

These problems are dealt with for national RDD samples carried out by the US National Center for Health Statistics in a series of papers by Thornberry and Massey (1978); Botman, Massey and Shimizu (1982); and Massey and Botman (1988). They describe the weighting adjustments carried out for the RDD US National Health Interview Survey (NHIS) and for a smoking survey to account for multiple telephones per household, for telephone coverage and for nonresponse. The adjustments were based on external data for race and geographic region and on survey information on nonresponse and on multiple telephones. Several alternative adjustment and weighting procedures are compared and evaluated. Chapman and Roman (1985) compare substitution with nonresponse adjustment in a feasibility study for the RDD NHIS and report that the results with respect to bias and variance are similar. Drew
and Groves (1989) compare altemative adjustment procedures for unit nonresponse based on external administrative data, on an explicit response prediction model and on response probabilities estimated on the basis of callback data. Casady and Sirken (1980) propose a multiplicity estimator for a multiple-frame sampling design applied to data from the US National Health Interview Survey. Brick (1990) compares the multiplicity estimator with the traditional multiple frame estimator for an educational RDD survey.

Goksel, Judkins and Mosher (1991) report on adjustments, based on modeling nonresponse propensities, for a telephone follow-up of a face-to face interview in the US National Survey of Family Growth. Adjustment based on response propensities by intensity of follow-up effort and by smoking status are proposed for a Canadian survey of attitudes to smoking restrictive legislation by Bull, Pederson and Ashley (1988).

Following a comparison by Keeter (1995) of nontelephone households with 'transient'households (those who recently gained or lost telephone service), Brick, Waksberg and Keeter (1996) propose the use of data on interruptions in telephone service in order to adjust for the undercoverage due to non-telephone households. Their results indicate that such adjustment can lead to a reduction of mean square error. Hoaglin and Battaglia (1996) compare a modified poststratification method and a model-based estimation with simple poststratification for adjusting for noncoverage in an RDD survey of vaccination coverage. The modified poststratification uses national data on vaccination rates for telephone and non-telephone children in addition to demographic and socioeconomic data used for simple poststratification, while the modelased adjustment is based on a logit model to estimate the probability of residing in a telephone household. The results show gains from the use of the modified poststratification but only slight differences between the modified poststratification and the model based adjustment. A similar adjustment based on telephone interruption data is applied by Frankel, Srinath, Battaglia, Hoaglin, Wright and Smith (1999) to NHIS data and shows conclusively a substantial reduction in bias.

### 3.4 Data Quality - Response Errors and Mode Effects

The quality of information obtained over the telephone has always been a controversial issue. As mentioned in section 2, apprehensions on the supposed inferiority of the quality of data from telephone interviewing were allayed at an early stage, to a large degree by some of the extensive empirical appraisals carried out in the sixties and seventies. However there was still some conflicting evidence from different studies on the relative quality of telephone and face-to-face interviewing. Although the intensive analysis of large omnibus surveys carried out under the two modes by the University of Michigan Survey Research Center
(Groves and Kahn 1979), provided important information on data quality and other issues, the mode comparisons and a comparison with external data were not conclusive. In an attempt to resolve the issue, de Leeuw and van der Zoowen (1988) carried out an extensive meta-analysis of 28 major empirical studies in which comparisons of face-to-face and telephone interviewing were investigated. The studies, carried out between 1952 and 1986 on a variety of topics, were primarily from the US but some European studies were also covered. Data quality indicators used were response validity (based on validation studies), absence of social desirability bias, item response, amount of information (for open questions or check-lists) and similarity of response. The overall finding is that if there are any differences in quality between the two modes, they are definitely very minor and that other considerations, such as costs and convenience, should be used in decisions on the use of the telephone for survey work. Similar conclusions are reached for the UK by Sykes and Collins (1988), on the basis of four comparative studies; for income data in Denmark by Körmendi (1988), in a validation study, based on administrative data; and in a comparison of financial data in a Canadian Farm Financial Survey (Caron and Lavallée 1998).

Other recent studies on mode effects concentrate on specific issues and topics but reach similar conclusions. Thus Herzog and Rodgers (1988) report on a mode comparison in a study of older adults and find only small differences. Similar results are reported by Foley and Brook (1990) for a survey on the last days of life. In a study of the sensitive topic of drug use Aquilino and Lo Sciuto (1990) find almost identical results for whites, but some significant differences for blacks, even after controlling for variables possibly related to telephone undercoverage. This may be explained by results reported by Johnson, Fendrich, Shaligram and Garey (1997) for a telephone survey of drug use, which supports a social distance model of interviewer effects.

There is little doubt that interviewers have a great effect on quality, both in face to face and in telephone surveys. The use of central telephone interviewing facilities provides more opportunities to control and monitor interviewer effects than in field interviewing. Some of the issues involved are treated by Stokes and Yeh (1988), who propose a Bayesian model for interviewer effects and methods for estimating the model parameters. A betabinomial model for the interviewer variance component and methods of estimation of its parameters are proposed by Pannekoek (1988).

An effective way of reducing response errors in face-to-face interview surveys has been the use of records provided by the respondent to verify and recall information on income, insurance, health events etc. Obviously, the extension of this method to telephone interviewing involves some problems, since the interviewer cannot see the documents and even asking the respondent to get them may
involve a disruptive break in the telephone interview more frequently than in a face-to-face interview. However the use of records by respondents in telephone surveys can help to reduce response bias. Battaglia, Shapiro and Zell (1996) report on an attempt to ask respondents to use vaccination records in one of the rounds of the US National Immunization Survey and to compare the information obtained with provider records. Some $47 \%$ of the respondents did in fact use vaccination records but substantial underreporting bias was still found, possibly due to the fact that the vaccination reports were not always up to date. Similar effects are found in face-to-face surveys - see Brick, Kalton, Nixon, Givens and Ezzati-Rice (2000).

## 4. CURRENT AND FUTURE TECHNOLOGICAL DEVELOPMENTS

Together with almost complete telephone coverage, the very intensive technological development and the diversity of communications possibilities are continuously opening up new opportunities and potentials for using novel communication options for survey work. On the other hand, some of these developments may cause difficulties for telesurveys under the conventional methodology of today. Thus the increased sophistication of filtering devices and algorithms (as a development of the simple answering machines and caller ID devices mentioned in section 3.3) may make it easier than ever for respondents not to cooperate. In the following we examine present applications and conjectured future developments and comment on the methodological problems involved in their use.

### 4.1 E-Mail and Web Surveys

Internet access for households has experienced a very rapid increase in recent years. For instance in the US the proportion of households with access to the Internet has risen from $26 \%$ in December 1998 to $42 \%$ in August 2000 - NTIA (2000). Other countries have reached somewhat lower levels - the UK $28 \%$ (in August, 2000 - OFTEL 2000), Canada $25 \%$, Finland $22 \%$, France $7 \%$ and Belgium $5 \%$ in 1999, according to Rouquette (2000), Israel $12 \%$ (in 1999 - Central Bureau of Statistics 2000) and Germany $11 \%$ (Federal Republic of Germany 1999). This rapid increase in coverage, is still far off from attaining completeness. Furthermore, there are also some indications that, together with the increase in total use, there is also a growing category of ex-users. Katz and Aspden (1998) report that the proportion of former users of the Internet increased from $8 \%$ to $11 \%$ between 1995 and 1996. However the overall increase in access has encouraged the use of e-mail and the Internet for survey work. While coverage for an e-mail survey (EMS) is comparable to that of a Web (or Intemet) survey and both are based on the use of a computer self-administered questionnaire (CSAQ), there is a basic difference between these two types of
telesurveys. The e-mail survey is very similar to a mail survey, in that it is based on sending out a text questionnaire and asking the respondent to send back the completed questionnaire. The advantage over the mail survey is in cost and in the ease and simplicity of transmission and receipt. The Web survey is, in general, based on interaction between the respondent and the survey instrument, via the use of Java, XML, or a similar instrument. It allows multiple enhancements, such as colour and animation, and extensive possibilities for sophisticated skip patterns and real-time editing. The exciting potential for innovative collection systems based on ever-developing Web tools cannot yet overcome the basic problem inherent in both e-mail and Web surveys that current coverage is completely inadequate for most human populations of interest (Dillman 2000).

Nonetheless, e-mail and Internet surveys can and are being used, with varying degrees of success, for certain populations where coverage is virtually complete or in conjunction with other modes of collection. Thus Couper, Blair, and Triplett (1999) report on an experimental study comparing e-mail and regular mail for a survey of employees in several U.S. government statistical agencies. The sampled employees were randomly assigned to a mail or e-mail mode of data collection and comparable procedures were used for advance contact and follow-up of subjects across modes. The results indicated somewhat higher response rates for mail than for e-mail, but data quality (item missing data) was similar across the two modes. In field tests for the 1999 US National Study of Postsecondary Faculty both administrators and faculty were offered the choice between completing and mailing a conventional paper questionnaire or completing a CSAQ via the Web (Abraham, Steiger and Sullivan 1998). Although it may be assumed that practically all respondents had access to the Web, only $8 \%$ of responding faculty and $17 \%$ of the institution administrators opted for the CSAQ mode. The US National Science Foundation is planning to use a Web-based option in its 1999 National Survey of Recent College Graduates, under the hypothesis that most of the survey population would be relatively computer literate and have access to the Web (Meeks, Lanier, Fecso and Collins 1998). For a review of the use of CSAQ by government agencies and private survey organizations and the problems involved, see Ramos, Sedivi and Sweet (1998).

However, most current Web surveys of general populations are based on non-probability sampling - mostly by some form of self-selection. Fischbacher, Chappel, Edwards and Summerton (1999) report on a meta-analysis of 28 surveys in the health field using e-mail and the Internet. Many of these were epidemiological studies aimed at patients of specific diseases and the problem of selection bias meant that most of the results could not be generalized. One of the largest Web surveys is the WWW User Survey carried out by the Graphics Visualization and Usability

Center at Georgia Institute of Technology (Kehoe, Petkow, Sutton, Aggarwal and Rogers 1999). Although the survey population is defined as Internet users, the lack of any sample framework for this population implies that respondents had to be solicited by various methods (Web and other media announcements, advertising banners, incentive cash prizes etc.), rather than sampled with known probabilities. Although some 20,000 users participated, the survey report points out that the data is biased towards experienced and more frequent users and recommends the augmentation of their data with random sample surveys. In an attempt to overcome the bias inherent in basing surveys on samples of those with internet access only, some commercial survey organizations distribute devices, which let users access the Internet through television sets, to all of its panelists on an RDD sample, to ensure consistent results (Felson 2001). However Poynter (2000) predicts that by the year $200595 \%$ of market research surveys will be conducted via the intemet but that $80 \%$ will be based on respondents who have 'opted in', rather than on probability sampling.

On the other hand, there is evidence that Web-based data collection can be applied with relative success for establishment surveys. Nusser and Thompson(1998) report on its use for the US Department of Agriculture's National Resources Inventory Surveys; Rosen, Manning and Harrel (1998) on Web-based collection from establishments for the US Current Employment Statistics Survey and Meeks et al. (1998) on its use for data collection from academic institutions, federal agencies and private corporations for US National Science Foundation surveys. Assuming that the problem of coverage and sampling will eventually be resolved for households and individuals, this holds hope for Web-based collection for household surveys at some point in the future.

### 4.2 Other Computer Self Administered Questionnaire (CSAQ) and Computer

## Assisted Self Interviewing (CASI) Methods

Couper and Nichols (1998) differentiate between computer self administered questionnaire (CSAQ) collection, in which an interviewer is not present, and computer assisted self interviewing (CASI), in which an interviewer is present or delivers the survey instrument. Thus both e-mail and Internet surveys are based on CSAQ with the assistance of telecommunications technology. Other CSAQ methods are touchtone data entry (TDE), whereby respondents enter data using their touchtone telephones, and interactive voice recognition (IVR) or voice recognition entry (VRE). Both are based on respondents initiating calls to report at their convenience, after initial contact has been established, and have been extensively tested and successfully used by the US Bureau of Labor Statistics for data collection from establishments for its Current Employment Statistics program - Werking, Tupek and Clayton (1988), Winter and Clayton (1990) and Clayton
and Winter (1992). Phipps and Tupek (1991) report on a study of the quality of TDE collection, by means of a record check. Their results show that there are few problems with the method and that response errors diminish with experience. More recently US statistical agencies have initiated tests of the possibility of applying these CSAQ methods to household surveys. McKay, Robison and Malik (1994) report on initial laboratory testing of TDE for the Current Population Survey. Malakhoff and Appel (1997) report on the development of an IVR prototype at the US Bureau of Census, albeit for a listing operation by field staff. It should be noted that while TDE is obviously unique to telephone surveys, IVR could be used for other modes of collection.

Computer assisted self interviewing (CASI) methods include audio (ACASI) and video (VCASI) modes of collection and have long been regarded as the natural extensions of mail surveys that benefit from modern day technology (Dillman 2000). Their usefulness has been especially emphasized for surveys of sensitive and embarrassing topics, where the presence of the interviewer during the interview may make respondents reluctant to answer in a face-to-face interview. For a review of recent advances in these methods see Baker (1998), O'Reilly, Hubbard, Lessler, Biemer and Tumer (1994), Rogers, Miller, Forsyth, Smith and Turner (1996) and Tourangeau and Smith (1998). Practically all the reported applications are of surveys in which the survey instrument is brought to the respondent's home by field staff. The use of the telephone for ACASI (T-ACASI) collection has already been tried - Turner, Forsyth, O'Reilly, Cooley, Smith, Rogers and Miller (1998). The long-expected development of videotelephony to become a widespread common form of telephone service for households has not yet materialized. If and when it occurs it should make telephone VCASI (T-VCASI) possible in the future, with important implications for telesurvey work. The addition of a visual element will help to overcome many of the problems of present day telephone surveys that are not present in face-to-face interviews (eye contact with the interviewer, use of cue cards and other visual aids). The use of videotelephony will probably not be universal for a very long time, so that at least for the time being, T-VCASI will only be able to serve as a supplementary mode of collection.

### 4.3 Mobile Telephones

The problems envisaged for coverage of fixed line RDD surveys due to the rapid proliferation of mobile telephones have been mentioned in section 3.3.1. In the future it is obvious that mobile telephones will have to be used to reach the ever-increasing numbers of households without fixed telephone lines. Present levels of mobile telephone coverage imply that mobile telephone surveys can, in general, only be used for specific populations or for supplementing fixed line RDD surveys. For instance Perone, Matrundola and Soverini (1999) report on a mobile telephone survey for a naturally accessible population - that
of mobile telephone subscribers in order to assess customer satisfaction. Refusal rates were found not to exceed those found in fixed line telephone surveys. However, noncontact rates were high, primarily due to subscribers being outside the signal range or shutting down their telephones. An additional problem associated with mobile phone surveys is that in many cases in North America the subscriber has to pay for received calls - Casady and Lepkowski (1999).

As mentioned above, Cunningham, et al. (1997) report on the use of mobile telephones to interview nontelephone households (primarily in rural areas), with the mobile telephone brought to the respondent by field interviewers. This was designed to minimize mode effects by having telephone interviews conducted by the same interviewers as those conducted for telephone households. The response rates were high, even though in some cases the interviews had to be conducted outdoors in order to obtain reasonable reception. The most intensive use of mobile phones for household surveys is no doubt for the Finish Labour Force Survey - Kuusela and Notkola (1999). Out of some $97 \%$ of interviews completed by telephone, over $20 \%$ are carried out by mobile telephone. Although the average duration of mobile telephone interviews is somewhat longer than those of conventional telephone interviews, this is probably due to socio-demographic differences between the respondent groups.

### 4.4 Future Technological Developments and their Effect on Telesurvey Methodology

The rapid advances in technological developments in the areas of telecommunications and information systems make it very difficult to forecast their influence on survey work. Not all these technological changes will necessarily increase the potential for using advanced telecommunications technology for survey work. The problems raised by persons who have opted to 'drop-out' from the Internet (Katz and Aspden 1998) or from fixed line telephone service (see e.g., Gabler and Haeder 2000; and Kuusela and Vikki 1999) have already been mentioned. Furthermore, in some areas, such as market research and official statistics, technological developments may lead to a reduced reliance on surveys to gather information for decision-making. Thus Baker (1998) and Poynter (2000) predict that techniques such as data mining of existing data resources may become predominant for market research. Similarly, Scheuren and Petska (1993) discuss the possibilities for the use of administrative record systems for official statistics. However, there still remain important areas (for instance for opinions and unobservable behaviour) in which surveys will remain the predominant source of data. The technological advances will open new possibilities for telesurvey work, though the required methodology might become more complex than that used today.

One of the expected developments forecast for the near future is the integration of multiple communication devices
and methods - telephony (fixed line and wireless), fax, internet, e-mail, videotelephony, data transmission, television transmissions etc. - Baker (1998). This implies that each individual will have access to a variety of telecommunication services possibly via the same physical instrument, which could be a mobile phone (e.g., via WAP technology), a PC or a TV set or some combination of these. Similarly, the survey taker may be able to gain access to respondents via several different modes. See Ranta-aho and Leppinen (1997) for some of the issues involved in this plethora of possible avenues of access. It is envisaged that the recipient will have a large degree of control over whether to receive communications at all and, if, so by which mode. This is already now ensured for many users by means of sophisticated devices for screening, forwarding, message transfer, multiple message transmission etc. On the other hand, the degree of control of mode of transmission by the sender will probably decrease as a result.

The implications of these developments for survey work are that mixed mode surveys and possibly multi-frame methodology will have to become predominant. Although we consider that overall telecommunications coverage will increase to some saturation point that is close to universal coverage, it seems unlikely that any given mode of telecommunication will by itself provide virtual complete coverage. Furthermore, even when a single mode may provide practically complete coverage, it is not clear that a mixed mode approach, taking into account respondents' mode preferences, is not preferable. The increased reliance of survey work on the voluntary cooperation of respondents practically dictates that we should offer the respondent the choice of mode. However it should be pointed out that mixed mode surveys are very expensive and that the present technology does not allow the simple transfer of questionnaires developed for one mode (e.g., the CAI Blaise questionnaire) to another mode - e.g., to a paper form.

The major problem that the new developments in telecommunications pose for survey design will probably be the choice of relevant frameworks and the allocation of sample units to modes of collection. Eventually it is envisaged that each individual will have a unique, permanent, personal communication number (or ID) through which he/she can be reached by a multiplicity of modes (written, oral or visual), via a variety of fixed line or wireless devices which could be at home, in the office or mobile. The choice of mode will be ultimately controlled by the joint decision of recipient and sender. While the idea of such a universal number (which would basically be an identity number) is no doubt anathema to libertarians, there is little doubt that it will eventually become acceptable, even if small activist groups may attempt to evade its use and even disrupt its proliferation. In fact standard universal identity number systems have been operating and are well accepted for several decades in many countries in Northern Europe and in Israel. The identity number in these countries is not regarded as confidential information and is widely used for
many administrative and commercial purposes. For example, in Israel personal cheques are required by law to include the person's ID number, name, address and telephone number.

Once such a system of unique communication numbers is operable, standard methods of sampling can be used. It may well be that complete lists of these numbers will be generally available - possibly with only limited geographical or other information. This is the situation with respect to ID's in many national registration systems. There are reasons to expect that a similar situation may prevail for communication numbers - initially at least in Europe rather than in North America. This could come about since the need for unlisted status might well be made redundant because of sophisticated screening techniques. Although screening may enhance the ease of non-response, the possibility of transmitting prior written messages by e-mail or voice mail could reduce the problem.

Sampling from such lists would be simple but in most cases might be inefficient, since it could benefit only marginally from auxiliary information. While differentiation between personal and business contacts might be ensured by the listings, it is doubtful that any household information would be available. This dictates that the sampling and reporting unit would be the individual rather than a household. This is in any case the aim of many surveys and the usefulness of the household as a sampling unit for telesurveys is definitely doubtful, even under current practice. Household information, if required, would have to be obtained from the individual and include information on household size to ensure proper weighting for household characteristics. If the communications numbering system ensures the allocation of a single number to each individual, no information is required on the modes of communication or their multiplicity.

If listings of communication numbers are not available or if the problem of unlisted numbers does persist, some form of RDD will have to be used. This should not differ much from the RDD techniques currently employed. Assuming that the communication numbering system is indeed unique and universal and also arranged by some logic, efficient methods for sampling could easily be developed. Hopefully the numbering system will still bear some relationship to geography, via the individual's permanent address. Otherwise local or even national RDD surveys will become extremely difficult to design efficiently. If sufficient information on the numbering system is available, the extent of out-of-scope numbers could be minimized.

Since it is likely that choice of the mode of communications will be largely under the control of the recipient, the question of allocation of sample units to mode of communication will probably hardly arise. The survey taker will have to prepare a whole range of collection instruments suitable for the different modes of communication. These would have to include written instruments, such as faxed, e-mail
and Internet versions of questionnaires, oral instruments, such as traditional voice interviews and automated interviewing, and combinations of these. The integration of the data obtained from these modes of collection into a uniform data set would be a formidable but surmountable technological challenge.

The almost utopian situation described above will probably take a long time to reach and in the interim suitable methodologies will have to be developed to deal with the problems arising from the short-term developments in communications technology and their application. The necessity to move from telephone surveys based uniquely on fixed line telephones to some combination of mobile and fixed-line telephone situation will have to be dealt with very shortly, as pointed out in section 4.3. Basically multiple frame methodology developed to cover both telephone households and non-telephone households can easily be extended to deal with this. The development of suitable frames and/or RDD sampling methods for mobile telephones still has to be carried out, but the necessary principles are available. The problem of combining data obtained from mobile phones which are basically personal devices with that obtained from fixed-line telephones, which are still fundamentally household devices, will have to be worked out to ensure proper weighting. To ensure this, sufficiently complete information on all the communication devices available to the household is required.

In conclusion, the advances in telesurvey methodology over the past few decades, which have made telephone surveys a viable and predominant survey instrument, will have to be continually updated to deal with the everchanging developments in telecommunications technology and it usage. However the basic elements for these new developments are available and will continue to allow the use of advanced options to obtain high quality survey data

## REFERENCES

ABRAHAM, S.Y., STEIGER, D.M. and SULLIVAN, C. (1998). Electronic and mail self-administered questionnaires: A comparative assessment of use among elite populations. Proceedings of the Section on Survey Research Methods, American Statistical Association, 833-841.

ALEXANDER, C.H. (1988). Cutoff rules for secondary calling in a random digit dialing survey. Telephone Survey Methodology, (R.M. Groves, et al. - Eds.). New York: John Wiley and Sons, 113-126.
AMERICAN STATISTICAL ASSOCIATION (1999). More About Telephone Surveys. ASA series: what is a survey? Section on Survey Research Methods [http://www.amstat.org/sections/srms/ brochures/telephone.pdf].
ANDERSON, J.E., NELSON, D.E. and WILSON, R.W. (1998). Telephone coverage and measurement of health risk indicators: data from the National Health Interview Survey. American Journal of Public Health, 88, 1392-1395.

AQUILINO, W.S., and LO SCIUTO, L.A. (1990). Effects of interview mode on self-reported drug use. Public Opinion Quarterly, 54, 362-395.
BAKER, R.P. (1998). The CASIC future. Computer Assisted Survey Information Collection, (M.P. Couper, et al. - Eds.). NewYork: John Wiley and Sons, 583-604.
BANKS, M.J., and HAGAN, D.E. (1984). Reducing interviewer screening and controlling sample size in a local-area telephone survey. Proceedings of the Section on Survey Research Methods, American Statistical Association, 271-273.
BANKS, R., CHRISTIE, C., CURRALL, J., FRANCIS, J., HARRIS, P., LEE, B., MARTIN, J., PAYNE, C. and WESTLAKE, A. (Eds.) (1999). ASC'99 - Leading Survey \& Statistical Computing into the New Millennium. Proceedings of the ASC International Conference. Association for Survey Computing Chesham, Bucks, UK.

BATTAGLIA, M.P., SHAPIRO, G. and ZELL, E.R. (1996). Substantial response bias may remain when records are used in a telephone survey. Proceedings of the Section on Survey Research Methods, American Statistical Association, 452-455.
BENNET, C.T. (1961). A telephone interview: A method for conducting a follow-up study. Mental Hygiene, 45, 216-220.
BERCINI, D.H., and MASSEY, J.T. (1979). Obtaining the household roster in a telephone survey: The impact of names and placement on response rates. Proceedings of the Social Statistics Section, American Statistical Association, 136-140.
BERRY, S.H., and O'ROURKE, D. (1998). Administrative designs for centralized telephone survey centers: Implications of the transition to CATI. Telephone Survey Methodology, (R.M. Groves, et al. - Eds.). New York: John Wiley and Sons, 457-474.
BIEL, A. L. (1967). Abuses of survey research techniques: the phony interview. Public Opinion Quarterly, 31, 298.
BIEMER, P.P. (1983). Optimal dual frame sample design: Results of a simulation study. Proceedings of the Section on Survey Research Methods, American Statistical Association, 630-635.
BINSON, D., CANCHOLA, J.A. and CATANIA, J.A. (2000). Random selection in a national telephone survey: A comparison of the Kish, next-birthday, and last-birthday methods. Journal of Official Statistics, 16, 53-59.

BLAIR, J., and CZAJA, R. (1982). Locating a special population using random digit dialing. Public Opinion Quarterly, 46, 585-590.

BLANKENSHIP, A.B. (1977a). Professional Telephone Surveys. New York: McGraw Hill.
BLANKENSHIP, A.B. (1977b). Listed versus unlisted numbers in telephone-survey samples. Journal of Advertising Research, 39-42.
BOTMAN, S.L., and ALLEN, K. (1990). Some effects of undercoverage in a telephone survey of teenagers. Proceedings of the Section on Survey Research Methods, American Statistical Association, 396-400.
BOTMAN, S.L., MASSEY, J.T. and SHIMIZU, I.M. (1982). Effect of weighting adjustments on estimates from a random-digit-dialed telephone survey. Proceedings of the Section on Survey Research Methods, American Statistical Association, 139-144.

BRICK, J.M. (1990). Multiplicity sampling in an RDD telephone survey. Proceedings of the Section on Survey Research Methodology, American Statistical Association, 296-301.

BRICK, J.M., and COLLINS, M.A. (1997). A response rate experiment for RDD surveys. Proceedings of the Section on Survey Research Methods, American Statistical Association, 1052-1057.

BRICK, J.M., KALTON, G., NIXON, M., GIVENS, J. and EZZATI-RICE, T. (2000). Statistical issues in a record check study of childhood immunizations. Proceedings of the 1999 Federal Committee on Statistical Methodology Research Conference: Statistical Policy Working Paper, 30, 625-634.

BRICK, J.M., and WAKSBERG, J. (1991). Avoiding sequential sampling with random digit dialing. Survey Methodology, 17, 27-42.
BRICK, J.M., WAKSBERG, J. and KEETER, S. (1996). Using data on interruptions in telephone service as coverage adjustments. Survey Methodology, 22, 185-197.

BRICK, J. M., WAKSBERG, J., KULP, D. and STARER, A. (1995). Bias in list-assisted telephone samples. Public Opinion Quarterly, 59, 218-235.

BRUNNER, J.A., and BRUNNER, G.A. (1971). Are voluntary unlisted telephone subscribers really different? Journal of Marketing Research, 8, 121-124.
BRYANT, B.E. (1975). Respondent selection in a time of changing household composition. Journal of Marketing Research, 12, 129-135.
BRYSON, M.C. (1976). The literary digest poll: Making of a statistical myth. The American Statistician, 30, 184-185.
BULL, S.B., PEDERSON, L.L. and ASHLEY, M.J. (1988). Intensity of follow up; effects on estimates in a population telephone survey with an extension of Kish's (1965) approach. American Journal of Epidemiology, 127, 552-561.
BURKE, J., MORGANSTEIN, D. and SCHWARTZ, S. (1981). Toward the design of an optimal telephone sample. Proceedings of the Section on Survey Research Methods, American Statistical Association, 448-453.

CAHALAN, D. (1960). Measuring newspaper readership by telephone: two comparisons with face to face interviews. Journal of Advertising Research, 1, 2, 1-6.
CAHALAN, D. (1989). Comment: The digest poll rides again! Public Opinion Quarterly, 53, 129-133.
CAMPBELL, J., and PALIT, C.D. (1988). Total digit dialing for a small area census by phone. Proceedings of the Section on Survey Research Methods, American Statistical Association, 549-551.
CANNELL, C.F., GROVES, R.M., MAGILAVY, L.J., MATHIOWETZ, N.A., MILLER, P.V. and THORNBERRY, O.T. (1987). An experimental comparison of telephone and personal health interview surveys. Vital and Health Statistics, Series 2, 106, Public Health service.
CARON, P., and LAVALLEE, P. (1998). Comparison study on the quality of financial data collected through personal and telephone interviews. Proceedings of the Section on Survey Research Methods, American Statistical Association, 208-213.

CASADY, R.J., and LEPKOWSKI, J.M. (1991). Optimal allocation for stratified telephone survey design. Proceedings of the Section on Survey Research Methods, American Statistical Association, 111-116.

CASADY, R.J., and LEPKOWSKI, J.M. (1993). Stratified telephone survey designs. Survey Methodology, 19, 103-113.
CASADY, R.J., and LEPKOWSKI, J.M. (1998). Telephone Sampling. Encyclopedia of Biostatistics. New York: John Wiley and Sons, 4498-4511.

CASADY, R.J., and LEPKOWSKI, J.M. (1999). Telephone Sampling. Sampling of Populations: Methods and Applications third edition, (P.S. Levy and S. Lemeshow, Eds.). New York: John Wiley and Sons, 455-479.

CASADY, R.J., and SIRKEN, M.G. (1980). A multiplicity estimator for multiple frame sampling. Proceedings of the Section on Survey Research Methods, American Statistical Association, 601-609.

CASRO (1982). Report of the Council of American Survey Research Organization Completion Rate Task Force. New York: Audits and Surveys Inc. (Unpublished report).

CENTRAL BUREAU OF STATISTICS (2000). The Household Expenditure Survey 1999. Special Publication 1147. Jerusalem.

CHAPMAN, D.W., and ROMAN, A.M. (1985). An investigation of substitution for an RDD survey. Proceedings of the Section on Survey Research Methods, American Statistical Association, 269-274.

CHOUDHRY, G.H. (1989). Cost-variable optimization of dual frame design for estimating proportions. Proceedings of the Section on Survey Research Methods, American Statistical Association, 566-571.

CLAYTON, R.L., and WINTER, D. L. S. (1992). Speech data entry: results of a test of voice recognition for survey data collection. Journal of Official Statistics, 8, 377-388.

COLLINS, M. (1983). Computer assisted telephone interviewing in the UK. American Statistical Association, Proceedings of the Section on Survey Research Methods, 636-639.

COLLINS, M. (1999). Editorial: sampling for UK telephone surveys. Journal of the Royal Statistical Society, A 162, 1-4.
COLLINS, M., SYKES, W., WILSON, P. and BLACKSHAW, N. (1988). Nonresponse: the UK experience. Telephone Survey Methodology, (R.M. Groves, et al. - Eds.). New York: John Wiley and Sons, 213-232.

COLOMBOTOS, J. (1965). The effects of personal vs. telephone interviews on socially acceptable responses. Public Opinion Quarterly, 29, 457-458.

COOMBS, L., and FREEDMAN, R. (1964). Use of telephone interviews in a longitudinal fertility study. Public Opinion Quarterly, 28, 112-117.
COOPER, S. L. (1964). Random sampling by telephone: an improved method. Journal of Marketing Research, 1, 45-48.

COUPER, M.P., BAKER, R.P., BETHLEHEM, J., CLARK, C.Z F., MARTIN, J., NICHOLLS, W.L., II and O'REILLY, J.M. - (Eds.) (1998). Computer Assisted Survey Information Collection. New York: John Wiley and Sons.

COUPER, M.P., BLAIR, J. and TRIPLETT, T. (1999). A comparison of mail and e-mail for a survey of employees in US statistical agencies. Journal of Official Statistics, 15, 39-56.

COUPER, M.P., and NICHOLLS, W.L., II (1998). The history and development of computer assisted survey information collection methods. Computer Assisted Survey Information Collection, (M.P. Couper, et al. - Eds.). NewYork: John Wiley and Sons, 1-22.
CUNNINGHAM, J.M., WESTERMAN, H.H. and FISCHOFF, J. (1956). A follow-up study of patients seen in a psychiatric clinic for children. American Journal of Orthopsychiatry, 26, 602-610.
CUNNINGHAM, P., BERLIN, M., MEADER, J., MOLLOY, K., MOORE, D. and PAJUNEN, S. (1997). Using cellular telephones to interview nontelephone households. Proceedings of the Section on Survey Research Methods, American Statistical Association, 250-254.
CUNNINGHAM, P., BRICK, J.M. and MEADER, J. (2000). 1999 NSAF In-Person Survey Methods Report No. 5. Washington, DC: Urban Institute. [http://newfederalism.urban.org/nsaf/ methodology_rpts/ 1999_Methodology_5.pdf].

CUMMINGS, M.K. (1979). Random digit dialing: a sampling technique for telephone surveys. Public Opinion Quarterly, 43, 233-244.

CZAJA, R., BLAIR, J., and SEBESTIK, J.P. (1982). Respondent selection in a telephone survey: a comparison of three techniques, Journal of Marketing Research, 19, 381-385.
DECISION ANALYST (1997). More households using answering machines. News Release, October 15, 1997. [http://www.decisionanalyst.com/publ_data/1997/ansmachi.htm].
DEKKER, F., and DORN, P.K. (1984). Computer Assisted Telephone Interviewing: A research project in the Netherlands. Paper presented at: Conference of the Institute of British Geographers.
DE LEEUW, E.D., and VAN DER ZOOWEN, J. (1988). Data quality in telephone and face to face surveys: a comparative meta-analysis. Telephone Survey Methodology, (R.M. Groves, et al. - Eds.). New York: John Wiley and Sons, 283-299.
DIEHR, P., KOEPSELL, T.D., CHEADLE, A., and PSATY, B.M. (1992). Assessing response bias in random-digit dialing surveys: The telephone-prefix method. Statistics in Medicine, 11, 1009-1021.
DILLMAN, D.A. (1978). Mail and Telephone Surveys: The Total Design Method. New York: John Wiley and Sons.

DILLMAN D. A. (2000). Mail and Internet Surveys: The Total Design Method (2nd edition). New York: John Wiley and Sons.

DREW, J.D., CHOUDHRY, G.H. and HUNTER, L.A. (1988). Nonresponse issues in government telephone surveys. Telephone Survey Methodology, (R.M. Groves, et al. - Eds.). New York: John Wiley and Sons, 233-246.

DREW, J.H., and GROVES, R.M. (1989). Adjusting for nonresponse in a telephone subscriber survey. Proceedings of the Section on Survey Research Methods, American Statistical Association, 452-456.

DUTKA, S., and FRANKEL, L. R. (1980). Sequential survey design through the use of computer assisted telephone interviewing. Proceedings of the Business and Economic Statistics Section, American Statistical Association, 73-76.

EASTLACK, J.O., JR. (1964). Recall of advertising by two telephone samples. Journal of Advertising Research, 4, 25-29.
EASTLACK, J.O., JR., and ASSAEL, H. (1966). Better telephone surveys through centralized interviewing. Journal of Advertising Research, 6, 1, 2-7.
FEDERAL COMMITTEE ON STATISTICAL METHODOLOGY. (1984). The Role of Telephone Data Collection in Federal Statistics. Statistical Policy Working Paper 12, Washington, D.C.
FEDERAL COMMITTEE ON STATISTICAL METHODOLOGY. (1990). Computer Assisted Survey Information Collection, Statistical Policy Working Paper 19, Washington, D.C.

FEDERAL REPUBLIC OF GERMANY (1999). Continuous family budget surveys for January 1999. Statistiches Bundesamt: Press release, 20 December, 1999. [http://www.statistik-bund.de/presse/ englisch/pm1999/p4350024.htm].
FELSON, L. (2001). Netting limitations: online researchers' new tactics for tough audiences. Marketing News (American Marketing Association), 35, 5 [http://www.ama.org/pubs/ article.asp?id=4881].
FINK, J.C. (1983). CATI's first decade: The Chilton experience. Sociological Methods and Research, 12, 153-168.
FISCHBACHER, C., CHAPPEL, D., EDWARDS, R. and SUMMERTON, N. (1999). The use and abuse of the Internet for survey research. Proceedings of the Association for Survey Computing 3rd International Conference, Edinburgh, 501-507.
FITTI, J.E. (1979). Some results from the telephone health interview system. Proceedings of the Section on Survey Research Methods, American Statistical Association, 244-249.

FOLEY, D.J., and BROCK, D.B. (1990). Comparison of in-person and telephone responses in a survey of the last days of life. Proceedings of the Section on Survey Research Methods, American Statistical Association, 382-386.
FORD, E.S. (1998). Characteristics of survey participants with and without a telephone; findings from the third National Health and Nutrition Examination Survey. Journal of Clinical Epidemiology, 51, 55-60.

FORSMAN, G. (1993). Sampling individuals within households in telephone surveys. Proceedings of the Section on Survey Research Methods, American Statistical Association, 1113-1118.
FORSMAN, G., and DANIELSSON, S. (1997). Can plus digit sampling generate a probability sample? Proceedings of the Section on Survey Research Methods, American Statistical Association, 958-963.

FOX, A., and RILEY, J. P.(1996). Telephone coverage, housing quality and rents: RDD survey biases. Proceedings of the Section on Survey Research Methods, American Statistical Association, 515-519.

FRANKEL, M.R., SRINATH, K.P., BATTAGLIA, M.P., HOAGLIN, D.C., WRIGHT, R.A. and SMITH, P.J. (1999). Reducing nontelephone bias in RDD surveys. Proceedings of the Section on Survey Research Methods, American Statistical Association, 934-937.

FREEMAN, H.E., and SHANKS, J. M. - Eds. (1983). The emergence of computer-assisted survey research. Sociological Methods and Research, 12 (special issue), 115-230.

FREJEAN, M., PANZANI, J-P. and TASSI, P. (1990). Les menages inscrits en liste rouge et les enquêtes par téléphone. Journal de la Société de Statistique de Paris, 131, Nos. 3-4, 86-102.
FREY J. H. (1989). Survey Research by Telephone (2nd edition). Beverly Hills, CA: Sage Publications.
FRY, H.G., and MCNAIRE, S. (1958). Data gathering by long distance telephone. Public Health Records, 73, 831-835.
GABLER, S., and HAEDER, S. (2000). Telephone sampling in Germany. Paper presented at Fifth International Conference of Social Science Methodology, Koln.
GENESYS (1996). Unlisted numbers: what's really important. Genesys News (Genesys Sampling Systems, Fort Washington, PA), 1-2.
GHOSH, D. (1984). Improving the plus 1 method of random digit dialing. Proceedings of the Section on Survey Research Methods, American Statistical Association, 285-288.
GIESBRECHT, L.H., KULP, D.W. and STARER, A.W. (1996). Estimating coverage bias in RDD samples with current population survey data. Proceedings of the Section on Survey Research Methods, American Statistical Association, 503-508.
GLASSER, G.J., and METZGER, G.D. (1972). Random digit dialing as a method of telephone sampling. Journal of Marketing Research, 9, 59-64.
GLASSER, G. J., and METZGER, G. D. (1975). National estimates of nonlisted telephone households. Joumal of Marketing Research, 12, 359-361.
GOKSEL, H., JUDKINS, D.R. and MOSHER, W.D. (1991). Nonresponse adjustments for a telephone follow-up to a national in-person survey. Proceedings of the Section on Survey Research Methods, American Statistical Association, 581-586.

GROVES, R.M. (1977). An Empirical Comparison of Two Telephone Sample Designs. Unpublished report of the Survey Research Center, the University of Michigan, Ann Arbor, MI.
GROVES, R.M., BIEMER, P.P., LYBERG, L.E., MASSEY, J.T., NICHOLLS, W.L., II and WAKSBERG, J. - Eds. (1988). Telephone Survey Methodology. New York: John Wiley and Sons.
GROVES R.M., and KAHN, R.L. (1979). Surveys by Telephone: A National Comparison With Personal Interview. New York: Academic Press.
GROVES, R.M., and LEPKOWSKI, J.M. (1985). Dual frame, mixed mode survey designs. Journal of Official Statistics, 1, 264-286.
GROVES, R.M., and LEPKOWSKI, J.M. (1986). An experimental implementation of a dual frame telephone sample design. Proceedings of the Section on Survey Research Methods, American Statistical Association, 340-345.
GROVES, R.M., and LYBERG, L.E. (1988a). An overview of nonresponse issues in telephone surveys. Telephone Survey Methodology, (R.M. Groves, et al. - Eds.). New York: John Wiley and Sons, 191-211.
GROVES, R.M., and LYBERG, L.E. - Eds. (1988b). Telephone survey methodology. Journal of Official Statistics (special issue), 4, 283-416.
GUNN, W.J., and RHODES, I.N. (1981). Physician response rates to a telephone survey: effects of monetary incentive level. Public Opinion Quarterly, 45, 109-115.

HAGAN, D. E., and MEIER C. C. (1983). Must respondent selection procedures for telephone surveys be invasive? Public Opinion Quarterly, 47, 547-556.
HARLOW, B.L., CREA, E.C., EAST, M.A., OLESON, B., FRAER, C.J. and CRAMER, D.W. (1993). Telephone answering machines: the influence of leaving messages on telephone interviewing response rates. Journal of Epidemiology, 4, 380-383.
HARTGE, P., BRINTON, L.A., ROSENTHAL, J.F., CAHILL, J.I., HOOVER, R.N. and WAKSBERG, J. (1984). Random digit dialing in selecting a population-based control group. American Journal of Epidemiology, 120, 825-833.
HAUCK, M., and COX, M. (1974). Locating a sample by random digit dialing. Public Opinion Quarterly, 38, 253-260.
HERZOG, A.R., and RODGERS, W.L. (1988). Interviewing older adults: mode comparison using data from a face-to-face survey and a telephone resurvey. Public Opinion Quarterly, 52, 84-99.
HOAGLIN, D.C., and BATTAGLIA, M.P. (1996). A comparison of two methods of adjusting for noncoverage of nontelephone households in a telephone survey. Proceedings of the Section on Survey Research Methods, American Statistical Association, 497-502.
HOCHSTIM, J.R. (1967). A critical comparison of three strategies of collecting data from households. Journal of the American Statistical Association, 62, 976-989.
HOGUE, C.R., and CHAPMAN, D.W. (1984). An investigation of PSU cutoff points for a random digit dialing survey. Proceedings of the Section on Survey Research Methods, American Statistical Association, 286-291.

HOX, J.J., DE LEEUW, E.D. and KREFT, I.G.G. (1991). The effect of interviewer and respondent characteristics on the quality of survey data: a multilevel model. Measurement errors in surveys (P.P. Biemer, L.E. Lyberg, N.A. Mathiowetz and S. Sudman Eds.). New York: John Wiley and Sons, 439-461.
INGLIS, K.M., GROVES, R.M. and HEERINGA, S.G. (1987). Telephone sample designs for the U.S. Black household population. Survey Methodology, 13, 1-14.
JANOFSKY, A.I. (1971). Affective self-disclosure in telephone versus face-toface interviews. Journal of Humanistic Psychology, 11, 93-103.
JOHNSON, T., FENDRICH, M., SHALIGRAM, C. and GAREY, A. (1997). A comparison of interviewer effects models in an RDD telephone survey of drug use. Proceedings of the Section on Survey Research Methods, American Statistical Association, 887-892.
KALSBEEK, W.D., and DURHAM, T.A. (1994). Nonresponse and its effects in a followup telephone survey of low-income women. Proceedings of the Section on Survey Research Methods, American Statistical Association, 943-948.

KALTON, G. (2000). Developments in survey research in the past 25 years. Survey Methodology, 26, 3-10.
KATZ, D., and CANTRIL, H. (1937). Public opinion polls. Sociometry, 1, 155-179.
KATZ, J.E., and ASPDEN, P. (1998). Internet dropouts in the USA. Telecommunications Policy, 22, 4/5, 327-339.

KEETER, S. (1995). Estimating telephone noncoverage bias with a telephone survey. Public Opinion Quarterly, 59, 196-217.
KEHOE, C., PITKOW, J., SUTTON, K., AGGARWAL, G. and ROGERS, J.D. (1999). Results of GVU's Tenth World Wide Web User Survey. Atlanta, GA: Graphics Visualization and Usability Center, College of Computing, Georgia Institute of Technology. [http://www.gvu.gatech.edu/user_surveys].
KHURSHID, A., and SAHAI, H. (1995). A bibliography on telephone survey methodology. Journal of Official Statistics, 11, 325-367.
KISH, L. (1949). A procedure for objective respondent selection within the household. Journal of the American Statistical Association, 44, 380-387.
KOEPSELL, T.D., MCGUIRE, V., LONGSTRETH, Jr., W.T., NELSON, L.M. and VAN BELLE, G. (1996). Randomized trial of leaving messages on telephone answering machines for control recruitment in an epidemiological study. American Journal of Epidemiology, 144, 704-706.
KÖRMENDI, E. (1988). The quality of income information in telephone and face to face surveys. Telephone Survey Methodology (R.M. Groves, et al. - Eds.). New York: John Wiley and Sons, 341-356.

KUUSELA, V., and NOTKOLA, V. (1999). Survey quality and mobile phones. Paper presented at Intemational Conference on Survey Nonresponse, Portland OR.

KUUSELA, V., and VIKKI, K. (1999). Change of telephone coverage due to mobile phones. Paper presented at International Conference on Survey Nonresponse, Portland OR.
LANDON, E.L., JR., and BANKS, S.K. (1977). Relative efficiency and bias of plus-one telephone sampling. Journal of Marketing Research, 14, 294-299.

LARSON, O. N. (1952). The comparative validity of telephone and face-to-face interviews in the measurement of message diffusion from leaflets. American Sociological Review, 17, 471-476.
LAVRAKAS, P.J. (1993). Telephone Survey Methods: Sampling, Selection and Supervision (2nd edition). Newbury Park, CA: Sage Publications.

LAVRAKAS, P.J., BAUMAN, S.L. and MERKLE, D.M. (1993). The last-birthday method and within-unit coverage problems. Proceedings of the Section on Survey Research Methods, American Statistical Association, 1107-1 112.
LEPKOWSKI, J.M. (1988). Telephone sampling methods in the United States. Telephone Survey Methodology, (R.M. Groves, et al. - Eds.). New York: John Wiley and Sons, 73-98.
LEPKOWSKI, J.M., and GROVES, R.M. (1984). The impact of bias on dual frame survey design. Proceedings of the Section on Survey Research Methods, American Statistical Association, 265-270.
LEPKOWSKI, J.M., and GROVES, R.M. (1986a). A two phase probability proportional to size design for telephone sampling. Proceedings of the Section on Survey Research Methods, American Statistical Association, 357-362.

LEPKOWSKI, J.M., and GROVES, R.M. (1986b). A mean square error model for dual frame, mixed mode survey design. Journal of the American Statistical Association, 81, 930-937.

LEUTHOLD, D.A., and SCHEELE, R. (1971). Patterns of bias in samples based on telephone directories. Public Opinion Quarterly, 35, 249-257.

LOCANDER, W., SUDMAN, S. and BRADBURN, N. (1976). An investigation of interview method, threat and response distortion. Journal of the American Statistical Association, 71, 269-275.
LYBERG, L., and KASPRZK, D. (1991). Data collection methods and measurement error: an overview. Measurement Errors in Surveys (P.P. Biemer, L.E. Lyberg, N.A. Mathiowetz and S. Sudman - Eds.). New York: John Wiley and Sons, 237-258.
MCCARTHY, W.F., and BATEMAN, D.V. (1988). The use of mathematical programming for designing dual frame surveys. Proceedings of the Section on Survey Research Methods, American Statistical Association, 652-653.
MCKAY, R.B., ROBISON, E.L. and MALIK, A.B. (1994). Touch-tone data entry for household surveys: research findings and possible applications. Proceedings of the Section on Survey Research Methods, American Statistical Association, 509-511.
MAFFEO, C., FREY, W. and KALTON, G. (2000). Survey design and data collection issues in the Disability Evaluation Study. Proceedings of the Section on Survey Research Methods, American Statistical Association, forthcoming.
MAKLAN, D., and WAKSBERG, J. (1988). Within household coverage in RDD surveys. Telephone Survey Methodology (R. M. Groves, et al. - Eds). New York: John Wiley and Sons, 51-69.

MALAKHOFF, L.A., and APPEL, M.V. (1997). The development of a voice recognition prototype for field listing. Proceedings of the Section on Survey Research Methods, American Statistical Association, 234-238.

MASON, R.E., and IMMERMAN, F.W. (1988). Minimum cost sample allocation for Mitofsky-Waksberg random digit dialing. Telephone Survey Methodology, (R.M. Groves, et al. - Eds.). New York: John Wiley and Sons, 127-141.

MASSEY, J.T. (1995). Estimating the response rate in a telephone survey with screening. Proceedings of the Section on Survey Research Methods, American Statistical Association, 673-677.
MASSEY, J T., and BOTMAN, S.L. (1988). Weighting adjustments for random digit dialed surveys. Telephone Survey Methodology, (R.M. Groves et al. - Eds.). New York: John Wiley and Sons, 143-160.
MASSEY, J.T., O'CONNOR, D. and KROTKI, K. (1997). Response rates in random digit dialing (RDD) telephone surveys. Proceedings of the Section on Survey Research Methods, American Statistical Association, 707-712.
MEEKS, R.L., LANIER, A.T., FECSO, R.S. and COLLINS, M.A. (1998). Web-based data collection in national science foundation surveys. Proceedings of the Section on Survey Research Methods, American Statistical Association, 349-353.

MERKLE, D.M., BAUMAN, S.L. and LAVRAKAS, P.J. (1993). The impact of callbacks on survey estimates in an annual RDD survey. Proceedings of the Section on Survey Research Methods, American Statistical Association, 1070-1075.
MITCHELL, G.H., and ROGERS, E.M. (1958). Telephone interviewing. Journal of Farm Economics, 40, 743-747.

MITOFSKY, W. (1970). Sampling of Telephone Households. Unpublished CBS memorandum.
MOHADJER, L. (1988). Stratification of prefix areas for sampling rare populations. Telephone Survey Methodology, (R.M. Groves, et al.- Eds.). New York: John Wiley and Sons, 161-173.
MULLET, G.M. (1982). The efficacy of plus-one dialing: selfreported status. Proceedings of the Section. on Survey Research Methods, American Statistical Association, 575-576.
NATHAN, G., and AFRAMIAN, N. (1996). An experiment with CATI in Israel. Paper presented at InterCasic 96 Conference, San Antonio, TX.
NATHAN, G., and ELIAV, T. (1988). Comparison of measurement errors for telephone interviewing and home visits by misclassification models. Journal of Official Statistics, 4, 363-374.
NICHOLLS, W.L., II (1983). CATI research and development at the Census Bureau. Sociological Methods and Research, 12, 191-198.
NICHOLLS, W.L., II (1988). Computer-assisted telephone interviewing: A general introduction. Telephone Survey Methodology, (R.M. Groves, et al. - Eds.). New York: John Wiley and Sons, 377-386.
NICOLAAS, G., LYNN, P. and LOUND, C. (2000). Random digit dialling in the UK: viability of the sampling method revisited. Paper presented at Fifth International Conference of Social Science Methodology, Koln.
NORRIS, D.A., and PATON, D.G. (1991). Canada's General Social Survey: five years of experience. Survey Methodology, 17, 227-240.
NTIA (2000). Falling Through the Net, Toward Digital Inclusion. Washington DC: National Telecommunications and Information Administration.
NUSSER, S., and THOMPSON, D. (1998). Web-based survey tools. Proceedings of the Section on Survey Research Methods, American Statistical Association, 951-956.
OAKES, R.H. (1954). Differences in responsiveness in telephone versus personal interviews. Journal of Marketing, 19, 169.
OKSENBERG, L., and CANNEL, C. (1988). Effects of interviewer vocal characteristics on nonresponse. Telephone Survey Methodology, (R.M. Groves, et al. - Eds.). New York: John Wiley and Sons, 257-269.
OLDENDICK, R.W., BISHOP, G.F., SORENSON, S.B. and TUCHFARBER, A.J. (1988). A comparison of the Kish and last birthday methods of respondent selection in telephone surveys. Journal of Official Statistics, 4, 307-318.
OLDENDICK R.W., and LINK, M.W. (1994). The answering machine generation: who are they and what problem do they pose for survey research? Public Opinion Quarterly, 58, 264-273.
OFTEL (1999). Homes Without a Fixed Line Phone - Who Are They? [http://www.oftel.gov.uk/publications/research/unph0400.htm].
OFTEL (2000). Consumers' use of Internet. [http://www.oftel.gov.uk/publications/research/int1000.htm]
O'REILLY, J.M., HUBBARD, M.L., LESSLER, J.T., BIEMER, P.P. and TURNER, C.F. (1994). Audio and video computer assisted self-interviewing; preliminary tests of new technologies for data collection. Journal of Official Statistics, 10, 197-214.

O'ROURKE, D., and BLAIR, J. (1983). Improving random respondent selection in telephone surveys. Journal of Marketing Research, 20, 428-432.
PALIT, C.D. (1980). A microcomputer based computer assisted interviewing system. Proceedings of the Section on Survey Research Methods, American Statistical Association, 243-244.
PALIT, C.D. (1983). Design strategies in RDD sampling. Proceedings of the Section on Survey Research Methods, American Statistical Association, 627-629.
PALIT, C.D., and BLAIR, J. (1986). Some alternatives for the treatment of first phase telephone numbers in a Waksberg-Mitofsky RDD sample. Proceedings of the Section on Survey Research Methods, American Statistical Association, 363-369.

PALIT, C.D., and SHARP, H. (1983). Microcomputer-assisted telephone interviewing. Sociological Methods and Research, 12, 169-189.
PANNEKOEK, J. (1988), Interviewer variance in a telephone survey. Journal of Official Statistics, 4, 375-384.
PAYNE, S. L. (1956). Some advantages of telephone surveys. Journal of Marketing, 20, 278-281.
PERNEGER, T.V., MYERS, T.L., KLAG, M.J. and WHELTON, P.K. (1993). Effectiveness of the Waksberg telephone sampling method for the selection of population controls. American Journal of Epidemiology, 138, 574-584.
PERONE, C., MATRUNDOLA, G. and SOVERINI, M. (1999). A quality control approach to mobile phone surveys; the experience of Telecom Italia Mobile. Proceedings of the Association for Survey Computing 3rd International Conference, Edinburgh, 180-187.
PERRY, J. B. (1968). A note on the use of telephone directories as a sample source. Public Opinion Quarterly, 32, 691-695.
PHIPPS, P.A., and TUPEK, A.R.(1991). Assessing measurement errors in a touchtone recognition survey. Survey Methodology, 17, 15-26.
PIAZZA, T. (1993). Meeting the challenge of answering machines. Public Opinion Quarterly, 57, 219-231.

POTTER, F.J., MCNEILL, J.J., WILLIAMS, S.R. and WAITMAN, M.A. (1991). List-assisted RDD telephone surveys. Proceedings of the Section on Survey Research Methods, American Statistical Association, 117-122.

POTTHOFF, R.F. (1987a). Some generalizations of the Mitofsky-Waksberg technique of random digit dialing. Journal of the American Statistical Association, 82, 409-418.
POTTHOFF, R.F. (1987b). Generalizations of the Mitofsky-Waksberg technique for random digit dialing: some added topics. Proceedings of the Section on Survey Research Methods, American Statistical Association, 615-620.

POYNTER, R. (2000). We've Got Five Years. London: Association for Survey Computing meeting on Survey Research on the Internet.

RAMOS, M., SEDIVI, B.M. and SWEET, E.M. (1998). Computerized self-administered questionnaires. Computer Assisted Survey Information Collection (M.P. Couper, et al. Eds.). NewYork: John Wiley and Sons, 389-408.

RANTA-AHO, M., and LEPPINEN, A. (1997). Matching telecommunication services with user communication needs. Proceedings of the International Symposium on Human Factors in Telecommunications, (K. Nordby and L. Grafisk - Eds.). Oslo, Norway, 401-408. [http://www.comlab.hut.fi/hft/publications/ matcharticle.pdf].
RICH, C.L. (1977). Is random digit dialing really necessary? Journal of Marketing Research, 14, 300-305.
ROGERS, T.F. (1976). Interviews by telephone and in person: quality of responses and field performance. Public Opinion Quarterly, 40, 51-65.
ROGERS, S.M., MILLER, H.G., FORSYTH, B.H., SMITH, T.K. and TURNER, C.F. (1996). Audio-CASI: the impact of operational characteristics on data quality. Proceedings of the Section on Survey Research Methods, American Statistical Association, 1042-1047.
ROMUALD, K.S., and HAGGARD, L.M.(1994). The effect of varying the respondent selection script on respondent self-selection in RDD telephone surveys. Proceedings of the Section on Survey Research Methods, American Statistical Association, 1299-1304.
ROSEN, R.J., MANNING, C.D. and HARRELL, L.J., Jr. (1998). Web-based data collection in the current employment statistics survey. Proceedings of the Section on Survey Research Methods, American Statistical Association, 354-359.
ROSLOW, S., and ROSLOW, L. (1972). Unlisted phone subscribers are different. Journal of Advertising Research, 7, 4, 35-38.

ROUQUETTE, C. (2000). La percée du telephone portable et d'Internet. INSEE Première No. 200. INSEE, Paris. [http://www.insee.ff/ff/ffc/docs_ffc/ip700.pdf].
ST. CLAIR, J., and MUIR, J. (1997). Household adoption of digital technologies. Year Book Australia 1997. Canberra: Australian Bureau of Statistics.
SALMON, C.T., and NICHOLS, J.S. (1983). The next birthday method of respondent selection. Public Opinion Quarterly, 47, 270-276.
SCHEUREN, F., and PETSKA, T. (1993). Turning administrative systems into information systems. Journal of Official Statistics, 9, 109-119.

SCHMIEDESKAMP, J. W. (1962). Reinterviews by telephone. Journal of Marketing, 26, 28-34.
SEBOLD, J. (1988). Survey period length, unanswered numbers, and nonresponse in telephone surveys. Telephone Survey Methodology, (R.M. Groves, et al. - Eds.). New York: John Wiley and Sons, 247-256.
SHANKS, J.M. (1983). The current status of computer assisted telephone interviewing: recent progress and future prospects. Sociological Methods and Research, 12, 119-142.
SHANKS, J.M., NICHOLLS, W.L., II and FREEMAN, H.E. (1981). The California Disability Survey: design and execution of a computer-assisted telephone study. Sociological Methods and Research, 10, 123-140.
SHAPIRO, G. M., BATTAGLIA, M. P., HOAGLIN, D. C., BUCKLEY, P., and MASSEY, J. T. (1996). Geographical variation in within-household coverage of households with telephones in an RDD survey. Proceedings of the Section on Survey Research Methods, 491-496.

SHURE, G.E., and MEEKER, R.J. (1978). A mini-computer system for multi-person computer-assisted telephone interviewing. Behavior Methods and Instrumentation. (April) 196-202.
SMITH, C., and FRAZIER, E. L. (1993). Comparison of traditional and modified Waksberg. Proceedings of the Section on Survey Research Methods, American Statistical Association, 926-931.
SQUIRE, P. (1988). Why the 1936 Literary digest poll failed. Public Opinion Quarterly, 52, 125-133.
STATISTICS CANADA (2000). Selected Dwelling Characteristics and Household Equipment. Income Statistics Division. [http://www.statcan.ca/english/Pgdb/People/Families/famil09b. htm ].
STOCK, J. S. (1962). How to improve samples based on telephone listings. Journal of Advertising Research, 2, 3, 50-51.
STOKES, L., and YEH, M.-Y. (1988). Searching for causes of interviewer effects in telephone surveys. Telephone Survey Methodology, (R.M. Groves, et al. - Eds.). New York: John Wiley and Sons, 357-376.
SUDMAN, S. (1966). New uses of telephone methods in survey research. Journal of Marketing Research, 3, 107-120.

SUDMAN, S. (1973). The uses of telephone directories for survey sampling. Journal of Marketing Research, 10, 204-207.
SUDMAN, S. (1978). Optimum cluster designs within a primary unit using combined telephone screening and face-to-face interviewing. Journal of the American Statistical Association, 73, 300-304.

SURVEY RESEARCH CENTER (2000). Sample Design for Household Telephone Surveys: A Bibliography 1949-1996. College Park, MD: University of Maryland. [http://www.bsos.umd.edu/ src/sampbib.html].
SURVEY SAMPLING INC. (1998). Random digit samples - part 1. [http://www.ssisamples.com/ssi.x20\$ssi_gen.search_item?id=1 19].
STATISTICS NETHERLANDS (1987). Automation in Survey Processing. Voorburg/Heerlen: Netherlands Central Bureau of Statistics (CBS Select 4).
SYKES, W.M., and COLLINS, M. (1987). Comparing telephone and face-to-face interviewing in the United Kingdom. Survey Methodology, 13,15-28.
SYKES, W.M., and COLLINS, M. (1988). Effects of mode of interview: experiments in the UK. Telephone Survey Methodology, (R.M. Groves, et al. - Eds.). New York: John Wiley and Sons, 301-320.

THORNBERRY, O.T. JR., and MASSEY, J.T. (1978). Correcting for undercoverage bias in random digit dialed National Health Surveys. Proceedings of the Section on Survey Research Methods, American Statistical Association, 224-229.
THORNBERRY, O.T. JR., and MASSEY, J.T. (1983). Coverage and response in random digit dialed national surveys. Proceedings of the Section on Survey Research Methods, American Statistical Association, 654-659.

THORNBERRY, O.T. JR., and MASSEY, J.T. (1988). Trends in United States telephone coverage across time and subgroups. Telephone Survey Methodology, (R.M. Groves, et al. - Eds.). New York: John Wiley and Sons, 25-49.

TORTORA, R.D. (1985). CATI in an agricultural statistics agency. Journal of Official Statistics, 1, 301-314.
TOURANGEAU, R., and SMITH, T.W. (1998). Collecting sensitive information with different modes of data collection. Computer Assisted Survey Information Collection, (M.P. Cooper, et al. Eds.). NewYork: John Wiley and Sons, 431-453.
TRAUGOTT, M.W., GROVES, R.M. and LEPKOWSKI, J.M. (1987). Using dual frame designs to reduce nonresponse in telephone surveys. Public Opinion Quarterly, 51, 522-539.
TREWIN, D., and LEE, G. (1988). International comparisons of telephone coverage. Telephone Survey Methodology, (R.M. Groves, et al. - Eds.). New York: John Wiley and Sons, 3-24.
TROLDAHL, V.C., and CARTER, R.E. (1964). Random selection of respondents within households in phone surveys. Journal of Marketing Research, 1, 71-76.
TUCKEL, P.S., and FEINBERG, B.M. (1991) The answering machine poses many questions for telephone survey researchers. Public Opinion Quarterly, 55, 200-217.
TUCKEL, P.S., and O'NEILL, H. (1996), New technology and nonresponse bias in RDD surveys. Proceedings of the Section on Survey Research Methods, American Statistical Association, 889-894.
TUCKEL, P., and SHUKERS, T. (1997). The effect of different introductions and answering machine messages on response rates. Proceedings of the Section on Survey Research Methods, American Statistical Association, 1047-1051.
TUCKER, C., CASADY, R. and LEPKOWSKI, J. (1992). Sample allocation for stratified telephone sample designs. Proceedings of the Section on Survey Research Methods, American Statistical Association, 291-296.
TURNER, C.F., FORSYTH, B.H., O'REILLY, J.M., COOLEY, P. C., SMITH, T.K., ROGERS, S.M., and MILLER, H.G. (1998). Automated self-interviewing and the survey measurement of sensitive behaviors. Computer Assisted Survey Information Collection, (M.P. Cooper, et al. - Eds.). NewYork: John Wiley and Sons, 455-473.
WAKSBERG, J. (1978). Sampling methods for random digit dialing. Journal of the American Statistical Association, 73, 40-46.
WAKSBERG, J. (1983). A note on 'Locating a special population using random digit dialing'. Public Opinion Quarterly, 47, 576-578.
WAKSBERG J. (1984). Efficiency of Alternative Methods of Establishing Cluster Sizes in RDD Sampling. Unpublished Westat memorandum.
WAKSBERG, J., BRICK, J.M., SHAPIRO, G., FLORES-CERVANTES, I. and BELL, B. (1997). Dual-frame RDD and area sample for household survey with particular focus on low-income population. Proceedings of the Section on Survey Research Methods, American Statistical Association, 713-718.
WERKING, G., TUPEK, A. R., and CLAYTON, R. L. (1988). CATI and touchtone self-response applications for establishment surveys. Journal of Official Statistics, 4, 349-362.
WHITE, A. A. (1983). Response rate calculation in RDD telephone health surveys: current practices. American Statistical Association, Proceedings of the Section on Survey Research Methods, 277-282.

Whitmore, R. W., MASON, R. E., and HARTWELL, T. D. (1985). Use of geographically classified telephone directory lists in multi-mode surveys. Joumal of the American Statistical Association, 80, 842-844.
WILSON, P., BLACKSHAW, N., and NORRIS P. (1988). An evaluation of telephone interviewing on the British Labour Force Survey. Journal of Official Statistics, 4, 385-400.
WINTER, D. L. S., and CLAYTON, R. L. (1990). Speech data entry: results of the first test of voice recognition for data collection. American Statistical Association, Proceedings of the Section on Survey Research Methods, 387-392.

WISEMAN, F. (1972). Methodological bias in public opinion surveys. Public Opinion Quarterly, 36, 105-108.
XU, M. BATES, B.J., and SCHWEITZER, J.C. (1993). The impact of messages on survey participation in answering machine households. Public Opinion Quarterly, 57, 232-237.

# Regression Composite Estimation for the Canadian Labour Force Survey with a Rotating Panel Design 

AVINASH C. SINGH, BRIAN KENNEDY and SHIYING WU ${ }^{1}$


#### Abstract

We consider the regression composite estimation introduced by Singh (1994, 1996; termed earlier as "modified regression composite" estimation), a version of which (suggested by Fuller 1999) has been implemented for the Canadian Labour Force Survey (CLFS) beginning in January 2000. The regression composite (rc) estimator enhances the generalized regression (gr) estimator used earlier for the CLFS and the well known Gurney-Daly $a k$-composite estimator in several ways. The main features of the rc-estimator are: (a) it considerably improves the efficiency of level and change estimates for key study variables resulting into less volatile estimate series; (b) it is calculated like the gr-estimator as a calibration estimator such that all the usual poststratification controls used in gr as well as the new controls corresponding to correlated variables from the previous time point are met; and (c) it respects the internal consistency of estimators without having to calculate part estimates differently as residuals. The main innovations used in re-class of estimators entail: (a) using the idea of working covariance matrix in estimating functions as an alternative to superpopulation modeling for defining regression coefficients for the predictors in the gr-estimator, (b) treating random controls (the ones based on the key correlated variables from past) as fixed, while computing the regression coefficients, similar to two-phase estimation, and motivated from the working covariance idea, and (c) that of the use of micro-matching to obtain previous time point's micro-level auxiliary information for realizing higher correlation with the present time point's study variables. As a by product, a new version of the akestimator which uses the micro-matching based predictors from past rather than the traditional macro-level is recommended in the interest of higher efficiency gains. The paper also presents an interesting heuristic justification of the smoothness feature of composite estimates using the amortization idea. Empirical results based on the Ontario 1996 CLFS data are presented for comparison of various estimators.


KEY WORDS: Generalized regression; Modified regression; Estimating functions; Regression calibration.

## 1. INTRODUCTION

In the case of repeated surveys with partially overlapping samples, it is well known ( see, e.g., Cochran 1977, Ch. 12) that estimates of level at a point in time and change between two time points can be improved by regressing the usual cross-sectional estimator (typically regression or simply Horvitz-Thompson) on the new predictors provided by the correlated observations on the overlapping subsample from the previous time point. Such methods of estimation belong to the class of composite estimation, and a simple version of which known as the $k$-composite estimator was proposed some time ago by Hansen, Hurwitz and Madow (1953), and examined further by Rao and Graham (1964), Binder and Hidiroglou (1988) provide an excellent review of the literature on estimation with repeated surveys. Note that there is an associated loss of efficiency in estimates aggregated over several time points due to increased positive correlation between composite estimates of successive time points. This is, however, probably a small price to pay because it is not the aggregate, but the level and change estimates that need more precision. The akcomposite estimator of Gurney and Daly (1965) provides an improved version of the $k$-composite estimator by reducing the variance further, an alternative simpler justification of which was provided by Wolter (1979).

The composite estimator considered in this paper was developed in the context of the Canadian Labour Force Survey (CLFS). The CLFS is a monthly survey that follows a rotating panel design with six panels. In any two consecutive months, five sixth of the households form the overlapping sample. It was in January 2000 that the CLFS started using a version (suggested by Fuller 1999) of the composite estimators introduced by Singh $(1994,1996)$ termed originally as "modified regression composite" estimators, which will be referred to in this paper as simply "regression composite" or rc-estimators. Before January 2000 , CLFS used the generalized regression (gr) estimators of Cassel, Särndal, and Wretman (1976) and Särndal (1980) which were based on only cross-sectional (i.e., present month's) data. It has long been felt that the estimator for CLFS could be improved using the composite estimation idea in the sense that estimates of level and change would be more efficient, and hence the resulting series would be more stable, i.e., less volatile. There are four goals that the rc-estimator attempts to meet in modifying the gr-estimator:
(i) It should considerably increase the efficiency of level and change estimates so that the estimate series becomes smoother or less volatile.
(ii) It can be computed as a calibration estimator like the gr -estimator so that the existing estimation software system can be used with little modification,

[^1](iii) The final calibrated weights should continue to satisfy the usual demographic and geographic controls used in the gr-estimator in addition to some new controls based on past month's variables, and
(iv) The estimator should have the internal consistency property in that the part composite estimators add up to the whole, e.g., estimates for Employed ( E ), Unemployed (U), and Not in the Labour Force (N) should add up to the total eligible population in the domain of interest.
The $a k$-estimator was studied by Kumar and Lee (1983) in the context of CLFS, and it was found that it didn't give substantial gains in efficiency as required by goal (i). The goal (iv) was, of course, known to be not satisfied by the $a k$-estimator because the (optimal) coefficients $a$ and $k$ used for combining several present month's estimators (in fact three of them, one is the usual estimator based on the present month, and the other two are built on predictors from the past month) tum out to be specific to the characteristic such as E. A solution (although rather undesirable) is to designate one of the components as least important (say, N ) and then obtain its estimate as a residual. The goals (ii) and (iii) can, however, be met by the $a k$-composite weighting suggested by Fuller (1990), and studied for the US Current Population Survey context by Lent, Miller, and Cantwell (1994, 1996). The goal (ii) is important especially for unplanned study variables for which the coefficients ( $a, k$ ) are not known in advance. The rc-estimator meets all the four goals, in particular the goals (i) and (iv), by making use of the following three innovations:
(i) . The design-based estimation in the presence of correlated predictors can be cast in an estimating functions framework as defined by Godambe and Thompson (1989), and then use the idea of working covariance matrix as in Liang and Zeger (1986) to obtain an alternative to the superpopulation modelling to compute regression coefficients. The resulting regression estimates, like gr, are only suboptimal under the design randomization.
(ii) The previous month's full sample composite estimates used as regression controls for present month's estimation can be treated as fixed using the working covariance idea for computational simplicity without violating the design consistency property. For variance estimation, the extra variation due to random controls should, of course, be accounted for.
(iii) Using micro-matching of the present month's overlapping subsample with the previous month, information about key study variables from the previous month is augmented to the present month's data. These now serve as additional covariates deemed to be highly correlated with the present month's study variable.

These innovations allow for computation of all estimates using the gr-system, thus avoiding the need of having to compute parts of estimates as residuals in the interest of internal consistency. The feature of micro-matching gives rise to desired gains in efficiency. In practice, it would often be the case that some of the present month's respondents in the overlapping sample were nonrespondents in the previous month, and so imputation might be necessary. In the case of CLFS, this is a small fraction, and the Hot Deck method with donor classes defined by demographic, geographic (subprovincial economic regions), type of area (rural/urban), present month's employment status, and industry group is used to fill in the missing values. It may be noted that sometimes imputation may be necessary not due to nonresponse at the previous time point, but due to the household's move. Assuming that on the average, households that move in the dwellings sampled at the present time $t$ are similar to the households that move out at $t$, then even though movers may have different employment characteristics than nonmovers, the imputation for movers is not expected to introduce any new bias as current month's employment status among other covariates is taken into account.

In the concluding section 6 , a method is suggested to diagnose the impact of this imputation. This impact may be serious for surveys with high fraction of previous month's missing values for the present month's respondents in the overlapping subsample. A possibly simple way out would be to redesign the questionnaire so that the interviewer is prompted by the instrument CATI software (computer assisted telephone interviewing commonly used now-adays) while administering the interview in second or later months, whether the respondent was nonrespondent at the previous month. If so, then the interviewer administers a rather short supplementary questionnaire in order to elicit the respondent's employment status for the previous month. This idea is similar to the method suggested by Hansen-Hurwitz-Madow for completely nonoverlapping repeated surveys, but each respondent is asked questions for the present as well as the previous time point, see Cochran (1977, page 355).

The organization of this paper is as follows. Section 2 presents a heuristic motivation using the amortization idea of why composite estimation, in general, is expected to provide desired smoothing of the estimate series. Section 3 defines various estimators, and discusses their computation via the gr-system. A new version of the $a k$-estimator, denoted by $a k^{*}$, is also proposed. The estimator uses predictors from previous month based on micro-matching, and is expected to give high gains in efficiency. Section 4 considers variance estimation by the currently used method of jackknife. An empirical comparison of the estimators is presented in section 5 using the Ontario 1996 CLFS data. Finally section 6 contains concluding remarks.

## 2. SERIES SMOOTHING BY COMPOSITE ESTIMATION: HEURISTICS

In this section, we present an interesting heuristic justification (based on the amortization idea rather than the shrinkage) of why smoothing of the estimate series is expected by composite estimation. (Using only the shrinkage idea, the series can be smoothed but it may not cross the original series often enough. With amortization, however, the left-over part after shrinkage is accounted for gradually over time, thus allowing for the smoothed series to cross the original one more often.) Consider the panel rotation scheme similar to that of the CLFS and let $\gamma$ denote the fraction of the panels rotated out; in the case of CLFS, $\gamma$ is $1 / 6$. Denote the cross-sectional estimator (typically gr) at time $t$ based on all panels, i.e., the full sample, by $F_{t}$, the estimator based on only the birth (i.e., rotate-in) panel by $B_{i}$, and the one based on nonbirth panels (i.e., the subsample at $t$ overlapping with the past sample at $t-1$ ) be $\bar{B}_{i}$. Similarly, denote the estimator based only on the death (i.e., rotate-out) panel by $D_{r}$, and the one based on nondeath panels (i.e., the subsample at $t-1$ overlapping with the present sample at $t$ ) be $\bar{D}_{t}$. We have

$$
\begin{gather*}
F_{t}=\gamma B_{t}+(1-\gamma) \bar{B}_{t}  \tag{2.1a}\\
F_{t-1}=\gamma D_{t-1}+(1-\gamma) \bar{D}_{t-1} . \tag{2.1b}
\end{gather*}
$$

Suppose, the series $\left\{F_{t}\right\}$ is too volatile, and we wish to smooth it. In the following it is assumed that there is no rotation group bias (Bailar 1975), i.e., different rotation groups have the same expected value. Thus $F_{i}$ is unbiased but may be unstable. This set-up is the traditional one for composite estimation in which different unbiased estimates are combined optimally to get a more efficient estimate. However, see the discussion at the end of this section for an alternative perspective on composite estimation in the presence of rotation group bias. Now denote the smoothed series by $\left\{C_{t}\right\}$, and consider the identity:

$$
\begin{equation*}
F_{t}=C_{t-1}+\left(F_{t}-F_{t-1}\right)+\left(F_{t-1}-C_{t-1}\right) . \tag{2.2}
\end{equation*}
$$

The above relation can be interpreted as follows. The estimate $C_{t-1}$ at $t-1$ is adjusted by the fluctuation ( $F_{\mathrm{t}}-F_{t-1}$ ) at the next time point $t$ in the $F$-series, and the existing gap ( $F_{t-1}-C_{t-1}$ ) at the time point $t-1$. If we define $C_{t}$ after full adjustments for these two differences, then $C_{t}$ would be the same as $F_{\mathrm{t}}$ and there would be no smoothing of the $F$-series. This suggests that the adjustments for the differences $\left(F_{t}-F_{t-1}\right)$ and ( $F_{t-1}-C_{t-1}$ ) should be accounted for only partially as $C$-series moves from $C_{t-1}$ to $C_{t}$. The remaining portions of the differences should be amortized gradually over future time points. All these adjustments should be done without affecting unbiasedness of the estimator $C_{f}$. The difference ( $F_{t-1}-C_{t-1}$ ) is zero in expectation assuming unbiasedness of $C_{t-1}$ and $F_{t-1}$ (which is so under the assumption of no rotation group bias) and therefore amortizing parts of it
would not affect unbiasedness of future estimates $C_{1}$. However, the difference $F_{t}-F_{t-1}$ is not zero in expectation, and care should be exercised in amortizing part of this difference. Observe that

$$
\begin{equation*}
F_{t}-F_{t-1}=\left(\bar{B}_{t}-\bar{D}_{t-1}\right)+\gamma\left(B_{t}-\bar{B}_{t}\right)+\gamma\left(\bar{D}_{t-1}-D_{t-1}\right) . \tag{2.3}
\end{equation*}
$$

The first term on the RHS is the change estimate based on common panels, while the second and third terms represent birth and death effects at $t$ and $t-1$ respectively. The last two terms are zero functions (i.e., are zero in expectation) but the first one is not. (Fortunately, the first term is expected to be stable as it is a difference of two highly correlated estimates.) Therefore, it is the second and third terms that should be amortized. Now, write (2.2) as

$$
\begin{align*}
F_{t}=C_{t-1} & +\left(\bar{B}_{t}-\bar{D}_{t-1}\right)+\gamma\left(B_{t}-\bar{B}_{t}\right) \\
& +\left[\gamma\left(\bar{D}_{t-1}-D_{t-1}\right)+\left(F_{t-1}-C_{t-1}\right)\right] \\
=C_{t-1} & +\left(\bar{B}_{t}-\bar{D}_{t-1}\right)+\gamma\left(B_{t}-\bar{B}_{t}\right) \\
& +\left[\left(\bar{D}_{t-1}-F_{t-1}\right)+\left(F_{t-1}-C_{t-1}\right)\right] \\
=C_{t-1} & +\left(\bar{B}_{t}-\bar{D}_{t-1}\right)+\gamma\left(B_{t}-\bar{B}_{t}\right)+\left(\bar{D}_{t-1}-C_{t-1}\right) . \tag{2.4}
\end{align*}
$$

and define two amortization factors $\delta_{1}, \delta_{2 t}$, between 0 and 1 , and then define the smoothed series $\left\{C_{t}\right\}$ as

$$
\begin{equation*}
C_{t}=C_{t-1}+\left(\bar{B}_{t}-\bar{D}_{t-1}\right)+\delta_{1 t} \gamma\left(B_{t}-\bar{B}_{t}\right)+\delta_{2 t}\left(\bar{D}_{t-1}-C_{t-1}\right) . \tag{2.5}
\end{equation*}
$$

The term with $\delta_{1 t}$ in (2.5) represents shrinkage of the birth effect at $t$ which $C_{t}$ tries to account for, while the term with $\delta_{2 t}$ refers approximately to shrinkage of the death effect at the past time $(t-1)$ which $C_{t}$ tries to make up for the present time $t$. Also, it would be desirable to set $\delta_{2 t}<\delta_{1 t}$ in order for the series $\left\{C_{t}\right\}$ to track $\left\{F_{t}\right\}$ better so that they have similar trend over time, i.e., give more importance to the current birth effect than the past death effect. (In fact, a rigorous justification under fairly general conditions of why one should set $\delta_{2 t} \leqslant \delta_{1 t}$ comes from optimality considerations in which variance of $C_{t}$ is minimized to obtain the best linear combination of three unbiased estimators, $F_{1}$, $C_{t-1}+\bar{B}_{t}-\bar{D}_{t-1}$, and $F_{t}+C_{t-1}-\bar{D}_{t-1}$ of the present month's population total; see (2.8) at the end of this section for the actual expression.) Now, to see the connection with the well known composite estimates defined in the next section, define $0<a_{t}, b_{t}<1$, so that $\delta_{1 t}=1-b_{t}, \delta_{2 t}=1-b_{t}-a_{t}$. We have

$$
\begin{align*}
C_{t}=C_{t-1} & +\left(\bar{B}_{t}-\bar{D}_{t-1}\right)+\left(1-b_{t}\right) \gamma\left(B_{t}-\bar{B}_{t}\right) \\
& +\left(1-b_{t}-a_{t}\right)\left(\bar{D}_{t-1}-C_{t-1}\right) . \tag{2.6}
\end{align*}
$$

It is interesting to note that if $b_{t}=0$, there would be no dampening of the birth effect, and the $C$-series is expected to be closer to $F$-series, i.e., there is less smoothing and the two would cross each other more often. If $a_{t}=0$, the past
effect represented by ( $\bar{D}_{t-1}-C_{t-1}$ ) is dampened less. This would imply more smoothing of the $F$-series, and the two series are expected to cross each other less frequently. Finally, if $a_{t}, b_{t}>0$, then the behaviour of the $C$-series relative to the $F$-series would be somewhere in the middle. Moreover, if $b_{t}$ is high (close to 1 ), there would be quite a bit of smoothing of the $F$-series because there is high amortization of both the birth and death effects. In these situations, one would expect sustained gaps between $F$ and $C$ series over time before they cross each other. Notice that parts of the term $\gamma\left(B_{t}-\bar{B}_{t}\right)$ that get amortized over $t, t+1, \ldots$ decrease as $t$ increases. They are given by $b_{t} \gamma\left(B_{t}-\bar{B}_{t}\right),\left(b_{t+1}+a_{t+1}\right) b_{t} \gamma\left(B_{t}-\bar{B}_{t}\right), \ldots$. Similarly, the amortized parts of ( $\bar{D}_{t-1}-C_{t-1}$ ) are

$$
\left(b_{t}+a_{t}\right)\left(\bar{D}_{t-1}-C_{t-1}\right),\left(b_{t+1}+a_{t+1}\right)\left(b_{t}+a_{t}\right)\left(\bar{D}_{t-1}-C_{t-1}\right), \ldots
$$

Clearly, when $b_{\mathrm{t}}$ is large, it will take several time points for completing the amortization. However, as explained earlier, this would not introduce bias because the effects being amortized are zero functions under the assumption of no rotation group bias.

The expression (2.6) can be cast into a more familiar expression of the composite estimator as follows:

$$
\begin{align*}
& C_{t}=C_{t-1}+\left(\bar{B}_{t}-\bar{D}_{t-1}\right)+\left(1-b_{t}\right)\left(F_{t}-\bar{B}_{t}\right) \\
&+\left(1-b_{t}\right)\left(\bar{D}_{t-1}-C_{t-1}\right)+a_{t}\left(C_{t-1}-\bar{D}_{t-1}\right)  \tag{2.7a}\\
&= C_{t-1}+\left(\bar{B}_{t}-\bar{D}_{t-1}\right)+\left(1-b_{t}\right)\left(F_{t}-\bar{B}_{t}+\bar{D}_{t-1}-C_{t-1}\right) \\
&+a_{t}\left(C_{t-1}-\bar{D}_{t-1}\right)  \tag{2.7b}\\
&= F_{t}+b_{t}\left[C_{t-1}-\left(F_{t}+\bar{D}_{t-1}-\bar{B}_{t}\right)\right]+a_{t}\left(C_{t-1}-\bar{D}_{t-1}\right)  \tag{2.7c}\\
&= F_{t}+\left(b_{t}+a_{t}\right)\left(C_{t-1}-\bar{D}_{t-1}+\bar{B}_{t}-F_{t}\right)+a_{t}\left(F_{t}-\bar{B}_{t}\right) \tag{2.7d}
\end{align*}
$$

The expression ( 2.7 d ) coincides with the $a k$-estimator (see next section) when $a_{t}=a$ and $b_{t}+a_{t}=k$. In practice, the values of $a_{t}$ and $b_{t}$ can be determined optimally or suboptimally using regression (see next section). The partial regression coefficients $a_{t}, b_{t}$ satisfy $0<a_{t}<b_{t}<1$ in general, because the direct estimator $F_{t}$ is expected to be more positively correlated with the predictor $F_{t}+\left(\bar{D}_{1-1}-\bar{B}_{t}\right)$, i.e, $\bar{D}_{t-1}+\gamma\left(B_{t}-\bar{B}_{t}\right)$ than with the predictor $\bar{D}_{t-1}^{t}$; both predictors being unbiased estimates, like $C_{t-1}$, of the population total parameter at the previous time point $t-1$. It follows from (2.7c) that the estimator $C_{t}$ can be written as a linear combination of the three unbiased estimators mentioned earlier, and is given by

$$
\begin{gather*}
C_{t}=\left(1-b_{t}-a_{t}\right) F_{t}+b_{t}\left(C_{t-1}+\bar{B}_{t}-\bar{D}_{t-1}\right) \\
+a_{t}\left(F_{t}+C_{t-1}-\bar{D}_{t-1}\right) . \tag{2.8}
\end{gather*}
$$

The above heuristic motivation corresponds to the variance reduction considerations under the assumption of no rotation group bias when combining three unbiased estimators of the population total at $t$. In the presence of rotation group bias, however, all the three estimators become biased with possibly different magnitude and direction, and what composite estimation does is to adjust each one of them so that the adjusted value for each is equal to a common value given by the composite estimator. (For example, in the case of two estimators $\hat{\theta}_{1}$ and $\hat{\theta}_{2}$ of $\theta$, the linear combination $\lambda \hat{\theta}_{1}+(1-\lambda) \hat{\theta}_{2}$ can be written as $\hat{\theta}_{2}+\lambda\left(\hat{\theta}_{1}-\hat{\theta}_{2}\right)$ or $\hat{\theta}_{1}+(1-\lambda)\left(\hat{\theta}_{2}-\hat{\theta}_{1}\right)$ implying that the two original estimators are adjusted appropriately to converge to a common value.) The relative weight in combining the three estimators depends on the criterion of minimum variance. Ideally, it should be based on the minimum MSE criterion, but it is hard to get a handle on bias because it can't be estimated. Clearly the composite estimator is not bias free, and it can only be speculated that the overall bias of the estimator is reduced by compositing. Similarly if, instead, a suboptimal regression is used in constructing the composite estimator (as in rc-estimation, see the next section), then what composite estimation does is to adjust the sampling weights in the full sample (which are generally gr-weights) so that $F_{t}-\left(\bar{B}_{t}-\bar{D}_{t-1}\right)$, and $\bar{D}_{t-1}$ with adjusted weights become equal to $C_{t-1}$; the $C_{t-1}$ serve as new controls in the calibration step. This is another way of adjusting the three estimators to a common value, but again bias of the resulting composite estimator remains unknown. The above discussion of two perspectives on composite estimation has some similarity with the dual property of poststratification in terms of both variance and (coverage) bias reduction, see Singh and Folsom (2000).

## 3. COMPOSITE ESTIMATORS: NEW AND OLD

We start with the cross-sectional estimator at time $t$ of the total $\tau_{y}(t)$ defined as gr , which is given by

$$
\begin{equation*}
\hat{\tau}_{y(g r)}(t)=\sum_{k \in s(t)} y_{k}(t) w_{g r}(t, k), \tag{3.1}
\end{equation*}
$$

$$
\begin{align*}
& w_{\mathrm{gr}}(t, k) \\
& \quad=d(t, k)\left[1+x_{k}(t)^{\prime}\left(X(t)^{\prime} \Delta(t) X(t)\right)^{-1}\left(\tau_{x}(t)-\hat{\tau}_{x}(t)\right)\right], \tag{3.2}
\end{align*}
$$

where $d(t, k)$ 's are the initial design weights adjusted for nonresponse, $x_{k}(t)$ is a $p$-vector of covariates used for calibration (or poststratification), $X(t)$ is the $n(t) x p$ matrix of $x$-observations, $n(t)$ is the sample size, $\Delta(t)$ is diag $(d(t, k)), \tau_{x}(t)$ is the known $p$-vector of calibration controls, and $\hat{\tau}_{x}(t)$ is the corresponding vector of expansion estimates based on $d$-weights. In terms of the notation $F_{s}, \bar{B}_{t}$, and $B_{t}$ of the previous section, $F_{t}$ here can be taken as the gr-estimator (3.1), and $\bar{B}_{\mathrm{r}}$ is gr-estimator based on nonbirth panels given by

$$
\begin{equation*}
\bar{B}_{t}=(1-\gamma)^{-1} \sum_{k \in s(\mid t-1)} y_{k}(t) w_{\mathrm{gr}}(t, k), \tag{3.3}
\end{equation*}
$$

where $s(t \mid t-1)$ is the subsample at $t$ matched with the sample at $t-1$. The estimator $B_{t}$ is also a gr-estimator, and is given by

$$
\begin{equation*}
B_{t}=\gamma^{-1} \sum_{k \in s(t)-s(t \mid t-1)} y_{k}(t) w_{\mathrm{gr}}(t, k), \tag{3.4}
\end{equation*}
$$

where the sum is over the subsample defined by the birth panel at $t$.

The $a k$-composite estimator uses the macro-level past information for the new predictors, and can be defined as

$$
\begin{align*}
& C_{t(a k)}=F_{t}+k\left(C_{t-1(a k)}-\bar{D}_{t-1}+\bar{B}_{t}-F_{t}\right)+a\left(F_{t}-\bar{B}_{t}\right) \\
= & F_{t}+(k-a)\left(C_{t-1(a k)}-\bar{D}_{t-1}+\bar{B}_{t}-F_{t}\right)+a\left(C_{t-1(a k)}-\bar{D}_{t-1}\right) . \tag{3.5}
\end{align*}
$$

Here the coefficients $a, k$ for level estimation are obtained by optimally regressing $F_{t}$ on the two predictor zero functions, based on the past information, namely, $C_{t-1(a k)}-\left(F_{t}+\bar{D}_{t-1}-\bar{B}_{t}\right)$, and $C_{t-1(a k)}-\bar{D}_{t-1}$. Thus, $a, k$ depend on the sample design as well as on the study variable $y$, in particular, they are not even the same for level and change estimates for the same $y$. For change estimation, $F_{t}-C_{t-1(a k)}$, and not $F_{t}$ is regressed optimally on the above predictors. In practice, $a, k$ are estimated by performing a grid search on the interval $(0,1)$ such that the variance of $C_{t}$ is minimized. As mentioned earlier, typically $a$ is smaller than $k$. In defining the above two new predictor zero functions based on past information, two estimators of $\tau_{y}(t-1)$ are first formed: one is $\bar{D}_{t-1}$ based on the nondeath panels at $t-1$ (i.e., subsample at $t-1$ matched with the sample at $t$ ), and the other is $F_{t}+\left(\bar{D}_{t-1}-\bar{B}_{t}\right)$ which is the grestimator at time $t$ adjusted for change from $t-1$ to $t$ estimated from the common sample. Clearly, if there is no overlap in the panel design, then all the predictor zero functions become no longer meaningful resulting in no change in $F_{t}$ by composite estimation. Similarly, if there is a complete overlap, then $\bar{B}_{t}=F_{t}$, and again there is no effect on $F_{f}$ of composite estimation. This may at first seem counter-intuitive, because the past data $\left(y_{t-1}\right)$ is correlated with the present $\left(y_{t}\right)$ due to sample overlap. However, complete overlap amounts, in principle, to collecting a single sample of multivariate data on $y$ with elements corresponding to $y$ at different time points. Using this analogy, there is no room for improvement (in the designbased framework) as there is no larger sample with additional information. In the case of no overlap, additional information is there but it doesn't help as it is uncorrelated. Note, however, that at the first stage, psu's ( primary sampling units) in CLFS remain common over several years before they are rotated out. Therefore, efficiency gains due to partial overlap are realized mainly from the reduction of the second stage variance component.

Furthermore, note that the estimator $C_{\text {rak) }}$ uses past information in the univariate sense in that for the study variable $y$, past information about only $y_{t-1}$ is used. If new predictors based on several variables such as $y_{t-1}, z_{t-1}, \ldots$ from the past are also used for the study variable $y$, then the composite estimation becomes multivariate. However, the optimal choice of the ( $a, k$ ) coefficients for the multivariate case can be quite cumbersome.

The rc-class of estimators is given by

$$
\begin{align*}
C_{t(\mathrm{rc})}=F_{\mathrm{t}} & +\dot{b}_{t(\mathrm{rc})}\left(\tilde{C}_{t-1(\mathrm{rc})}-\bar{D}_{t-1}^{*}+\bar{B}_{t}-F_{t}\right) \\
& +a_{t(\mathrm{rc})}\left(\tilde{C}_{t-1(\mathrm{rc})}-\bar{D}_{t-1}^{*}\right) \tag{3.6}
\end{align*}
$$

where $\tilde{C}_{t-1(\mathrm{rc})}$ denotes the $t-1$ estimator for the study variable ( $y$ ) after the ( $t-1$ )-calibration weights are further calibrated to meet the controls used for poststratification by gr at time $t$. Thus $\tilde{C}_{t-1(\mathrm{rc})}$ is an estimate of the population total at $t$ for the $y$-variable at $t-1$. The starred $\bar{D}_{t-1}^{*}$ signifies that it is based on the subsample at $t$ matched with the sample at $t-1$, but uses the gr-weights at $t$ as the $y$ values from $t-1$ are augmented to the sample at $t$ by micromatching. (Note that the estimator $\bar{D}_{t-1}^{*}$ involves, in general, imputed values, and may suffer from bias due to imputation. For a diagnosis and adjustment for this bias, see section 6.) The coefficients $b_{t(\mathrm{rc})}$ and $a_{t(\mathrm{rc})}$ are computed similar to gr of (3.1); see below for more details. These coefficients are suboptimal unlike ( $a, k$ ). However, like ( $a, k$ ), they are $y$-specific, and in the case of multivariate they depend on the key set of study variables chosen from past for new controls, but they can be computed easily as they are suboptimal in nature. Thus with rc-estimation, it is fairly easy to introduce more predictors. The predictors $\left(C_{t-1}-\bar{D}_{t-1}\right)$ and ( $C_{t-1}-\bar{D}_{t-1}+\bar{B}_{t}-F_{t}$ ) can be termed respectively as level-driven and change-driven as in Singh, Kennedy, Wu and Brisebois (1997). The reason for this is that not only the former is a difference of two level estimates, and the latter a difference of two change estimates, ( $C_{t-1}-F_{t}$ ) and ( $\bar{D}_{t-1}-\bar{B}_{t}$ ), but that the former tends to provide high efficiency gains in level estimation over what can be obtained in the presence of the latter, and similarly, the latter provides high efficiency gains in change estimation over what can be achieved in the presence of the former.

The idea of using the micro-level past information for the new predictors in rc-estimation can be applied to the $a k$ estimator, and thus a new estimator $a k^{*}$ can be proposed.

$$
\begin{align*}
C_{t\left(a k^{*}\right)}=F_{t} & +\left(k^{*}-a^{*}\right)\left(\tilde{C}_{t-1\left(a k^{*}\right)}-\bar{D}_{t-1}^{*}+\bar{B}_{t}-F_{t}\right) \\
& +a^{*}\left(\tilde{C}_{t-1\left(a k^{*}\right)}-\bar{D}_{t-1}^{*}\right) . \tag{3.7}
\end{align*}
$$

The control $\left.\tilde{C}_{r-1(a k}{ }^{*}\right)$ denotes the $(t-1)$ calibration estimator for $y$ after the $a k^{*}$-composite weights are further
calibrated to meet the controls used for poststratification by gr at $t$. (Here the $a k^{*}$-composite weights are similar to the $a k$-composite weights of Fuller (1990) where the composite estimators for a set of key $y$-variables serve as additional controls in the usual gr to obtain a set of final calibration weights. This allows for the $a k$-composite estimator to be computed as a calibration estimator.) The main differences between the various estimators defined above lie in the definition of regression coefficients (optimal vs. suboptimal), and that of the predictors (macro-level vs. micro-level use of past information). Special cases of the above composite estimators can be obtained as described in Singh, et al. (1997) by using only one of the two predictors. For $C_{t(a k)}$, if $a=0$ (i.e., only change-driven predictor is used), we get the well known $k$-composite estimator which can be termed as the $a k 2$-estimator in the present context. If $a=k$, i.e., only level-driven predictor is used, we get a new composite estimator $C_{\text {t(ak1) }}$ which can be termed as the $a k 1$-estimator. Similarly for $C_{t(a k *)}$, we get two more new composite estimators $a k^{*} 1$ and $a k^{*}$. For $C_{\text {t(rc) }}$, with only leveldriven predictor, we get the rc1-estimator, termed earlier as MR1 in Singh and Merkouris (1995). With only changedriven predictors, we get the rc2-estimator termed earlier as MR2 in Singh, et al. (1997).

As mentioned earlier, the rc-estimator is computed as a gr-estimator of (3.1), and therefore, it can be expressed as $\hat{\tau}_{y(\mathrm{rc})}(t)=\sum_{k \in s(t)} y_{k}(t) w_{\mathrm{rc}}(t, k)$. The $X(t)$-matrix is expanded to $n(t) \times(p+2 q)$ matrix $X(t)^{*}$ where $2 q$ represents the number of new predictors, the factor 2 signifying the pair of level-driven and change-driven predictors. The (random) control totals $\tilde{C}_{t-1(\mathrm{rc})}$ corresponding to the key set of $y$-variables from $t-1$ selected for composite estimation are treated as fixed (during the computation of regression coefficients) like the other (nonrandom) grcontrols. Now, since the level-driven predictor can be written as

$$
\begin{align*}
\bar{D}_{t-1}^{*} & =(1-\gamma)^{-1} \sum_{k \in s(t k-1)} y_{k}(t-1) w_{\mathrm{gr} t}(t, k) \\
& =\sum_{k \in s(t)}(1-\gamma)^{-1} y_{k}(t-1) 1_{k \in s(t t-1)} w_{\mathrm{gr}}(t, k) \tag{3.8}
\end{align*}
$$

the column of the $X(t)^{*}$-matrix corresponding to this predictor consists of $n(t)$-values of $(1-\gamma)^{-1} y_{k}(t-1) 1_{k \in s(n t-1)}$. Similarly the change-driven predictor can be written as

$$
\begin{align*}
& F_{t}+\bar{D}_{t-1}^{*}-\bar{B}_{t}= \\
& \sum_{k \in s(t)}\left(y_{k}(t)+(1-\gamma)^{-1}\left(y_{k}(t-1)-y_{k}(t)\right) 1_{k \in s(t t-1)}\right) w_{\mathrm{gr}}(t, k) \tag{3.9}
\end{align*}
$$

and the corresponding column of the $X(t)^{*}$ matrix consists of the $n(t)$-values of $y_{k}(t)+(1-\gamma)^{-1}\left(y_{k}(t-1)-y_{k}(t)\right) 1_{k \in s(l k-1)}$. Once the $X(t)^{*}$ matrix is defined, the gr-system can be used to compute the calibration weights $w_{\mathrm{rc}}(t, k)$ as in (3.2). Note that the calibration weights $w_{\mathrm{rc}}(t, k)$ can be used for estimation of all study variables although they depend explicitly only on the key set of study variables chosen for the new predictors from correlated past information. Also
note that although the rc-estimator of (3.6) was defined as the gr-estimator plus regression-adjustments for the new predictors, computationally it is convenient to perform a grcalibration on the design weights when all the old and new calibration controls are considered simultaneously. This way computation for the multivariate rc-estimator is not much different from the univariate rc-estimator. Alternatively, one could compute the rc-estimator as an adjusted gr as in (3.6), but the coefficients for the new predictors would be partial regression coefficients, and therefore do not have the standard form of the gr-coefficients.

Finally we note that with composite estimation, one would expect higher efficiency gains for change estimates $\left(C_{t}-C_{t-1}\right.$ vs. $F_{t}-F_{t-1}$ ) than those for level estimates ( $C_{i}$ vs. $F_{t}$ ). To see this, consider a simple identity: $V\left(F_{t}-F_{t-1}\right)=V\left(F_{t}\right)+V\left(F_{t-1}\right)-2 \operatorname{Cov}\left(F_{t}, F_{t-1}\right)$. Typically $V\left(F_{t}\right) \approx V\left(F_{t-1}\right)=\sigma_{\mathrm{gr}}^{2}($ say $)$, then the above can be reduced to $V\left(F_{t}-F_{t-1}\right) \approx 2 \sigma_{\mathrm{gt}}^{2}\left(1-\rho_{\mathrm{gr}}\right)$. Similarly, $V\left(C_{t}-C_{t-1}\right) \approx 2 \sigma_{r c}^{2}\left(1-\rho_{r c}\right)$. Thus the change efficiency is approximately the level efficiency times $\left(1-\rho_{\mathrm{gr}}\right) /\left(1-\rho_{\mathrm{rc}}\right)$. It follows that if the new predictors for composite estimation increase considerably the (positive) correlation between $C_{t}$ and $C_{t-1}$, then the change efficiency will highly dominate the level efficiency.

## 4. VARIANCE ESTIMATION

The CLFS currently uses delete-one psu jackknifing to find variance of the gr-estimate. The method of jackknifing is valid (for cross-sectional surveys) if the psu-level estimates have identical mean and variance, and the psu selection can be treated as with replacement. When psu selection is without replacement the variance estimate becomes conservative if the (common) covariance between the psu-level estimates is negative. This is generally the case. For repeated surveys, a third condition that psu's are common (or connected) over time is needed. When this is the case the survey can be viewed as cross-sectional by treating the vector of observations (psu-level estimates) over time as a single observation collected at the conceptual end point in time. In the rotating panel design of the CLFS, psu's are not rotated out for a number of years, but the within psu units are rotated every six months. Each psu in the CLFS corresponds to a single panel which is either birth or non-birth. Note that to meet the conditions of jackknifing, it is not necessary that the same set of units be used to obtain psu-level estimates. The condition that psulevel estimates have common mean and variance within a stratum is reasonable on the grounds that the panel estimates have common mean and variance. For composite estimation, although birth and non-birth panels are treated differently, panel-level composite estimates should have identical mean and variance unconditionally on the panel assignment. This is so because the panels are assigned at random; a panel could have been birth with probability
$\gamma=1 / 6$ and non-birth with probability $1-\gamma=5 / 6$. The resulting unconditional variance estimate will not be smaller than the one obtained conditionally on the panel assignment. Thus the method of jackknifing is expected to provide a conservative variance estimate in the CLFS context. Note that the above considerations for measures of uncertainty do not involve rotation group bias that may be present.

## 5. EVALUATION RESULTS

The numerical results are based on 1996 Ontario CLFS data, see Singh, et al. (1997). The auxiliary variables for gr are population counts corresponding to 16 age-sex groups, 11 economic regions, 10 census metropolitan areas, and 6 panels. Each panel control specifies $1 / 6$ of the $15+$ population. The new controls ( 30 in all) for rc corresponding to only change-driven predictors are: employed, unemployed and not in the labour force by age (young and old) by sex groups for a total of 12 , employment by industry categories for a total of 16 , and 2 employment by full/part time categories. In fact, these 30 new controls reduce to only 28 because of linear dependence. The multivariate rcestimator involves these 28 extra controls, while the univariate re involves just one extra control. The average relative efficiency shown in various tables is computed as the average variance of gr over 12 months of 1996 divided by the average variance of the composite estimator over 12 months.

### 5.1 Macro-level vs. Micro-level Predictors

For level-estimates, the correlation is computed between the current month level estimate (i.e., $F_{t}$ ) and the predictor (e.g., the level-driven $C_{t-1}-\bar{D}_{t-1}$ at the macro-level), whereas for the change estimate, it is computed between
$F_{t}-C_{t-1}$ and the predictors. The correlation is negative as expected because the estimate involving common panels is positively correlated with $F_{t}$ but expressed with a negative sign in the predictor. Recall that the composite estimator used is the $a k$ with macro-level and $a k^{*}$ with micro-level predictors.

It is seen from Table 1 for the four key variables (employed, unemployed, employed in Trade, and employed in Transportation and Communication (TRCO)), for each of the level-driven and change-driven predictors, microlevel predictors outperform macro-level in terms of high correlation.

Between level- and change-driven predictors at the micro-level, change-driven is seen to out-perform leveldriven. Similar results hold for other key variables. In view of these correlations, other evaluation results shown below pertain to only $a k 2, a k * 2$, and rc2 versions of composite estimates. The rc-estimator with both level- and changedriven predictors was not included in the interest of keeping down the number of extra controls.

## $5.2 a k \nu s . a k * v s$. rc (Efficiencies Relative to gr )

Table 2 shows the optimal coefficients (e.g., $k$ for $a k 2$ estimator) and the corresponding relative efficiency over gr . The optimal coefficients were found via grid-search using the same 1996 data. (In practice, this should be based on past data). It is seen that the efficiency gains can be considerable as one moves from $a k$ to $a k^{*}$. The optimal coefficients vary for level and change estimates. The last two columns under each of level and change estimates show the reduction in efficiency if level-optimal coefficients are used for change estimates and vice-versa. Level-optimal coefficients seem to perform quite well for change estimates, in contrast to a drop in efficiency of level estimates when change-optimal coefficients are used.

Table 1
Average Monthly Correlation between Composite Predictor and Estimates for Level and Change (Ontario, 1996)

| Variable | Level |  |  |  | Change |  |  |  |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
|  | Level-Driven Predictors |  | Change-Driven Predictors |  | Level-Driven Predictors |  | Change-Driven Predictors |  |
|  | Macro | Micro | Macro | Micro | Macro | Micro | Macro | Micro |
| Employed | -0.27 | -0.35 | -0.23 | -0.45 | -0.35 | -0.49 | -0.57 | -0.84 |
| Unemployed | -0.26 | -0.35 | -0.24 | -0.33 | -0.22 | -0.40 | -0.39 | -0.53 |
| Empl. Trade | -0.58 | -0.55 | -0.58 | -0.65 | -0.65 | -0.73 | -0.91 | -0.96 |
| Empl. TRC0 | -0.58 | -0.55 | -0.60 | -0.68 | -0.63 | -0.70 | -0.92 | -0.96 |

Table 2
Average Relative Efficiency of $a k$ and $a k^{*}$ over gr (Ontario, 1996)

| Variable | Coeff |  |  |  | Eff (Level) |  | Eff (Change) |  | Eff (Level) | Eff (Change) |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
|  | Level | timal | Chan | ptimal | Level | mal |  |  | Change Optimal | Level Optimal |
|  | $a k$ | $a k^{*}$ | $a k$ | $a k^{*}$ | $a k$ | $a k^{*}$ | ak | $a k^{*}$ | $a k^{*}$ | $a k^{*}$ |
| Employed | 0.42 | 0.72 | 0.48 | 0.95 | 1.05 | 1.25 | 1.28 | 2.43 | 0.72 | 2.21 |
| Unemployed | 0.40 | 0.50 | 0.54 | 0.69 | 1.06 | 1.12 | 1.11 | 1.29 | 1.05 | 1.26 |
| Empl. Trade | 0.79 | 0.84 | 0.95 | 0.98 | 1.43 | 1.67 | 2.36 | 4.97 | 0.88 | 4.22 |
| Empl. TRC0 | 0.84 | 0.87 | 0.95 | 0.98 | 1.59 | 1.88 | 3.60 | 7.59 | 1.11 | 6.51 |

Table 3 compares rc (univariate and multivariate) with $a k^{*}$. The possible values of $b_{t(\mathrm{rc} 2)}$ coefficients over the 12 month-period for the univariate rc2 are summarized via mean, minimum and maximum. They can be compared with the corresponding optimal coefficients for $a k^{*}$. The rccoefficients seem to provide a compromise and lie somewhere between level-optimal and change-optimal coefficient values. The rc-efficiencies for the change estimate are quite at par with those for $a k^{*}$ but for level estimates, are somewhat lower. The efficiency gains at the aggregate level for which gr had controls are low but are high for domains without gr-controls.

Table 4 presents possible loss in efficiencies for estimates obtained as residuals in $a k^{*}$-estimation in the interest of internal consistency. It shows that caution should be exercised in practice when choosing variables for residual estimation or using compromise coefficient values in $a k^{*}$-estimation of components of an aggregate.

### 5.3 Change $v s$. Level Efficiencies of rc Over gr

Table 5 shows that the approximate relation (see section 3) between change and level efficiencies holds fairly well. It is seen that month-to-month correlation for rc-estimates for domains not having a corresponding population control in gr can be quite high compared to the correlation for gr.

This, in tum, yields a high factor by which change efficiency exceeds level efficiency.

### 5.4 Point Estimate and SE of Difference Between rc and gr

Table 6 shows monthly estimates (and SE of level and change estimates) for the variable (employed in trade at the Ontario level) for gr and rc. The corresponding values for the monthly difference (rc-gr) are also shown. It is seen that the differences between rc and gr are not significant in general. Efficiencies (not shown here) of annual average and quarterly estimates of rc and gr were also computed. As expected, due to serial correlation, there may be a loss in efficiency over gr. However in terms of the coefficient of variation, this is likely to be of no practical consequence.

### 5.5 Time Series of Level Estimates

Figures 1(a) and (b) show level estimates of employment for Ontario for the period 88-96 for gr and re without and with seasonal adjustment. (The X11-ARIMA method was used.) Figures 2(a) and (b), show employment for the industry group "Trade". At the provincial level, aggregated over the industry group, there is similarity between gr and rc (seasonally adjusted or not) series because the grestimates have high precision to begin with. At the domain

Table 3
Average Relative Efficiency of rc over gr (Ontario, 1996)

| Coeff |  |  |  |  |  |  |  | Eff (Change) |  |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
|  | rc-univariate (level or change) |  |  | $a k^{*}$ |  | rc (univariate) | $\underset{\text { (multivariate) }}{\text { TC }}$ | $\begin{gathered} \text { rc } \\ \text { (univariate) } \end{gathered}$ | $\begin{gathered} \mathrm{rc} \\ \text { (multivariate) } \end{gathered}$ |
| Variable | Avg | Min | Max | Level | Change |  |  |  |  |
| Employed | 0.88 | 0.81 | 0.90 | 0.72 | 0.95 | 1.05 | 1.05 | 2.39 | 2.46 |
| Unemployed | 0.60 | 0.53 | 0.65 | 0.50 | 0.69 | 1.12 | 1.12 | 1.31 | 1.33 |
| Empl. Trade | 0.96 | 0.94 | 0.98 | 0.84 | 0.98 | 1.17 | 1.22 | 4.98 | 5.07 |
| Empl. TRC0 | 0.95 | 0.93 | 0.97 | 0.87 | 0.98 | 1.37 | 1.42 | 7.47 | 7.52 |

Table 4
Average Relative Efficiency of $a k^{*}$ and rc over gr from Ontario, 1996 (Regular vs. Residual)

| Variable |  | Level |  |  | Change |  |  |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
|  |  | $a k^{*}$ Coeff | Eff ( $a k^{*}$ ) | Eff (rc) | $a k^{*}$ Coeff | Eff ( $a k^{*}$ ) | Eff (rc) |
| Agriculture | (regular) | 0.91 | 2.55 | 2.32 | 0.97 | 4.88 | 5.22 |
| Agriculture | (residual) | NA | 0.63 | 2.32 | NA | 3.90 | 5.22 |
| NILF | (regular) | 0.74 | 1.26 | 1.07 | 0.95 | 1.96 | 2.01 |
| NILF | (residual) | NA | 1.21 | 1.07 | NA | 1.95 | 2.01 |

Table 5
Relation Between Level and Change Efficiencies for rc (multivariate) over gr (Ontario, 1996)

| Variable | Change Eff | Level Eff | Change Eff/Level Eff | $\left(1-\rho_{\mathrm{gr}}\right)\left(1-\rho_{\mathrm{rc}}\right)$ | $\rho_{\mathrm{gr}}$ | $\rho_{\mathrm{c}}$ |
| :--- | :---: | :---: | :---: | :---: | :---: | :---: |
| Employed | 2.46 | 1.05 | 2.34 | 2.65 | 0.77 | 0.91 |
| Unemployed | 1.33 | 1.12 | 1.19 | 1.21 | 0.50 | 0.59 |
| Empl Trade | 5.07 | 1.22 | 4.16 | 3.80 | 0.79 | 0.95 |
| Empl TRCO | 7.54 | 1.42 | 5.31 | 5.66 | 0.80 | 0.97 |

Table 6
Monthly Point Estimates for gr and rc and Their Differences (Ontario, 1996)
(Level and Change for Employment in Trade, Ontario, 1996)

| Month | Type | gr |  | rc |  | rc-gr |  |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| January | Level | 886.5 | (21.0) | 858.9 | (17.3) | -27.6 | (23.0) |
|  | Change | -25.8 | (13.2) | -21.0 | (5.6) | 4.8 | (11.4) |
| February | Level | 906.5 | (22.9) | 867.9 | (17.6) | 38.6 | (24.6) |
|  | Change | 20 | (14.2) | 9.0 | (4.7) | -11.0 | (12.5) |
| March | Level | 927.1 | (20.8) | 874.1 | (18.3) | -52.9 | (23.1) |
|  | Change | 20.6 | (13.3) | 6.2 | (4.7) | -14.4 | (12.5) |
| April | Level | 914.8 | (20.3) | 872.5 | (17.7) | -42.3 | (22.4) |
|  | Change | -12.3 | (13.4) | -1.6 | (5.1) | 10.7 | (12.5) |
| May | Level | 912.8 | (18.9) | 887.6 | (17.0) | -25.1 | (21.8) |
|  | Change | -2.1 | (13.0) | 15.1 | (5.7) | 17.2 | (11.6) |
| June | Level | 908.1 | (17.8) | 888.6 | (17.2) | -19.5 | (21.5) |
|  | Change | -4.7 | (12.3) | 0.9 | (4.9) | 5.6 | (11.9) |
| July | Level | 899.9 | (18.1) | 881.2 | (17.7) | -18.7 | (23.0) |
|  | Change | -8.2 | (12.8) | -7.4 | (6.7) | 0.8 | (10.7) |
| August | Level | 913.9 | (16.9) | 888.1 | (18.3) | -25.8 | (22.6) |
|  | Change | 14.0 | (11.5) | 6.9 | (5.3) | -7.1 | (10.3) |
| September | Level | 886.6 | (20.4) | 876.4 | (19.7) | -10.2 | (23.1) |
|  | Change | -27.3 | (12.6) | -11.8 | (6.3) | 15.6 | (11.1) |
| October | Level | 898.6 | (22.9) | 889.3 | (19.3) | 9.3 | (26.1) |
|  | Change | 12.1 | (13.4) | 12.9 | (6.6) | 0.9 | (11.8) |
| November | Level | 911.2 | (20.3) | 902.3 | (19.3) | -8.9 | (25.9) |
|  | Change | 12.6 | (13.9) | 13.0 | (7.0) | 0.4 | (12.6) |
| December | Level | 917.9 | (20.5) | 916.3 | (19.0) | -1.5 | (26.0) |
|  | Change | 6.7 | (12.5) | 14.0 | (6.1) | 7.4 | (10.9) |

Note: SEs are shown in parentheses.
level defined by Trade, however, the series are quite different. (Note that among numerous series that were examined, this particular series was chosen here to depict the extreme scenario for gaps between gr and re series. For almost all other series, the two series crossed each other fairly often.) Since the gr-series is highly volatile, there is room for considerable smoothing by rc. Also note that because of expected high signal-to-noise ratio, seasonally adjusted rc series at the Trade-domain level looks considerably smoother than that for the gr-series; in fact, there is very little difference between with and without seasonally adjusted gr-series. It is also observed that there tends to be runs of consecutive periods where rc is either larger or smaller than gr. This is expected because of high values of the $b_{t \mathrm{trc})}$ coefficients (Table 3), and high serial correlation in both series ( see Table 5). Interestingly, turning points in the gr and rc series tend to occur at (approximately) same time points though they appear somewhat dampened with rc due to higher serial correlation in rc-series. It may be noted that the gap between the two series would have been smaller if controls for level- driven predictors were also included.

## 6. CONCLUDING REMARKS

The previously used gr-estimator in CLFS showed instability in change estimates and various domain level estimates. The rc-estimator provides smoother estimate series (which, in tum, renders change estimates more stable). The rc-method departs from the traditional akcomposite estimation in several ways, the main points being the use of micro-matching for collection of unit-level past information for common panels, and the use of regression calibration (like gr) to produce a set of final weights for use with all study variables. Three versions of rc were examined. Although this paper was mainly concemed with rc2, i.e., with change-driven predictors (because of the desired resulting smoothness in estimate series), it was found (although not reported here) that level estimates of some key variables can be further improved (in comparison to rc2) by including corresponding level-driven predictors. Thus, in practice, a good strategy might be to use a mixture of mostly change-driven and some level-driven predictors.

The version of the rc-estimator currently implemented for CLFS was suggested by Fuller (1999), and can be expressed as

Figure 1(a) Enployment in Ontario, actual


Figure 2 (a) Employment in Trade, Ontaria, actual


Figure 1(b) Enployment, Ontario, seasonally adjusted


Figure 2(b) Enployment in Trade, Ortario, seasonallyacj.



$$
\begin{align*}
C_{t(\mathrm{rca})}=F_{t} & +b_{t(\mathrm{rca})}\left[(1-\alpha)\left(\tilde{C}_{t-1(\mathrm{rca})}-\bar{D}_{t-1}^{*}+\bar{B}_{t}-F_{t}\right)\right. \\
& \left.+\alpha\left(\tilde{C}_{t-1(\mathrm{rca})}-\bar{D}_{t-1}^{*}\right)\right] \tag{6.1}
\end{align*}
$$

where $\alpha$ is prescribed ( $1 / 3$, say, but in general could be $y$ specific), and the coefficient $b_{t(\mathrm{rca})}$ is computed using the gr-system as in rc-class of estimates. A simple interpretation of (6.1) can be obtained by comparing with the $a k^{*}$ estimator of (3.7). First write (3.7) as

$$
\begin{align*}
C_{t\left(a k^{*}\right)}=F_{t} & +k^{*}\left[\left(1-a^{*} / k^{*}\right)\left(\tilde{C}_{t-1\left(a k^{*}\right)}-\bar{D}_{t-1}^{*}+\bar{B}_{t}-F_{t}\right)\right. \\
& \left.+\left(a^{*} / k^{*}\right)\left(\tilde{C}_{t-1\left(a k^{*}\right)}-\bar{D}_{t-1}^{*}\right)\right] . \tag{6.2}
\end{align*}
$$

Now, for (6.1), $\alpha$ can be roughly viewed as the ratio of the two optimal coefficients $a^{*}, k^{*}$, and the factor $k^{*}$ outside the square brackets of (6.2) is replaced by the
(suboptimal) regression coefficient $b_{t(\mathrm{rca})}$. Thus $C_{(\mathrm{rca})}$ is not equivalent to the optimal $a k^{*}$-estimator, but some optimality could be preserved (if $\alpha$ is made $y$-specific) in setting the relative contribution of change and level driven predictors. Note, however, that the problem of internal inconsistency as mentioned in the introduction might arise if $\alpha$ is $y$-specific. Other attractive features of this version are that the value of $\alpha$ can be chosen to be well bounded away from zero (this should help to avoid sustained gaps between gr and rc series), and the number of extra controls is not doubled when both level and change driven predictors are included, thus allowing for introducing more controls as well as more degrees of freedom in variance estimation.

As a diagnostic of the impact of bias due to imputation of the previous month's employment status in view of the nonresponse of some of the present month respondents, the following simple check can be performed. The basic idea is to compute a multiplicative bias adjustment factor to the
estimator $\bar{D}_{t-1}^{*}$ involving imputed values. The factor is defined as the ratio of two gr-estimators of the previous month's characteristic based on the matched subsample. The denominator is a gr-estimator for the previous month (involving imputed values) while the numerator is a grestimator for the previous month (not involving imputed values), both computed in a somewhat nonstandard way. For the numerator, we use the time $t-1$ respondents with their time $t-1$ responses, and after nonresponse adjustment of the design weights, construct the gr-estimator with controls for time $t$. For the denominator, we assume that the subsets of each of the matched subsamples at $t-1$ and $t$ (here the matching is done with respect to each other, one forward in time and the other backward) not having the counterpart because of nonresponse, are statistically exchangeable with respect to each other. We then replace the time $t-1$ respondents who did not respond at time $t$ by the time $t-1$ nonrespondents who responded at $t$, along with their imputed time $t-1$ responses as well as design weights. Now the nonresponse, and gr-poststratification (with controls for $t$ ) weight adjustments are redone for this modified full sample at $t-1$. The gr-weights so obtained are used to compute the denominator mentioned above. One can now look at the time series of this factor over several months for diagnostics on the bias due to imputation. If this is not deemed close to one, then the average of the factor over several months can be treated as a nonrandom multiplicative bias adjustment to $\bar{D}_{t-1}^{*}$. In practice, instead of adjusting $\bar{D}_{t-1}^{*}$, it would be preferable computationally to adjust the new control $\tilde{C}_{\mathrm{f}-1(\mathrm{rc})}$ (of equation 3.6 ) for the corresponding characteristic by inverse of the above multiplicative factor. Altematively, the need for imputation can be avoided altogether if the questionnaire can be modified to obtain the necessary past information as suggested in the introduction.

The study of Lent, Miller and Cantwell $(1994,1996)$ considers the ak-composite weighted estimator for the U.S. Current Population Survey as an alternative to the currently used ak-estimator with $\mathrm{a}=0.2, \mathrm{k}=0.4$. Based on our experience with $a k^{*}$, it may be recommended that the $a k^{*}$ composite weighted estimator might be a better alternative in the interest of efficiency gains.

## ACKNOWLEDGEMENTS

The bulk of this research work was done when the first and third authors were at Statistics Canada. The authors are indebted to M. Sheridan, J.D. Drew, J. Gambino and especially M.P. Singh for their encouragement and several useful discussions. They are grateful to Jon Rao and Wayne Fuller for comments and suggestions. They are also grateful to J.M. Levesque, P. Lorenz, and especially T. Merkouris (with whom this work initially got underway) for their assistance in analysis and interpretation of results. Thanks are also due to the referee and the assistant editor Harold

Mantel for their useful suggestions for revision of the paper. The first author's research was supported in part by an NSERC grant held at Carleton University under an adjunct research professorship.

## REFERENCES

BAILAR, B.A. (1975). The effect of rotation group bias on estimates from panel surveys. Journal of the American Statistical Association, 70, 23-29.

BINDER, D.A., and HIDIROGLOU, M.A. (1988). Sampling in time. Handbook of Statistics, 6: Sampling, Elsevier Science, NY, 187211.

CASSEL, C.M., SÄRNDAL, C.-E. and WRETMAN, J.H. (1976). Some results on generalized difference estimation and generalized regression estimation for finite populations. Biometrika, 63, 615620.

COCHRAN, W.G. (1977). Sampling Techniques. $3^{\text {rd }}$ edition. John Wiley and Sons.

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

FULLER, W.A. (1999). The Canadian Regression Composite Estimation. Unpublished manuscript.

GODAMBE, V.P., and THOMPSON, M.E. (1989). An extension of quasi-likelihood estimation (with discussion). Journal Statistical Planning and Inference, 22, 137-172.
GURNEY, M., and DALY, J.F. (1965). A multivariate approach to estimation in periodic sample surveys. Proceedings of the Survey Research Methods Section, American Statistical Association, 247257.

HANSEN, M.H., HURWITZ, W.N. and MADOW, W.G. (1953). Sample Survey Methods and Theory, 2. John Wiley and Sons.

KUMAR, S., and LEE, H. (1983). Evaluation of composite estimation for the Canadian Labour Force Survey. Survey Methodology, 9, 178-201.

LENT, J., MILLER, S. and CANTWELL, P. (1994). Composite weights for the current population survey. Proceedings of the Survey Research Methods Section, American Statistical Assocation, 867-872.
LENT, J., MILLER, S. and CANTWELL, P. (1996). Effects of composite weights on some estimates from the Current Population Survey. Proceedings of the Section Survey Research Methods, American Statistical Association. I, 130-139.

LIANG, K.-Y., and ZEGER, S.L. (1986). Longitudinal data analysis using generalized linear models. Biometrika, 73, 13-22.
RAO, J.N.K., and GRAHAM, J.E. (1964). Rotation designs for sampling on repeated occasions. Journal of the American Statistical Association, 59, 492-509.
SÄRNDAL, C.-E. (1980). On $\pi$-inverse weighting versus best linear unbiased weighting in probability sampling. Biometrika, 67, 639 650.

SINGH, A.C. (1994). Sampling-design-based estimating functions for finite population totals. Invited paper, Abstracts of the Annual Meeting of the Statistical Society of Canada, Banff, Alberta, May 8-11, p. 48.

SINGH, A.C. (1996). Combining information in survey sampling by modified regression. Proceedings of the Survey Research Methods, American Statistical Association, 1, 120-129.

SINGH, A.C., and FOLSOM, R.E. Jr. (2000). Bias corrected estimating functions approach for variance estimation adjusted for poststratification. Proceedings of the Survey Research Methods, American Statistical Association, 610-615.

SINGH, A.C., KENNEDY, B., WU, S. and BRISEBOIS, F. (1997).
Composite estimation for the Canadian Labour Force Survey. Proceedings of the Survey Research Methods Section, American Statistical Association, 300-305.

SINGH, A.C., and MERKOURIS, P. (1995). Composite Estimation by modified regression for repeated surveys. Proceedings of the Survey Research Methods Section, American Statistical Association, 420-425.

WOLTER, K.M. (1979). Composite estimation in finite populations. Journal of the American Statistical Association, 74, 604-613.

# A Regression Composite Estimator with Application to the Canadian Labour Force Survey 

WAYNE A. FULLER and J.N.K. RAO ${ }^{1}$


#### Abstract

The Canadian Labour Force Survey is a monthly survey of households selected according to a stratified multistage design. The sample of households is divided into six panels (rotation groups). A panel remains in the sample for six consecutive months and is then dropped from the sample. In the past, a generalized regression estimator, based only on the current month's data, has been implemented with a regression weights program. In this paper, we study regression composite estimation procedures that make use of sample information from previous periods and that can be implemented with a regression weights program. Singh (1996) proposed a composite estimator, called MR2, which can be computed by adding $x$-variables to the current regression weights program. Singh's estimator is considerably more efficient than the generalized regression estimator for one-period change, but not for current level. Also, the estimator of level can deviate from that of the generalized regression estimator by a substantial amount and this deviation can persist over a long period. We propose a "compromise" estimator, using a regression weights program and the same number of $x$-variables as MR2, that is more efficient for both level and change than the generalized regression estimator based only on the current month data. The proposed estimator also addresses the drift problem and is applicable to other surveys that employ rotation sampling.


KEY WORDS: Survey sampling; Rotating samples; Combining estimators.

## 1. INTRODUCTION

Composite estimation is a term used in survey sampling to describe estimators for a current period that use information from previous periods of a periodic survey with a rotating design. When some units are observed in some of the periods, but not in all periods, it is possible to use this fact to improve estimates for all time periods.

Statistics Canada, U.S. Bureau of the Census and some other statistical agencies use a rotating design for labour force surveys. The current Canadian Labour Force Survey (LFS) is a monthly survey of about 59,000 households, which are selected according to a stratified multistage sampling design. The ultimate sampling unit is the household and a sample of households is divided into six panels (rotation groups). A rotation group remains in the sample for six consecutive months and is then dropped from the sample completely. Thus five-sixths of the sample of households is common between two consecutive months. Singh, Drew, Gambino and Mayda (1990) and Gambino, Singh, Dufour, Kennedy and Lindeyer (1998) contain detailed descriptions of the LFS design. In the U.S. Current Population Survey (CPS), the sample is composed of eight rotation groups. A rotation group stays in the sample for four consecutive months, leaves the sample for the succeeding eight months, and then returns for another four consecutive months. It is then dropped from the sample completely. Thus there is a 75 percent month-to-month sample overlap and a 50 percent year-to-year sample overlap (Hansen, Hurwitz, Nisselson and Steinberg 1955).

Patterson (1950), following the initial work by Jessen (1942), provided the theoretical foundations for design and
estimation for repeated surveys, using generalized least squares procedures. For the CPS, Hansen et al. (1955) proposed a simpler estimator, called the $K$-composite estimator. Gurney and Daly (1965) presented an improvement to the $K$-composite estimator, called the $A K$-composite estimator with two weighting factors $A$ and $K$. Breau and Ernst (1983) compared alternative estimators to the $K$-composite estimator for the CPS. Rao and Graham (1964) studied optimal replacement schemes for the $K$-composite estimator. Eckler (1955) and Wolter (1979) studied two-level rotation schemes such as the one used in the U.S. Retail Trade Survey. Yansaneh and Fuller (1998) studied optimal recursive estimation for repeated surveys. Fuller (1990) and Lent, Miller, Cantwell and Duff (1999) developed the method of composite weights for the CPS. The composite weights are obtained by raking the design weights to specified control totals that included population totals of auxiliary variables and $K$-composite estimates for characteristics of interest, $y$. Using the composite weights, users can generate estimates from microdata files for the current month without recourse to data from previous months.

The above authors used the traditional design-based approach, assuming the unknown totals on each occasion to be fixed parameters. Other authors (Scott, Smith and Jones 1977; Jones 1980; Binder and Dick 1989; Bell and Hillmer 1990; Tiller 1989 and Pfeffermann 1991) developed estimates for repeated surveys under the assumption that the underlying true values constitute a realization of a time series.

Statistics Canada considered $K$ and $A K$ composite estimation for the Labour Force Survey at several times during the past 25 years (Kumar and Lee 1983), but did not

[^2]adopt composite estimation. Instead, a generalized regression estimator, based only on the current months data, has been computed with a regression weights program. When composite estimation was considered in the 1990's, there was strong pressure to developed a composite estimation procedure that used the existing estimation program. Singh (1996) proposed an ingenious method, called Modified Regression (MR), to address this issue. This method leads to a composite estimator, called MR2 estimator, which uses the existing regression weights program. Singh suggested creating $x$-variables to be used as control variables in the regression program. With the created variables and the previous period estimator, the existing regression weights program is used to construct regression weights that define the estimator for the current period. Control variables with known population totals are also included.

An empirical study of the MR2 estimator identified several characteristics of the procedure. First, the estimated variance of a one-period change is much reduced. Second, the estimated variance of level is often similar to that for the direct estimator. Third, the estimator of level could deviate from the direct estimator by a substantial amount and this deviation could extend over a long period.

In this paper, we study the efficiency of MR estimators theoretically, under a simplified set-up. We propose also a "compromise" estimator that leads to significant gains in efficiency, for both level and change, over the estimator using only the current month's data. The composite estimator also addresses the "drift" problem mentioned above and can be implemented using the existing regression weights program. Gambino, Kennedy and Singh (2000) evaluated the efficiency of the composite estimates for the LFS data, using a jackknife method of variance estimation. Bell (2000) compared several composite estimators using data from the Australian Labour Force Survey.

## 2. COMPOSITE REGRESSION ESTIMATION

There are two types of observations used in composite estimation; those observed only at the current time, $t$, and those observed both at the current time and at the previous time, $t-1$. Sometimes information in previous observations is condensed in the estimate(s) for the previous period(s). Let $w_{i}$ be the sampling weight for observation $i$ at time $t$, let $A_{t}$ be the set of elements with observations at both the time periods $t$ and $t-1$, and let $B_{t}$ be the set of elements observed only at the current time period $t$. In this initial context, $i$ is the index for an individual respondent. If there is no nonresponse, the set $A_{t}$ for the LFS is composed of individuals in the five panels that were in the sample during the previous period, called the overlap panels. With no nonresponse, the set $B_{t}$ for the LFS contains individuals first observed in the current period, called the birth panel. Assume

$$
\sum_{i \in A_{t}} w_{i}+\sum_{i \in B_{t}} w_{i}=\hat{N}_{t}=\text { estimated population total. }
$$

Let $\theta_{\mathrm{f}}$ be the fraction of the sample in the overlap at time $t$

$$
\begin{equation*}
\theta_{t}=\hat{N}_{t}^{-1} \sum_{i \in A_{t}} w_{i} \tag{2.1}
\end{equation*}
$$

In the Labour Force Survey $\theta_{i}$ is about $5 / 6$ and is nearly constant over time. We will frequently omit the subscript $t$ on $A_{t}, B_{t}$ and $\theta_{i}$, for simplicity.

### 2.1 Estimator

Singh's (1996) MR2 estimator uses the control variable

$$
\begin{align*}
x_{1 t i} & =\theta_{t}^{-1}\left(y_{t-1, i}-y_{t i}\right)+y_{t i} & & \text { if } i \in A_{t} \\
& =y_{t i} & & \text { if } i \in B_{t}, \tag{2.2}
\end{align*}
$$

in the regression program, where $y_{t i}$ is the value of a characteristics of interest, $y$, for element $i$ at time $t$. Because of nonresponse in the LFS, Singh's original proposal used imputation for missing data and set $\theta=5 / 6$, after imputation for missing data. In our initial discussion we use the $\theta_{t}$ as defined in (2.1), assuming no nonresponse so that imputation is not required. Note that "micromatching" of individual data files at $t-1$ and $t$ is needed to calculate $x_{1 t i}$ and the resulting MR2 estimator. Additional control variables of the form (2.2) associated with other $y$-variables as well as auxiliary variables with known population totals are also included in the regression estimation. The auxiliary variables in the LFS include demographic variables such as age, sex and location.

The particular $x$-variables in (2.2) is designed such that the estimated total of $x_{1}$ is an estimator of the previous period total of $y$. Thus, the control total for $x_{1}$ in the regression procedure is the previous period estimator of the total of $y$.

Let $\hat{\rho}_{t-1}$ be the estimator of the mean of $y$ for period $t-1$, let $\bar{y}_{m, t-1}$ and $\bar{y}_{m t}$ be the means of the matched panels at time $t-1$ and $t$ respectively, let $\bar{y}_{t}$ be the grand mean of all sample panels at time $t$, and let $\bar{y}_{B_{i}, t}$ be the mean of the birth panel at time $t$. Assume the sample of size $n$ is divided into $g$ panels of equal size and denote the matched sampling fraction by $\theta$. To simplify the discussion we consider a single $y$-variable. Then Singh's (1996) MR2 estimator at time $t$, constructed with $x_{1 t i}$, can be written in a regression estimator form as

$$
\begin{equation*}
\hat{\mu}_{t}=\bar{y}_{t}+\left(\bar{x}_{C N t}-\bar{x}_{C t}\right) \hat{\beta}_{C_{t}}+\left[\hat{\mu}_{t-1}-\left(\bar{y}_{m, t-1}-\bar{y}_{m, t}+\bar{y}_{t}\right)\right] b_{r} \tag{2.3}
\end{equation*}
$$

where $\bar{x}_{C N t}$ is the population mean of the vector of auxiliary variables, such as age and sex, at time $t, \bar{x}_{C_{I}}$ is the weighted sample mean of the auxiliary variables, and ( $\left.\hat{\beta}_{c_{t}}^{\prime}, b_{t}\right)^{\prime}$ is the vector of regression coefficients for the regression of $y_{t}$ on ( $x_{C_{1}}, x_{1 t}$ ).

One can write

$$
y_{t, i}=\hat{y}_{t, i,(r)}+d_{t, i,(r)}
$$

where $\hat{y}_{t, i,(r)}$ is the predicted value in the regression of $y_{t, i}$ on $x_{C r}$ and $d_{t, 1,(r)}$ is the deviation from the regression predicted value. Then

$$
\begin{array}{cc}
x_{1 t}= & \theta^{-1}\left(\hat{y}_{t-1, i,(t-1)}+d_{t-1, i,(t-1)}\right. \\
& \left.-\hat{y}_{t, i,(t)}-d_{t, i,(t)}\right) \\
& +\hat{y}_{t, i,(t)}+d_{t, i,(t)} \\
=\hat{y}_{t, i,(t)}+d_{t, i,(t)} & \text { if } i \in A_{t} \\
& \text { if } i \in B_{t} .
\end{array}
$$

For demographic variables $X_{C t i}$, it is reasonable to believe that $\hat{y}_{t-1, i,(t-1)}$ is close to $\hat{y}_{t-1, i,(t)}$ and close to $\hat{y}_{t, i,(t)}$. Therefore the part of $x_{11}$ that is orthogonal to $x_{C T}$ is close to

$$
\begin{array}{rlr}
x_{d, 1, t}= & \theta^{-1}\left(d_{t-1, i,(t-1)}\right. & \left.-d_{t, i,(t)}\right) \\
& +d_{t, i,(t)} & \\
\text { if } i \in A_{t} \\
= & d_{t, i,(t)} & \text { if } i \in B_{r} .
\end{array}
$$

Thus the partial regression coefficient $b_{f}$ is close to the regression coefficient for the regression of $d_{t, i,(t)}$ on $x_{d, 1,1}$, and the value depends on the correlation between $d_{t, i,(t)}$ and $d_{t-1, i,(t-1)}$. A simple model for $d_{t, i,())}$ that has been used in the past, and the one we adopt in our analysis, is the assumption that the $d_{t, i,(t)}$ is the sum of a fixed $\mu_{t}$ and an error that is a first order autoregression with parameter $\rho$.

To simplify the presentation, we discuss the simple random sampling model without $x_{C}$. The results extend to the general case by considering the parameter $\rho$ to be the partial correlation between $y_{t}$ and $y_{t-1}$ after adjusting for $x_{C}$.

Under the autoregressive model with fixed $\rho$, an intercept and no other $\mathbf{x}_{C_{t}}$ in the model, it can be shown that $b_{t}$ converges in probability to

$$
b_{0}=\rho \lim _{n \rightarrow \infty} b_{t}=\theta \rho\left[2-\theta-2(1-\theta) \rho-(1-\theta) \sigma_{y}^{-2} \Delta_{t}^{2}\right]^{-1},
$$

where $\Delta_{t}^{2}=\left(\mu_{t}-\mu_{t-1}\right)^{2}$. Assuming ( $1-\theta$ ) $\sigma_{y}^{-2} \Delta_{t}^{2}$ is small relative to the other terms we get

$$
\begin{equation*}
b_{0} \doteq \theta \rho[2-\theta-2(1-\theta) \rho]^{-1} . \tag{2.4}
\end{equation*}
$$

For the LFS, $b_{0}=(7-2 \rho)^{-1} 5 \rho$.
Altemative representations for the estimator $\hat{\mu}_{t}$, omitting $x_{C l}$, are obtained using the formula $\bar{y}_{t}=\theta \bar{y}_{m t}+$ $(1-\theta) \bar{y}_{B r}$.Thus

$$
\begin{align*}
&=(1-b) \bar{y}_{t}+\left[\hat{\mu}_{t-1}+\left(\bar{y}_{m t}-\bar{y}_{m, t-1}\right)\right] b \\
&=\theta {\left[\bar{y}_{m t}+\left(\hat{\mu}_{t-1}-\bar{y}_{m, t-1}\right) b\right] } \\
& \quad+(1-\theta)\left(\hat{\mu}_{t-1}-\bar{y}_{m, t-1} y_{m, t-1}+\bar{y}_{m t}\right) b+(1-\theta)(1-b) \bar{y} . \\
&= {[\theta+(1-\theta) b]\left[\bar{y}_{m, t}+\left(\hat{\mu}_{t-1}-\bar{y}_{m, t-1}\right) b b^{*}\right] } \\
&+(1-\theta)(1-b) \bar{y}_{B, t} \\
&=\lambda_{A}\left[\bar{y}_{m, t}+\left(\hat{\mu}_{t-1}-\bar{y}_{m, t-1}\right) b b^{*}\right]+\left(1-\lambda_{A}\right) \bar{y}_{B, t}, \tag{2.5}
\end{align*}
$$

where

$$
1-\lambda_{A} \approx(1-\theta)\left(1-b_{0}\right)
$$

and

$$
\begin{equation*}
b^{*} \approx\left[\theta+(1-\theta) b_{0}\right]^{-1} b_{0} . \tag{2.6}
\end{equation*}
$$

The first expression on the right of the equality of (2.5) gives the MR2 estimator as a linear combination of the direct estimator $\bar{y}_{t}$ and the difference estimator $\hat{\mu}_{t-1}+$ $\left(\bar{y}_{m t}-\bar{y}_{m, t-1}\right)$ i.e., in the form of a composite estimator. The final expression of (2.5) gives the estimator as a linear combination of a "regression-type" estimator based on the overlap panels and the mean of the birth panels.

### 2.2 An Alternative Estimator

It is possible to define altemative regression variables to use in regression composite estimation. We present a particular regression variable in this subsection. The associated regression estimator is not suggested as the ultimate estimator, but the estimator is a member of a class for which efficiency calculations are given. An alternative to Singh's (1996) MR2 estimator is outlined in section 5.

Define a variable to be equal to the previous period value if the individual is in the overlap sample and to be equal to the estimated mean for the previous period if the individual is in the birth sample. The regression variable is

$$
\begin{align*}
x_{2, t i} & =y_{t-1, i,} & & \text { if } i \in A_{t} \\
& =\hat{\mu}_{t-1} & & \text { if } i \in B_{t} . \tag{2.7}
\end{align*}
$$

If this variable is used in a regression estimator, the control mean is $\hat{\mu}_{t-1}$, the previous period estimator, because the mean for the created variable is estimating the mean for period $t-1$. Singh (1996) used a variable $\tilde{x}_{2 t i}$ similar to $x_{2 t i}$. In Singh's variable, the $\hat{\mu}_{t-1}$ in (2.7) is $\bar{\gamma}_{m, t-1}$ if $i \in B_{t}$.

Consider a regression estimator constructed with $x_{21}$ and recall that the control mean of $x_{2 t}$ is $\hat{\mu}_{t-1}$. The regression estimator using $x_{2 t}$ can be written

$$
\begin{equation*}
\hat{\mu}_{\text {reg }, t}=\tilde{y}_{t}+\left(\hat{\mu}_{t-1}-\bar{x}_{2, t}\right) \hat{\beta}, \tag{2.8}
\end{equation*}
$$

where $\hat{\beta}$ is the regression coefficient for the regression of $y_{\text {t }}$ on $x_{2 t}$ (subscript $t$ is dropped on $\hat{\beta}_{t}$ for simplicity), $\bar{y}_{t}$ is the sample mean of $y$ at time $t$, and $\bar{x}_{2, t}$ is the sample mean of $x_{2, i}$ for all sample panels at time $t-1$. The regression coefficient $\hat{\beta}$ is, approximately, the regression of $y_{t}$ on $x_{2 t}$ in the set $A$. The coefficient is not exactly the regression coefficient for the set $A$ because $\bar{y}_{m, t-1}$ is not equal to $\hat{\mu}_{t-1}$, but the difference between the two estimators will usually be small. Singh (1996) called the regression estimator constructed with $\tilde{x}_{2 i t}$, the MR1 estimator.

Using $\bar{y}_{t}=\theta \bar{y}_{m t}+(1-\theta) \bar{y}_{B t}$, the regression estimator of $\mu_{t}$ using $x_{2 t}$ as a control variable is given by

$$
\begin{equation*}
\hat{\mu}_{t}=(1-\theta) \bar{y}_{B, t}+\theta\left\{\bar{y}_{m, t}+\left(\hat{\mu}_{t-1}-\bar{y}_{m, t-1}\right) \hat{\beta}\right\} . \tag{2.9}
\end{equation*}
$$

The expression within curly brackets in (2.9) is the regression estimator of $\mu_{t}$ using the estimator $\mu_{t-1}$ and only the data from the matched sample $A$. Note that the regression estimator

$$
\begin{equation*}
\hat{\mu}_{m, t}=\bar{y}_{m, t}+\left(\hat{\mu}_{t-1}-\bar{y}_{m, t-1}\right) \hat{\beta}, \tag{2.10}
\end{equation*}
$$

where $\beta$ is the regression of $y_{t}$ on $y_{t-1}$ in the set $A$, is the optimal linear estimator for $\mu_{t}$ based on $\hat{\mu}_{t-1}$ and the data of set $A$. Note that $\beta=\rho$ if the variances are the same at the two time periods. Hereafter, we often set $\beta=\rho$.

Using the variable $x_{2 t}$ gives the optimal estimator, $\hat{\mu}_{m}$, based on data in set $A$, but it does not combine that estimator with the mean of set $B$ in an optimal way. As can be seen in (2.10), the weight given to the mean of set $B$ is $1-\theta$. In general, this weight is too large because the variance of the regression estimator is less than the variance of the simple mean.

## 3. OPTIMAL ESTIMATION

The way in which one chooses to combine the regression estimator for set $A$ with the mean of set $B$ depends on one's objective function and on the variance of $\hat{\mu}_{t-1}$. We give some illustrative calculations based on some simplifying assumptions. For convenience let $V\left\{\hat{\mu}_{t-1}\right\}$ be expressed as a multiple of the variance of the birth panel,

$$
\begin{equation*}
V\left\{\hat{\mu}_{t-1}\right\}=q_{t}^{-1} V\left\{\bar{y}_{B, t}\right\} . \tag{3.1}
\end{equation*}
$$

Assume

$$
\begin{gather*}
V\left\{\bar{y}_{t}\right\}=g^{-1} V\left\{\bar{y}_{B, t}\right\},  \tag{3.2}\\
\operatorname{Cov}\left\{\mu_{t-1},\left(\bar{y}_{m, t}-\bar{y}_{m, t-1} \beta\right)\right\}=0,  \tag{3.3}\\
\operatorname{Cov}\left\{\hat{\mu}_{t-1}, \bar{y}_{B, t}\right\}=0, \tag{3.4}
\end{gather*}
$$

and

$$
\begin{equation*}
\operatorname{Cov}\left\{\bar{y}_{B, t},\left(\bar{y}_{m, t}-\bar{y}_{m, t-1} \beta\right)\right\}=0 \tag{3.5}
\end{equation*}
$$

where $g$ is the number of rotation groups (panels). Assumption (3.1) is reasonable if the original panels have a covariance function well approximated by that of a first order autoregressive process. For the LFS, the zero covariances in (3.4) and (3.5), and assumption (3.2) are only approximations because $\bar{\gamma}_{B, t}$ is not based on an entirely independent sample.

We write the regression estimator based on the overlap as

$$
\hat{\mu}_{m, t}=\bar{y}_{m, t}-\bar{y}_{m, t-1} \beta+\hat{\mu}_{t-1} \beta
$$

and, with the assumptions, obtain

$$
\begin{equation*}
V\left(\hat{\mu}_{m l}\right)=\left[g^{-1} \theta^{-1}\left(1-\rho^{2}\right)+q_{t}^{-1} \rho^{2}\right] V\left\{\bar{y}_{B, t}\right\} \tag{3.6}
\end{equation*}
$$

For the LFS, $g=6$ is the number of panels. Now consider an estimator that is a linear combination of $\hat{\mu}_{m, 1}$ and $\bar{y}_{B, r}$,

$$
\begin{align*}
\hat{\mu}_{t} & =\lambda \hat{\mu}_{m t}+(1-\lambda) \bar{y}_{B, t} \\
& =\lambda\left(\bar{y}_{m, t}-\bar{y}_{m, t-1} \beta+\hat{\mu}_{t-1} \beta\right)+(1-\lambda) \bar{y}_{B, t}, \tag{3.7}
\end{align*}
$$

where $0 \leq \lambda \leq 1$ is to be determined. To minimize the variance of current level, given $\hat{\mu}_{t-1}$ with variance $q_{t}^{-1} V\left\{\bar{y}_{B, t}\right\}$, one would minimize

$$
\begin{align*}
V\left\{\rho_{t}\right\} & =V\left\{\lambda \rho_{m t}+(1-\lambda) \bar{y}_{B, t}\right\} \\
& =\lambda^{2} V\left\{\hat{\mu}_{m t}\right\}+(1-\lambda)^{2} V\left\{\bar{y}_{B, t}\right\}, \tag{3.8}
\end{align*}
$$

with respect to $\lambda$. Under the assumptions (3.3), (3.4) and (3.5), the optimum $\lambda$ for current level is

$$
\lambda_{\mathrm{opt}}=\left[g^{-1} \theta^{-1}\left(1-\rho^{2}\right)+q_{t}^{-1} \rho^{2}+1\right]^{-1}
$$

However, if one is planning on using the estimator for a long period of time, one must realize that only certain values of $q_{t}$ are possible in the long run. The value of $\lambda$ chosen to estimate $\mu_{t}$ determines the variance of $\mu_{t}$ and hence, determines the variance that will go into the estimator of $\mu_{t+1}$. Assuming $\beta=\rho$, we have

$$
V\left\{\mu_{t}\right\}=\left\{g^{-1} \theta^{-1} \lambda^{2}\left(1-\rho^{2}\right)+q_{t}^{-1} \lambda^{2} \rho^{2}+(1-\lambda)^{2}\right\} V\left\{\bar{y}_{B, t}\right\}
$$

or

$$
\begin{equation*}
q_{t+1}^{-1}=g^{-1} \theta^{-1} \lambda^{2}\left(1-\rho^{2}\right)+(1-\lambda)^{2}+\lambda^{2} \rho^{2} q_{t}^{-1} . \tag{3.9}
\end{equation*}
$$

Thus, for a given $\lambda$, the limiting value for $q_{t}{ }^{-1}$ is

$$
\begin{align*}
\lim _{t \rightarrow \infty} q_{t}^{-1}=\left(1-\lambda^{2} \rho^{2}\right)^{-1}\left[g^{-1}\right. & \theta^{-1} \lambda^{2}\left(1-\rho^{2}\right) \\
& \left.+(1-\lambda)^{2}\right] \tag{3.10}
\end{align*}
$$

This result is equivalent to that given by Cochran (1977), page 352 equation (12.86).

Table 1 contains values of the limit variances as the number of periods becomes large, for selected values of $\rho$ and $\lambda$, where $\theta=5 / 6$ and $g \theta=5$ for the LFS. The variances are standardized so that the variance of the direct estimator based on the mean of six panels is 1.00 . Thus, the entries are six times the limiting value in (3.10). If the correlation is 0.95 and $\lambda$ is set equal to 0.96 , the long run variance of current level is $70 \%$ of that of the direct estimator. If $\lambda$ is set equal to 0.90 , the long run variance is $58 \%$ of that of the direct estimator when $p=0.95$.

The first line in Table 1 is for $\lambda=5 / 6=\theta$. This is the $\lambda$ corresponding to the use of $x_{21}$ in a regression estimator. The variance with $\lambda=5 / 6$ is always smaller than that of the direct estimator because of the improvement associated
with the use of the regression estimator $\hat{\mu}_{m, r}$. Thus, if $\rho \neq 0$, the regression estimator with $x_{21}$ leads to significant reduction in variance over the direct estimator, $\bar{y}_{t}$, that uses current data only.

Table 1
Standardized Limit Variances of Level:
LFS Rotation Pattern

|  |  | $\rho$ |  |  | $\rho$ |
| :--- | :--- | :--- | :--- | :--- | :--- |
| $\lambda$ | 0.70 | 0.80 | 0.90 | 0.95 | 0.98 |
| 0.833 | 0.897 | 0.840 | 0.743 | 0.665 | 0.600 |
| 0.840 | 0.895 | 0.836 | 0.734 | 0.650 | 0.581 |
| 0.860 | 0.894 | 0.830 | 0.714 | 0.614 | 0.527 |
| 0.880 | 0.903 | 0.835 | 0.705 | 0.588 | 0.481 |
| 0.900 | 0.921 | 0.851 | 0.711 | 0.575 | 0.444 |
| 0.920 | 0.951 | 0.882 | 0.736 | 0.582 | 0.420 |
| 0.940 | 0.992 | 0.928 | 0.785 | 0.617 | 0.420 |
| 0.960 | 1.046 | 0.994 | 0.867 | 0.698 | 0.465 |
| 0.980 | 1.115 | 1.083 | 0.997 | 0.861 | 0.619 |
| 0.990 | 1.155 | 1.138 | 1.087 | 0.998 | 0.803 |
| 0.995 | 1.177 | 1.168 | 1.140 | 1.089 | 0.960 |

The optimal $\lambda$ is a function of $\rho$ and increases slowly as $\rho$ increases. For $\rho=0.0$, the optimal $\lambda$ is 0.833 , for $\rho=0.7$ the optimal $\lambda$ is about 0.85 , for $\rho=0.95$ the optimal $\lambda$ is about 0.91 and for $\rho=0.98$ the optimal $\lambda$ is about 0.93 .

We now turn to the MR2 estimator (2.3) which can be written as

$$
\hat{\mu}_{t}=\lambda_{A}\left[\bar{y}_{m, t}+\left(\mu_{t-1}-\bar{y}_{m, t-1}\right) b^{*}\right]+\left(1-\lambda_{A}\right) \bar{y}_{B, r},
$$

where $\lambda_{A}$ and $b^{*}$ are defined in (2.6). While the MR2 estimator is not a member of the class (3.7), to the degree that $b^{*}$ is "close to" $\rho$, it is "close to" a member of the class. For example if $\rho=0.95$, then $b_{0} \doteq 0.9314$ and $b^{*} \doteq 0.9422$. If $\rho=0.90$, then $b_{0} \doteq 0.8659$ and $b^{*} \doteq 0.8853$.

Using the limiting value $b_{0}$ of $b$, we have $\left(1-\lambda_{A}\right)=$ $(1-\theta)\left(1-b_{0}\right)$, where $b_{0}$ is given by (2.4). Then $\lambda_{A}=$ $0.9375,0.9568,0.9776,0.9886$, and 0.9954 for $\rho=0.70$, $0.80,0.90,0.95$ and 0.98 , respectively. From Table 1, the standardized variances of $\mu_{t}$ for these values of $\lambda_{A}$ are $0.986,0.982,0.978,0.976$, and 0.975 , for $\rho=0.70,0.80$, $0.90,0.95$, and 0.98 , respectively. Thus, the MR2 estimator for current level has an efficiency for level that is essentially the same as that of the direct estimator, $\bar{y}_{t}$. The efficiency of the MR1 estimator is that for $\lambda=0.833$ in Table 1 and is always superior to that of $\bar{y}_{t}$.

## 4. VARIANCE OF ONE-PERIOD CHANGE

Given $\mu_{t-1}, \bar{y}_{m, t-1}, \bar{y}_{m, t}$ and $\bar{y}_{B, t}$ theoptimal estimator of $\mu_{t-1}$ is no longer $\mu_{t-1}$ because $\bar{y}_{m, t}$ contains information about $\mu_{t-1}$. However, it is not customary practice to revise the estimator of $\mu_{t-1}$. Given no revision, the estimator of change is $\hat{\mu}_{t}-\hat{\mu}_{t-1}$.

Under no revision in $\hat{\mu}_{t-1}$ and conditions (3.2) through (3.5), the variance of $\hat{\mu}_{t}-\hat{\mu}_{t-1}$, where $\hat{\rho}_{t}$ is defined in (3.7), is

$$
\begin{align*}
V\left\{\hat{\mu}_{t}-\hat{\mu}_{t-1}\right\}= & V\left\{\lambda\left[\bar{y}_{t}+\left(\hat{\mu}_{t-1}-\bar{x}_{2, t}\right) \rho\right]\right. \\
& \left.+(1-\lambda) \bar{y}_{B, t}-\hat{\mu}_{t-1}\right\} \\
= & {\left[g^{-1} \theta^{-1} \lambda^{2}\left(1-\rho^{2}\right)+(1-\lambda)^{2}\right.} \\
& \left.+(\rho \lambda-1)^{2} q_{t}^{-1}\right] V\left\{\bar{y}_{B, t}\right\} . \tag{4.1}
\end{align*}
$$

Table 2 contains standardized limit variances of the estimated change, $\hat{\mu}_{t}-\hat{\mu}_{t-1}$, for selected values of $g$ and $\lambda$, with $g \theta=5$. The entries in the table are the limiting variances of estimated change divided by the variance of change based on the common elements, $V\left\{\bar{y}_{m, 2}-\bar{y}_{m, r-1}\right\}$. The variance of change based on the common elements is $2 \theta^{-1}(1-\theta)(1-\rho) V\left\{\bar{y}_{B, r}\right\}$.

Table 2
Standardized Limit Variances of No-Revision One Period Change: LFS Rotation Pattern

|  |  | $\rho$ |  |  |  |
| :--- | :--- | :--- | :--- | :--- | :--- |
| $\lambda$ | 0.70 | 0.80 | 0.90 | 0.95 | 0.98 |
| 0.833 | 1.039 | 1.168 | 1.550 | 2.312 | 4.595 |
| 0.840 | 1.024 | 1.142 | 1.492 | 2.189 | 4.277 |
| 0.860 | 0.989 | 1.079 | 1.345 | 1.872 | 3.454 |
| 0.880 | 0.963 | 1.029 | 1.223 | 1.607 | 2.756 |
| 0.900 | 0.947 | 0.993 | 1.127 | 1.391 | 2.181 |
| 0.920 | 0.940 | 0.970 | 1.055 | 1.222 | 1.723 |
| 0.940 | 0.942 | 0.959 | 1.007 | 1.100 | 1.379 |
| 0.960 | 0.953 | 0.961 | 0.982 | 1.024 | 1.146 |
| 0.980 | 0.972 | 0.975 | 0.980 | 0.991 | 1.021 |
| 0.990 | 0.985 | 0.986 | 0.987 | 0.990 | 0.998 |
| 0.995 | 0.992 | 0.993 | 0.993 | 0.994 | 0.996 |

Tables 1 and 2 make clear the cost of not revising the estimate of $\hat{\mu}_{t-1}$. For example, if $\rho=0.95$, the variance of no-revision one period change is minimized with $\lambda \doteq 0.99$, but the variance of level is minimized with $\lambda \doteq 0.91$. A compromise value of $\lambda=0.95$ gives a variance of level that is about $14 \%$ larger than optimal and a variance of change that is about $7 \%$ larger than the smallest variance of Table 2.

Using the values of $\lambda_{A}$ associated with the MR2 estimator, the entries in Table 2 are 0.940, 0.960, 0.979, 0.989 , and 0.996 for $\rho=0.70,0.80,0.90,0.95$ and 0.98 , respectively. Thus, given no revision, and ignoring the difference between $b_{0}$ and $\rho$, the MR2 estimator is nearly optimal as an estimator of change, unlike the MR1 estimator, where the MR1 estimator corresponds to $\lambda=$ 0.833 in Table 2.

## 5. A COMPROMISE ESTIMATOR

On the basis of Table 2, the efficiency of the MR2 estimator of change for the LFS based on $x_{1 s}$, for the norevision case, is quite good. The MR1 no-revision estimator of change based on $x_{2 t}$ has relatively poor efficiency because it is a member of the class (3.7) with $\lambda=\theta=$ 0.8333 . On the other hand, the MR1 estimator of level based on $x_{2}$, is superior to the MR2 estimator based on $x_{1}$, and there are members of the class (3.7) that are much superior to the MR2 estimator of level.

Because the $\lambda$ in the MR2 estimator is relatively large and the $\lambda$ for the MR1 estimator is relatively small, we can create approximations to most interesting members of the class (3.7) as linear combinations of (2.10) and (2.5). Let

$$
\begin{equation*}
x_{3, t i}=\alpha x_{1, t i}+(1-\alpha) x_{2, t i}, \tag{5.1}
\end{equation*}
$$

where $0 \leq \alpha \leq 1$ is a fixed number. The regression estimator based on $x_{3 t i}$ gives an approximation to a member of the class (3.7) with

$$
\begin{equation*}
\lambda=\alpha \lambda_{A}+(1-\alpha) \theta, \tag{5.2}
\end{equation*}
$$

where $\lambda_{A}$ is defined in (2.6). Thus, if $\rho=0.95$,

$$
\lambda=\alpha(0.9886)+(1-\alpha)(5 / 6),
$$

for the LFS rotation pattern with $\theta=5 / 6$ and $b_{0}=(7-2 \rho)^{-1} 5 \rho ; \lambda=0.95$ if $\alpha=0.75$.

We choose $\alpha$ to give the desired combination of $\bar{y}_{B, z}$ and the "regression estimator" based on observations in set $A$. If one does not revise the estimator of $\mu_{t-1}$, the preferred combination depends on the relative importance assigned to the variance of level and to the variance of change.

Table 3 gives the variance of the MR2 estimator $(\alpha=1)$ relative to the variance of the estimator constructed using $\alpha$ $=0.75$ and the variance of the estimator constructed using $\alpha=0.65$. An entry in Table 3 for $\hat{\mu}_{t}$ is expression (3.10) evaluated at $\lambda_{A}$ of (2.6) and $\rho$, divided by (3.10) evaluated at $\lambda$ of (5.2) and $\rho$. An entry for $\hat{\mu}_{t}-\hat{\mu}_{t-1}$ is expression (4.1) evaluated at $\lambda_{A}$ of (2.6) and $\rho$, divided by (4.1) evaluated at $\lambda$ of (5.2) and $\rho$. These are approximations to actual efficiencies because $\rho$ is used for the coefficient of $x_{3}$. It is clear from Table 3 that the compromise estimator is slightly inferior to the MR2 estimator for one-period change, but is much superior to the MR2 estimator for level. For example, with $\rho=0.95$ and $\alpha=0.65$, the relative efficiency of the compromise estimator is 1.62 for level and 0.87 for oneperiod change.

For larger values of $\rho$, the variance of change is much smaller than the variance of level. Thus, for $\rho=0.95$, the variance of level and of change for $\alpha=1.00$ are about 1.00 and 0.12 , respectively, while the variance of level and of change for $\alpha=0.75$ are about 0.67 and 0.13 , respectively, when expressed in common units.

The smaller $\alpha$ has the advantage that the composite estimator will be closer to the direct estimator. Thus,
potential biases associated with the composite estimator should be smaller with the smaller $\alpha$.

Table 3
Approximate Efficiencies of Compromise Estimators Relative to $\alpha=1$

|  |  | $\alpha=0.75$ |  |  | $\boldsymbol{a}=0.65$ |  |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| $\rho$ | $b_{0}$ | $1-\lambda_{A}$ | $\hat{\mu}_{t}$ | $\hat{\mu}_{t}-\hat{\mu}_{t-1}$ | $\hat{\mu}_{t}$ | $\hat{\mu}_{t}-\hat{\mu}_{t-1}$ |
| 0.70 | 0.625 | 0.0625 | 1.052 | 0.999 | 1.069 | 0.995 |
| 0.80 | 0.741 | 0.0432 | 1.099 | 0.994 | 1.129 | 0.984 |
| 0.90 | 0.865 | 0.0224 | 1.238 | 0.975 | 1.303 | 0.946 |
| 0.95 | 0.931 | 0.0114 | 1.502 | 0.936 | 1.616 | 0.875 |
| 0.98 | 0.972 | 0.0046 | 2.177 | 0.833 | 2.321 | 0.712 |

## 6. DRIFT PROBLEM

As noted in the Introduction, the MR2 estimator could deviate from the direct estimator by a substantial amount and this deviation could extend over a long period. We now illustrate the basis for this phenomenon. We can express the deviation of the compromise regression estimator $\hat{\mu}_{r}$, based on $x_{3 t i}$, from the true mean $\mu_{t}$ as

$$
\begin{align*}
& \hat{\mu}_{t}-\mu_{t}=(\lambda \rho)^{t}\left(\mu_{0}-\mu_{0}\right)+\sum_{j=0}^{t-1}(\lambda \rho)^{j} \\
& \quad\left[\lambda \bar{r}_{m, t-j}+(1-\lambda)\left(\bar{y}_{B, t-j}-\mu_{t-j}\right)\right], \tag{6.1}
\end{align*}
$$

where $\mu_{0}$ is the mean at the initiation of the process and

$$
\bar{r}_{m t}=\bar{y}_{m t}-\mu_{t}-\rho\left(\bar{y}_{m, t-1}-\mu_{t-1}\right) .
$$

If $\rho$ is close to one and we use $\lambda=1$, then the emror $\hat{\mu}_{t}-\mu_{t}$ behaves roughly as a random walk which can lead to long periods in which $\hat{\mu}_{t}-\mu_{t}$ has the same sign. On the other hand, if $\alpha<1$ and $\rho=1$, then $\lambda<1$ and the error $\hat{\mu}_{t}-\mu_{t}$ exhibits less drift. For example, if $\alpha=0.70$, the correlation between adjacent errors $\hat{\mu}_{t}-\mu_{t}$, will be no greater than 0.95 under assumption (3.2)-(3.5). For the MR2 estimator, $\lambda \rightarrow 1$ as $\rho-1$ and hence the MR2 estimator can exhibit drift for $\rho$ close to one.

## 7. CONCLUDING REMARKS

For simplicity, we often assumed simple random sampling to obtain theoretical results. Similar results hold for complex designs and additional auxiliary variables, by considering $\rho$ to be a partial autocorrelation. Also, we used $x_{3}$-variables corresponding to only one variable $y$, but several $y$-variables can be used in constructing the corresponding $x$-variables for use in regression estimation. Gambino, Kennedy and Singh (2001) conducted an empirical study with LFS data using several $x_{3}$-variables with a common $\alpha$, and arrived at a compromise $\alpha$ for use in the LFS.

In section 2.1, we assumed no nonresponse so that imputation is not required. But in the LFS, nonresponse on an item $y$ may occur either at time $t-1$ or a time $t$ or at both time points. Gambino, Kennedy and Singh (2001) provide details of the imputation methods actually used in the LFS.

## ACKNOWLEDGEMENTS

The research of Wayne Fuller was partly supported by Cooperative Agreement 43-3AEU-3-80088 between Iowa State University, the National Agricultural Statistics Service and the U.S. Bureau of the Census. We thank Harold Mantel for a careful reading of the manuscript that led to improvements.

## REFERENCES

BELL, P. (2001). Comparisons of alternative Labour Force Survey estimators. Survey Methodology, 27, 53-63.

BELL, W.R, and HILLMER, S.C. (1990). The time series approach to estimation for periodic surveys. Survey Methodology, 16, 195215.

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

BREAU, P., and ERNST, L. (1983). Alternative estimators to the current composite estimator. Proceedings of the Section on Survey Research Methods, American Statistical Association, 397-402.

COCHRAN, W.G. (1977). Sampling Techniques. $3^{\text {nd }}$ Ed. New York : John Wiley and Sons.
ECKLER, A.R. (1955). Rotation sampling. Annals of Mathematical Statistics, 26, 664-180.
FULLER, W.A. (1990). Analysis of repeated surveys. Survey Methodology, 16, 167-180.

GAMBINO, J.G., KENNEDY, B. and SINGH, M.P. (2001). Regression composite estimatiuon for the Canadian labour force survey : Evaluation and implementation. Survey Methodology, 27, 65-74.

GAMBINO, J.G., SINGH, M.P., DUFOUR, J., KENNEDY, B. and LINDEYER, J. (1998). Methodology of the Canadian Labour Force Survey. Catalogue no. 71-526, Statistics Canada.
GURNEY, M., and DALY, J.F. (1965). A multivariate approach to estimation in periodic sample surveys. Proceedings of the American Statistical Association, Section on Social Statistics, 242-257.

HANSEN, M.H., HURWITZ, W.N., NISSELSON H. and STEINBERG, J. (1955). The redesign of the census current population survey. Journal of the American Statistical Association, 50, 701-719.

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

JONES, R.G. (1980). Best linear unbiased estimators for repeated surveys. Journal of the Royal Statistical Society, B, 42, 221-226.

KUMAR, S., and LEE, H. (1983). Evaluation of composite estimation for the Canadian Labour Force Survey. Survey Methodology, 9, 1-24.

LENT, J., MLLER, S.M., CANTWELL, P.J. and DUFF, M. (1999). Effect of composite weights on some estimates from the Current Population Survey. Joumal of Official Statistics, 14, 431-448.
PATTERSON, H.D. (1950). Sampling on successive occasions with partial replacement of units. Journal of the Royal Statistical Society, B, 12, 241-255.

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

RAO, J.N.K., and GRAHAM, J.E. (1964). Rotation designs for sampling on repeated occasions. Joumal of the American Statistical Association, 59, 492-509.

SCOTT, A.J., SMITH, T.M.F. and JONES, R.G. (1977). The application of time series methods to the analysis of repeated surveys. Intemational Statistical Review, 45, 13-28.

SINGH, A.C. (1996). Combining information in survey sampling by modified regression. Proceedings of the Section on Survey Research Methods, American Statistical Association, 120-129.

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

TILLER, R. (1989). A Kalman filter approach to labor force estimation using survey data. Proceedings of the Section on Survey Research Methods, American Statistical Association, 1625.

WOLTER, K.M. (1979). Composite estimation in finite populations. Journal of the American Statistical Association, 74, 604-613.

YANSANEH, I.S., and FULLER, W.A. (1998). Optimal recursive estimation for repeated surveys. Survey Methodology, 24, 31-40.

# Comparison of Alternative Labour Force Survey Estimators 

PHILIP BELL ${ }^{1}$


#### Abstract

This paper looks at a range of estimators applicable to a regularly repeated household survey with controlled overlap between successive surveys. The paper shows how the Best Linear Unbiased Estimator (BLUE) based on a fixed window of time points can be improved by applying the technique of generalised regression. This improved estimator is compared to the $A K$ estimator of Gumey and Daly (1965) and the modified regression estimator of Singh, Kennedy, Wu and Brisebois (1997), using data from the Australian Labour Force Survey.


KEY WORDS: Composite estimator; Best linear unbiased estimator; Modified regression; Repeated Surveys.

## 1. INTRODUCTION

This paper looks at a range of estimators applicable to a regularly repeated household survey with controlled overlap between successive surveys. The common theme of the estimators is to use data from previous times to improve current estimates, by taking advantage of correlations in the overlapping sample. I refer to all such estimators as composite estimators.

The estimators are evaluated for use in the Australian Labour Force Survey (LFS). In the LFS, overlap is controlled by dividing the first-stage sample of geographic areas into eight "rotation groups" from which dwellings are selected. In each month the same dwellings are selected from seven of the rotation groups, while new dwellings are selected in the remaining group. The sample consists of civilian persons aged 15 years old and over residing in the selected dwellings.

This sample design leads to high overlap of sample between two successive months within the seven "matched rotation groups". Using only data from these rotation groups rather than the whole sample can decrease the sampling error on an estimate of month to month movement. Composite estimation techniques seek to exploit this to give estimates with lower sampling error.

Section 2 of the paper introduces the Australian LFS and its current "generalised regression" estimator. The issue of time-in-survey bias (called rotation group bias by Bailar 1975) is also discussed.

Section 3 presents the " $A K$ composite" estimator proposed by Gurney and Daly (1965). This method has been used in the US Current Population Survey for many years. An extension known as " $A K$ composite weighting" has been used for the last few years; this was proposed by Fuller (1990) and studied by Lent, Miller and Cantwell (1994, 1996).

Section 4 presents the "modified regression" method of composite estimation (Singh and Merkouris 1995, Singh 1996). Here I focus on the MR2 estimator of Singh, et al.
(1997), which provides the largest reductions in sampling error. I also present a variant of this method suggested by Fuller (1999) for use in the Canadian Labour Force Survey.

Section 5 presents a "Best Linear Unbiased Estimator" (BLUE) based on data from a "window" containing a fixed number of successive months. This estimator was originally given by Jessen (1942) in the case of 2 occasions. A BLUE based on all occasions in a long series appears impractical, though a recursive approximation to this was developed by Yansaneh and Fuller (1998). This paper improves the fixed window BLUE described in Bell (1998) using the technique of generalised regression.

Section 6 gives the results of applying the different methods to the estimation of employed persons and unemployed persons in the LFS. Standard errors are estimated for longer-term indicators such as trend and trend movement, as well as for estimates of monthly level and its movement. Possible biases are explored, as well as evidence of change to seasonal patterns.

I conclude by comparing the advantages and disadvantages of the different types of estimator for application in the LFS. The improved BLUE estimator is found to be efficient, and when applied to the LFS is not subject to any large bias.

## 2. CURRENT ESTIMATES FOR THE LABOUR FORCE SURVEY

### 2.1 Overview of the LFS

The LFS has a multistage sample design, the first stage being a sample of small geographic areas known as "Census collector's districts" (CDs). A new sample of CDs is selected every five years, and the CDs are classified to eight "rotation groups". The dwellings selected from a CD remain in the sample for eight surveys, and then are replaced by other dwellings from the same $C D$. This replacement of dwellings is known as rotation, with all the dwellings in a rotation group being replaced at the same

[^3]time. Interviewers seek to collect data for all in-scope persons in the selected dwellings.

Of particular interest in the LFS is the person's labour force status - whether they are employed, unemployed or not in the labour force. The number of persons in each labour force status, for various categories of person, are key items to be estimated in the survey. Even more important to many users of the survey data than these level estimates are the estimates of movement in the figures between successive time points. It can be argued that longer-term indications of the direction of the series are even more important e.g., the movement of the X11 trend or of a similar smoother (Bell 1999).

The sample design ensures that the unconditional probability of selection $\pi_{t, j}$ is known for each sampled person $i$ at time $t$. This allows a simple estimator for a population total due to Horvitz and Thompson (1952). If $Y_{t}$ is the population item to be estimated at time $t$, and $y_{t i}$ is the same item as reported by the $i$-th unit at time $t$, the HorvitzThompson estimator is

$$
\begin{equation*}
\hat{y}_{t}^{\mathrm{H}}=\sum_{i} w_{t i}^{\pi} y_{t i} \tag{1}
\end{equation*}
$$

for $w_{t i}^{\pi}=\pi_{t i}^{-1}$, known as the selection weights.

### 2.2 The Generalised Regression (GR) Estimator

Generalised regression is a method for adjusting or "calibrating" a set of unit weights to add to a set of population attributes known as benchmarks. For a suitable choice of benchmarks the resulting weights give an improved estimate by taking account of externally available information.

In the LFS we start with the Horvitz-Thompson weights and calibrate them to add to demographic benchmarks that give numbers of people in the population for 560 poststrata ( 14 geographic regions classified by sex and 20 age groups). The weights from a given post-stratum are prorated to add to the stratum benchmark. This post-stratified ratio estimator is a particular case of the generalised regression or GR estimator.

Let $x_{t i}$ be a row vector of auxiliary variables for unit $i$ at time $t$, and $\hat{x}_{t}=\sum_{i} b_{t i} x_{t i}$ be estimates for the corresponding row vector of benchmark values $X_{t}$, based on some initial weights $b_{f i}$. The GR estimator based on these initial weights is then given by

$$
\begin{gather*}
\hat{y}_{t}^{G}=\hat{y}_{t}+\left(X_{t}-\hat{x}_{t}\right) \hat{\beta}  \tag{2}\\
\text { for } \hat{\beta}=\left(\sum_{i} b_{t i} x_{t i}^{\prime} x_{t i}\right)^{-1} \sum_{i} b_{t i} x_{t i}^{\prime} y_{t i} .  \tag{3}\\
\text { i.e., } \hat{y}_{t}^{G}=\sum_{i} w_{t i}^{G} y_{t i} \text { for } \\
w_{t i}^{G}=b_{t i}\left(1+\left(X_{t}-\hat{x}_{t}\right)\left(\sum_{i} b_{t i} x_{t i}^{\prime} x_{t i}\right)^{-1} x_{t i}^{\prime}\right) . \tag{4}
\end{gather*}
$$

In post-stratified ratio estimation the row vectors $x_{t i}$ contain zeroes except in the column corresponding to the unit's post-stratum, and $b_{t i}$ are the selection weights $w_{t i}^{\pi}$. In this case the regression parameters are just the post-stratum means, estimated using the selection weights.

### 2.3 Rotation Group Estimates

Each rotation group consists of a representative sample of dwellings, and so can provide a separate estimate. Number the rotation groups at a time point according to the number of times the dwellings in the rotation group have been sampled. Write $\mathrm{R}(t, i)=r$ if unit $i$ is in the rotation group sampled for the $r$-th time at time $t$. The HorvitzThompson estimate of $Y_{t}$ based on rotation group $r$ is

$$
\begin{equation*}
\hat{y}_{t}^{\mathrm{Hr}}=\sum_{i: R(t, i)=r} 8 w_{t i}^{\pi} y_{t i} . \tag{5}
\end{equation*}
$$

Generalised regression can be used to improve these estimators, by calibrating the weights to add to a set of benchmarks. Unfortunately the lower sample size in a single rotation group may require using a smaller number of benchmarks than in the overall case. In the LFS situation I used a single generalised regression step on the whole sample so that across the whole sample the weights add to the benchmarks for the current 560 poststrata, while in each rotation group the weights add to an eighth of the benchmarks for 71 collapsed poststrata. The resulting weights, when applied to a given rotation group $r$ and multiplied by eight, give the rotation group estimates $\hat{y}_{t}^{\mathrm{R} r}$.

### 2.4 Time-in-Survey Bias

Ideally rotation group estimates should have the same expectation $Y_{t}$, but in practice they have slightly different expectations, and hence different biases. Some of the difference is due to collection practices - for example, dwellings sampled for the first time are interviewed using a personal visit, while other rotation groups are mostly interviewed by telephone. It is not clear which rotation group is least affected by this sort of "time-in-survey" bias. The overall estimate will have a time-in-survey bias that is some mix of the biases from each rotation group. We rely on good survey methods to keep this bias small. Note that all the composite estimators will have different contributions from the rotation groups, and therefore different time-in-survey biases.

## 3. AK COMPOSITE ESTIMATION

### 3.1 AK Composite Estimator

The $A K$ composite estimator (Gurney and Daly 1965) is designed to put extra emphasis on the movement from the matched rotation groups (those rotation groups in which the same dwellings were selected in the current and previous months). The estimator has three components. The first is a mean of the rotation group estimates for the current month
data (time $t$ ). The second is last month's $A K$ composite plus a movement estimate based only on the matched rotation groups. The third component is the difference between estimates from the unmatched rotation group and from the matched ones. How much of each component to use is given by two parameters $A$ and $K$, as follows:

$$
\begin{align*}
\hat{y}_{t}^{\mathrm{AK}} & =(1-\mathrm{K}) \frac{1}{8} \sum_{r=1}^{8} \hat{y}_{t}^{\mathrm{Rr}} \\
& +\mathrm{K}\left(\hat{y}_{t-1}^{\mathrm{AK}}+\frac{1}{7} \sum_{r=2}^{8} \hat{y}_{t}^{\mathrm{R} r}-\frac{1}{7} \sum_{r=1}^{7} \hat{y}_{t-1}^{\mathrm{Rr}}\right) \\
& +\mathrm{A}\left(\hat{y}_{t}^{\mathrm{R} 1}-\frac{1}{7} \sum_{r=2}^{8} \hat{y}_{t}^{\mathrm{Rr}}\right) . \tag{6}
\end{align*}
$$

### 3.2 Choosing Parameter Values

The key parameter is $K$, which gives how much of the new estimate is based on the matched rotation group movement. The optimal $A$ and $K$ to use will depend on the variable being estimated. Higher $K$ values are appropriate for employment than for unemployment, since employment has a higher correlation between months.
$A K$ composite estimates of persons employed, unemployed and "not in the labour force" will not add correctly to the total population unless the same parameters are used for all the estimates. This leads to using a compromise choice of $A$ and $K$. The results in this report are based on $A=0.06$ and $K=0.7$. These values were found by trying a range of values of $A$ and $K$, and choosing those that gave optimal employed estimates. In this study no values of $A$ and $K$ gave unemployed estimates appreciably better than these values.

Our empirical study did not show particularly good sampling errors for the $A K$ estimator. The fine calibration that was used in obtaining the rotation group estimates may be to blame - it is possible that using broader categories would improve the sampling errors.

### 3.3 Properties of the $A K$ Estimator

The $A K$ estimator puts extra emphasis on the movement in the matched rotation groups. Thus the rotation group containing dwellings in sample for the first time contributes less than in the GR estimator. The $A K$ estimator thus has a different time-in-survey bias to the GR estimator.

The $A K$ estimator is recursive, in that last month's estimator is required in order to produce this month's. This is inconvenient for producing estimates for a new item or category. Also, the need to use the same values of $A$ and $K$ for all items can give sub-optimal estimates for any given item.

These concerns have led to the US Current Population Survey changing to a variant known as " $A K$ composite weighting" (Lent, Miller and Cantwell 1994). In $A K$ composite weighting, separate employed and unemployed estimates are produced for a number of important published
categories, using the $A K$ composite with optimal parameters for the estimate in question. The current data is then calibrated so that the unit weights add to these $A K$ estimates as well as demographic benchmarks. All estimates are then produced from the current dataset using these new " $A K$ composite weights".

The convenience of producing all estimates as a weighted sum of a single month's data is a major advantage of the $A K$ composite weighting approach. Another is that the most important estimates are $A K$ composite estimates with near-optimal choice of $A K$. A disadvantage is that only the most important estimates are true composite estimates. Any other estimates (including estimates of persons not in the labour force) are typically not much improved over the standard GR estimates (Lent, Miller and Cantwell 1996).

## 4. MODIFIED REGRESSION ESTIMATION

### 4.1 Overview of Modified Regression

The modified regression method is another way to provide composite estimates that can be obtained as weighted aggregates of the current survey dataset. The method targets a predetermined set of key items, for which it achieves particularly low sampling errors.

The modified regression technique uses generalised regression on the current month's dataset after attaching new auxiliary variables $z_{i i}$ to each unit $i$ at time $t$. Here $z_{t i}$ is a row vector with an element for each of the key items. Corresponding to these we have "pseudo-benchmarks" $Z_{\text {s }}$ based on the previous month's estimates for the key items. The modified regression estimator is then given by a generalised regression step applying both the demographic benchmarks and the pseudo-benchmarks.

$$
\begin{align*}
& \hat{y}_{t}^{\mathrm{M}}=\hat{y}_{t}^{\mathrm{H}}+\left(\left(X_{i}, Z_{t}\right)-\left(\hat{x}_{t}^{\mathrm{H}}, \hat{z}_{t}^{\mathrm{H}}\right)\right) \beta_{t}^{\mathrm{M}}  \tag{7}\\
& \text { for } \beta_{t}^{\mathrm{M}}=\left(\sum_{i} w_{t i}^{\pi}\left(x_{t i} z_{t i}\right)^{\prime}\left(x_{t i}, z_{t i}\right)\right)^{-1} \sum_{i} w_{t i}^{\pi}\left(x_{t i}, z_{t i}\right)^{\prime} y_{t i}  \tag{8}\\
& \text { i.e., } \hat{y}_{t}^{\mathrm{M}}=\sum_{i} w_{t i}^{\mathrm{M}} y_{t i} \text { for } \\
& w_{t i}^{\mathrm{M}}=w_{t i}^{\pi}\left\{1+\left(\left(X_{t}, Z_{t}\right)-\left(\hat{x}_{t}^{\mathrm{H}}, \hat{z}_{t}^{\mathrm{H}}\right)\right)\right. \\
& \left.\quad\left(\sum_{i} w_{t i}^{\pi}\left(x_{t i}, z_{t i}\right)^{\prime}\left(x_{t i}, z_{t i}\right)\right)^{-1}\left(x_{t i}, z_{t i}\right)^{\prime}\right\} \tag{9}
\end{align*}
$$

The key to the method is the definition of the auxiliary variables. Let D be the set of units in the matched rotation groups (those with dwellings selected at both time points) at time $t$. Let $y_{t i}^{*}$ be the vector of key items for unit $i$ at time $t$ and $Y_{t}^{*}$ the corresponding population totals. For $i \in \mathrm{D}$, let $y_{t-1, i}^{*}$ be the previous month's value for the vector of key items, or if no value was reported let $y_{t-1, i}^{*}$ be imputed -I used $y_{t-1, i}^{*}=y_{t, i}^{*}$ as suggested by Singh (1996).

I look at modified regression estimates for $z_{t i}$ of the following form, for $a \in[0,1]$ :

$$
\begin{array}{rlr}
z_{r i} & =(1-a) \frac{8}{7} y_{t-1, i}^{*}+a\left(y_{t i}^{*}-\frac{8}{7}\left(y_{t i}^{*}-y_{t-1, i}\right)\right) & \text { for } i \in \mathrm{D} \\
& =a y_{t, i}^{*} &  \tag{10}\\
\text { for } i \notin \mathrm{D} .
\end{array}
$$

Given this definition we have

$$
\begin{equation*}
\hat{z}_{t}^{\mathrm{H}}=(1-a) \hat{y}_{t-1}^{+\mathrm{HD}}+a\left(\hat{y}_{t}^{* H}-\left(\hat{y}_{t}^{* \mathrm{HD}}-\hat{y}_{t-1}^{+\mathrm{HD}}\right)\right), \tag{11}
\end{equation*}
$$

where $\hat{y}_{\mathrm{t}-1}^{+\mathrm{HD}}=8 / 7 \sum_{i \in \mathrm{D}} w_{t, i}^{*} y_{t-1, i}^{\pi}$ and $\hat{y}_{t}^{* H D}=8 / 7 \sum_{i \in \mathrm{D}}$ $w_{t, i}^{\pi} y_{t, i}^{*}$ are estimates of $Y_{t-1}^{*}$ and $Y_{t}^{*}$ respectively based on units in D only and using this month's selection weights. For $a=0, \hat{z}_{t}^{\mathrm{H}}$ is just the estimate $\hat{y}_{t-1}^{+\mathrm{HD}}$. For $a=1, \hat{z}_{t}^{\mathrm{H}}$ is this month's Horvitz-Thompson estimate minus an estimate of movement based on the matched rotation groups $\hat{y}_{t}^{* H D}-\hat{y}_{t-1}^{+\mathrm{HD}}$. Values $a=0$ and $a=1$ give the methods MR1 and MR2 respectively of Singh et al. (1997). Use of an intermediate $a$ was suggested by Fuller (1999).

An appropriate pseudo-benchmark $Z_{t}$ would be an estimate of $Y_{t-1}^{*}$ adjusted to agree with this month's weights. Following Singh et al. (1997) I used a step of generalised regression to adjust last month's modified regression estimator to add to this month's benchmarks:

$$
\begin{gather*}
Z_{t}=\hat{y}_{t-1}^{\bullet M}+\left(X_{t}-\hat{x}_{t-1}^{M}\right) \beta_{t}^{\text {adj }}  \tag{12}\\
\text { for } \beta_{t}^{\text {adj }}=\left(\sum_{i} w_{t-1, i}^{M} x_{t-1, i}^{\prime} x_{t-1, i}\right)^{-1} \sum_{i} w_{t-1, i}^{M} x_{t-1, i}^{\prime} y_{t-1, i}^{*} .
\end{gather*}
$$

Note that $Z_{t} \approx \hat{y}_{t-1}^{* M}$ since $\hat{x}_{t-1}^{\mathrm{M}}=X_{t-1} \approx X_{t}$. This completes the definition of the modified regression estimators.

### 4.2 Properties of Modified Regression Estimators

The movement $\hat{y}_{t}^{* H D}-\hat{y}_{t-1}^{\text {tHD }}$ at (11) is actually based on the matched sample only (i.e., units reporting at both times), since other units in the matched rotation groups D contribute zero to the movement (for the imputation used here). This may lead to the modified regression estimators having a lower sampling error than an $A K$ estimator, as this "matched sample movement" is not affected by units not present in both months.

Unfortunately, this also gives the possibility of a bias if persons not represented in the matched sample have different behaviour to those in the matched sample. This may well be the case - the matched sample excludes persons that changed dwelling between the two months, and it is possible that changes of dwelling are related to changes of employment. This "matched sample bias" will be in addition to any time-in-survey bias.

Another problem arises with the MR2 estimator (i.e., $a=1$ ). If the $k$-th key variable $y_{t i, k}^{*}$ has high month-tomonth correlation then it will also have a high correlation with the $k$-th new auxiliary variable $z_{\text {ri,k }}$. For such a
variable the element of $\beta_{t}^{\mathrm{M}}$ corresponding to $z_{t, k}$ will take some value $y_{t}$ close to one. Using (7), (11), and $Z_{t} \approx \hat{y}_{t-1}^{* M}$, the MR2 estimator takes the form

$$
\begin{align*}
\hat{y}_{t, k}^{* M} \approx\left(1-y_{t}\right) \hat{y}_{t, k}^{* H} & +y_{t}\left(\hat{y}_{t-1, k}^{* M}+\left(\hat{y}_{t, k}^{* H D}-\hat{y}_{t-1, k}^{+\mathrm{HD}}\right)\right) \\
& + \text { other terms. } \tag{14}
\end{align*}
$$

In this case it is possible that the matched sample movement at a given time will have a strong influence on estimates for many time points thereafter. In addition, any small bias in the movement will tend to accumulate over time. This danger was recognised by Fuller (1999), and referred to as "the drift problem". This was a motivation for his suggestion of the form of estimator given here, with a value of $a$ less than 1 .

In summary, modified regression has similar advantages to the $A K$ composite weighting approach, but with possibly lower sampling error. The method is not difficult to apply, and avoids the need to separately calibrate the rotation groups to the benchmarks.

## 5. BEST LINEAR UNBIASED ESTIMATION (BLUE)

### 5.1 Fixed Window BLUE

The fixed window BLUE estimator (denoted by $\hat{y}_{s}{ }^{\mathrm{B}}$ ) is obtained by choosing an "optimal" linear combination of the rotation group estimates $\hat{y}_{t}^{\mathrm{Rr}}$ (as defined in 2.3) from a window of $l+1$ months, as follows:

$$
\begin{equation*}
\hat{y}_{t}^{\mathrm{B}}=\sum_{s=t-1}^{1} \sum_{r=1}^{8} a_{s r} \hat{y}_{s}^{\mathrm{R} r} \tag{15}
\end{equation*}
$$

where the parameters $a_{\text {sfr }_{8}}$ are chosen to minimise $\operatorname{var}\left(\hat{y}_{t}^{\mathrm{B}}\right)$ under the constraints $\sum_{r=1}^{8} a_{s r}=1$ for $s=t$ and $\sum_{r=1}^{8} a_{s r}=0$ for $s=t-l, \ldots, t-1$. These constraints ensure that $\hat{y}_{t}^{g^{r}}$ will be unbiased for $Y_{t}$ provided that the rotation group estimates are unbiased, i.e., $\mathrm{E}\left(\hat{y}_{s}^{\mathrm{Gr}}\right)=Y_{s}$ for $s=t-l, \ldots, t$.

The minimisation requires knowing the variances and covariances of the rotation group estimates. In practice these are estimated based on historical data. The problem can then be written in a matrix form: we aim to choose the column vector $a$ (with elements $a_{s r}$ for $s=t-l, \ldots, t$ and $r=1, \ldots, 8$ ) so as to minimise a quadratic form $a^{\prime} \mathrm{V} a$ subject to constraints $C^{\prime} a=c$. The relevant standard result (Rao 1973 page 65) is that the minimum occurs for $a=\mathrm{V}^{-1} C q$ where $q$ is a solution of $\left(C^{\prime} \mathrm{V}^{-1} C\right) q=c$. In this study the matrix $V$ was replaced by a correlation matrix, under the assumption that all the rotation group estimates in the window had the same variance.

### 5.2 Correlation Structure of Rotation Group Estimates

Since different correlation patterns give different BLUE estimates, choosing a correlation pattern has similar issues
associated with it as choosing parameters A and K in the $A K$ composite. It is desirable to use the same linear combination for all estimates to assure additivity of the estimates.

I assumed a four parameter model for the correlation pattern:

$$
\begin{array}{rlrl}
\operatorname{corr}\left(\hat{y}_{t}^{\mathrm{Gr}}, \hat{y}_{s}^{\mathrm{Gr}}\right) & =\rho_{|t-s|}^{\mathrm{W}} & & \text { for } r-r^{\prime}=t-s \\
& =\rho_{|t-s|}^{\mathrm{B}} & & \text { for } r-r^{\prime}=t-s+8 m \\
& =0 & & \text { for integer } m \neq 0 \\
& =\text { otherwise. } \tag{16}
\end{array}
$$

Thus the correlation between estimates at lag $k$ from the same rotation group is $\rho_{k}^{W}$ if the rotation group contains the same dwellings at the two times, and $\rho_{k}^{\mathrm{B}}$ otherwise. Estimates from different rotation groups are uncorrelated. A four parameter model is used:

$$
\begin{gather*}
\rho_{k}^{\mathrm{W}}=\left(1-r_{U}^{2}\right)\left(\theta_{P}^{k} r_{P}^{2}+\theta_{B}^{k}\left(1-r_{P}^{2}\right)\right) \text { and }  \tag{17}\\
\rho_{k}^{\mathrm{B}}=\left(1-r_{U}^{2}\right) \theta_{B}^{k}\left(1-r_{P}^{2}\right) . \tag{18}
\end{gather*}
$$

Bell and Carolan (1998) discusses this model. The parameter values used in this paper were $\theta_{P}=0.87697$, $\theta_{B}=0.94, r_{U}=0.3101$ and $r_{P}=0.90456$. These values result from fitting the model to estimated autocorrelations for rotation group estimates of proportion employed.

It is important to note that the BLUE estimates are unbiased regardless of the correctness of the assumed correlation model. The model used here aims to be optimal for estimates of employed persons, but turns out to perform well for unemployed persons as well. Trying other values for the model parameters did not give any marked improvement in standard errors for unemployed persons.

### 5.3 Improved BLUE Estimates

A problem with the BLUE estimates above is that GR estimates are required at rotation group level. The lower sample size at rotation group level may limit the benchmarks that can be used, as discussed for the $A K$. For the BLUE, however, an alternative approach is available.

The B1 estimator is defined by forming a BLUE estimator based on the Horvitz-Thompson estimators at rotation group level, and then applying the generalised regression technique to improve this estimator. This proceeds as follows. Define $y_{t i}^{\#}=a_{t \mathrm{R}(t, i)} y_{t i}$ and $x_{t i}^{\#}=a_{t \mathrm{R}(t, i)} x_{t i}$, where $a_{t \mathrm{R}(, i)}$ is the BLUE multiplier applicable to the rotation group unit $i$ is in at time $t$. Then the BLUE estimator based on the Horvitz-Thompson estimators can be written

$$
\begin{equation*}
\hat{y}_{t}^{\mathrm{BH}}=\sum_{s=t-1}^{i} \sum_{i} w_{s i}^{\pi} y_{s i}^{\#} . \tag{19}
\end{equation*}
$$

Calibrating to the benchmarks we get the improved BLUE estimator B1:

$$
\begin{equation*}
\hat{y}_{t}^{\mathrm{B} 1}=\hat{y}_{t}^{\mathrm{BH}}+\left(X_{t}-\hat{x}_{t}^{\mathrm{BH}}\right) \hat{\beta} \tag{20}
\end{equation*}
$$

$$
\begin{equation*}
\text { for } \hat{\beta}=\left(\sum_{s=t-1}^{i} \sum_{i} w_{s i}^{\pi} \#_{s i}^{\prime^{\prime}} x_{s i}^{\#}\right)^{-1} \sum_{s=t-1}^{t} \sum_{i} w_{s i}^{\pi} i_{s i}^{\#^{\prime}} y_{s i}^{\#} \text {. } \tag{21}
\end{equation*}
$$

$$
\begin{equation*}
\text { i.e., } \hat{y}_{s}^{\mathrm{B} 1}=\sum_{s=t-1}^{t} \sum_{i} w_{s i}^{\mathrm{B1}} y_{s i} \tag{22}
\end{equation*}
$$

for $w_{s i}^{\mathrm{B} 1}=w_{s i}^{\pi} a_{s \mathrm{RR}(s, i)}\left\{1+\left(X_{t}-\hat{x}_{i}^{\mathrm{BH}}\right)\right.$

$$
\begin{equation*}
\left.\left(\sum_{u=s-1}^{s} \sum_{j} w_{w j}^{\pi} a_{u \mathrm{RR}(u j)} x_{u j}^{\prime} x_{k j}\right)^{-1} a_{s R(s, i)} x_{s i}^{\prime}\right)^{2} . \tag{23}
\end{equation*}
$$

## Properties of the Blue and B1 Estimators

The BLUE and B1 estimates are sums of weighted unit data from a window of months. Each estimate needs only data from this window, and can be produced independently from the estimates for previous months - so the method is not recursive.

The same month of data will contribute with different weights to the estimate for different times. A unit will contribute a sizeable weight to its current month estimate, and a weight near zero, often negative, to estimates for other months. The work required in producing a table is the same as for GR multiplied by the size of the window. There is also a possibility of negative estimates for tiny cells containing no current units.

Note that in the B1 estimator the weights applied to months other than the current one are not forced to sum to zero. Under the model assumptions the estimate $\hat{y}_{t}^{\mathrm{Bi}}$ remains unconditionally unbiased, since $\hat{y}_{t}^{\mathrm{BH}}$ and $\hat{x}_{t}^{\mathrm{BH}}$ are unbiased for $Y_{t}$ and $X_{t}$ respectively. In practice the current month contributes around 99.5 percent of the total weight. I consider the resulting bias to be small and not dangerous (leading as it does to some slight smoothing of the estimates over time).

For any estimate in which data from month to month is appreciably correlated, the BLUE and B1 estimates should have lower sampling error than the GR estimate. This is a theoretical advantage over a method that is designed for improving a predetermined set of estimates (like modified regression or the $A K$ with composite weighting). In practice this advantage may not be too important, as for the LFS much of our interest is in a small number of well-defined estimates.

The user must also determine the time period or "window" from which estimates will be used. Using too many time points will be expensive computationally, while too few will limit the gains available. The seven month window used here was sufficient to obtain nearly all the available gains, while smaller windows give noticeably higher standard errors.

## 6. COMPARING THE METHODS

### 6.1 Method of Comparison

Estimates for July 1993 to January 1999 were produced based on data from January 1993 to January 1999. Estimates were obtained classified by month, state, sex, marital status and employment status. Estimates were also obtained for lag one movement, quarterly average and movement between successive quarterly averages.

Standard errors for these estimates were calculated using the "delete-a-group jackknife" technique (Kott 1998). The geographic units that form the first stage of sample selection were divided systematically into $G=30$ groups, and "replicate groups" were formed consisting of the whole sample excluding the units from one of these groups. Each estimate studied was also produced for each of the $G$ replicate groups. Writing $e$ for the estimate and $e_{(g)}$ the estimate from replicate $g$, the delete-a-group jackknife estimate of standard error is given by

$$
\begin{equation*}
\mathrm{SE}_{(e)}=\sqrt{\frac{G-1}{G} \sum_{g=1}^{G}\left(e_{(g)-e}\right)^{2}} \tag{24}
\end{equation*}
$$

Estimates and standard errors were obtained for each of the following estimators (listed with short mnemonics for later reference):
GR: Generalised regression estimate as currently used in the LFS
$A K: A K$ estimate with $\mathrm{K}=0.7, \mathrm{~A}=0.06$
BL: BLUE based on 7 month window
B1: Improved BLUE based on 7 month window
MR2: MR2 estimator (modified regression with $a=1$ )
MF: Fuller's variant of modified regression ( $a=0.7$ )
The modified regression estimators require a choice of the key variables to be optimised for. In producing the modified regression estimates in this report, $z$ variables
were produced for estimates of employed and unemployed for each state and sex. This gives a total of 32 extra auxiliary variables, in addition to the usual 560 post-stratum benchmarks used in generalised regression.

### 6.2 Differences From GR Estimate

The current GR estimator can be used as a basis of comparison for the other estimators. Rather than present graphs of level estimates, I present the differences of the alternative estimates from the current GR estimates. Graphs 1 and 2 show these differences for estimates of employed persons and unemployed persons respectively. To put the size of these differences in perspective, note that the published standard errors for the current estimate were 25,200 for employed persons and 7,900 for unemployed persons in January 1999 (and similar for other months).

The $A K$, BL and B1 estimates are quite similar, since in all three methods the contribution of a unit depends on its rotation group. In both graphs the $A K, \mathrm{BL}$ and B 1 estimators appear to give lower values on average than the GR estimates. This indicates a change in the time-in-survey bias, resulting from putting less weight on the rotation group being sampled for the first time. The estimates vary up and down from their average difference for short periods.

The MR2 and MF estimates tend to be different to the other estimates since they emphasise the contribution of units from the matched sample. For employed persons, the MR2 and MF estimators are considerably larger on average than the GR estimates, up until September 1997. There is then a drop in the differences corresponding to the phase-in of a new sample from September 1997. For reasons that are not clear, over this period the matched sample behaved differently to the overall sample. This affects the difference between these modified regression series and the GR series. What may be of some concern is that the level change influences the level of the MR2 series for a considerable period thereafter, possibly a manifestation of the so-called "drift problem".


Graph 1. Difference of alternative estimates from GR estimate, employed persons ('000s), July 1993 to January 1999

For unemployed persons the M2 and MF estimates tend to be lower than the GR estimates. There is no evidence of a "drift problem" for unemployed persons, which is not surprising given the lower correlations involved.

### 6.3 Average Differences by Calendar Month

To quantify the likely change in bias from moving to a new estimator, the average difference over the period of each estimate from the GR estimate was calculated. It is possible that this difference is seasonal, so averages were obtained separately for each month of the calendar year, as well as overall. Average differences over the period July 1993 to January 1999 are given for employed persons in graph 3.

The graph shows that estimates of employed persons would have been higher on average using the MR2 or MF estimator. This upward difference for the modified regression estimators may actually be a feature of the particular period, since the difference has apparently dissipated since September 1997.

The other feature of the MR2 and MF estimates is that the difference for employed is highly seasonal. For example, the movement from December to January of the MR2 estimates is about 40,000 higher than the movement in the GR estimates. This suggests that the matched sample tends to miss people who were employed in December but not in January. The same seasonality shows up in looking at estimates from the matched sample directly. The matched rotation group movement does not show this large seasonal bias.

For the $A K$, BL and B1 estimates there is some seasonality in the differences, but the differences are much smaller.

Graph 4 shows the average differences of the various estimates from the GR estimate for unemployed persons over the same period. Here there appears to be a negative difference for all the estimators, though less pronounced for the $A K, \mathrm{BL}$ and B1 estimates than for the MR2 and MF. The change in seasonality from changing from the GR to the MR2 and MF estimators is again more extreme than for moving to the other estimators


Graph 2. Difference of alternative estimates from GR estimate, unemployed persons ('000s), July 1993 to January 1999


Graph 3. Average difference from GR estimate, overall and by calendar month, employed ('000)

### 6.4 Standard Errors

Standard errors (SEs) of estimates overall, by marital status and by sex are presented in the following graphs. The SE estimates are obtained as a percentage of the SE estimate for the same estimate using the GR method (i.e., the current LFS SEs), and these percentages are then averaged over the period for which they were produced (June 1993 to January 1999 for level estimates). Graphs 5,6, 7 and 8 show SEs of both employed and unemployed persons for level, movement, quarterly average and movement of quarterly average respectively.

For all these estimates the BLUE-class estimator B1 has lower sampling error than the $A K$ or BL estimators. Given that the B1 estimate appears to have similar bias and seasonality of bias it appears that the $A K$ and BL estimators used in this study are not competitive with the B1 estimator.

The modified regression estimators MR2 and MF, on the other hand, give much lower sampling errors than the B1 estimator for employed persons for overall estimates and estimates by sex. These are key estimates used in the modified regression - other key estimates such as state estimates also gave similarly improved standard errors. Estimates by marital status are not key estimates, and these have higher standard errors for MR2 and MF than for the B1 estimator.

For unemployed persons the improvement in SEs from using MR2 and MF are less consistent, disappearing altogether for estimates of quarterly average. The B1 estimator is more consistent in lowering the standard error, although the gains available for unemployed are lower than for employed.


Graph 4. Average difference from GR estimate, overall and by calendar month, unemployed ('000)


Graph 5. Standard Error of Level (\% of current SE)


Graph 6. Standard error of movement (\% of current SE)


Graph 7. Standard error of quarterly average (\% of current SE)


Graph 8. Standard error of movement of quarterly average (\% of current SE)

### 6.5 Seasonally Adjusted and Trend Series

The ABS uses the X11 package (Shiskin, Young and Musgrave 1967) to produce seasonally adjusted estimates that aim to remove various calendar effects from the series. The package also produces a trend, which is an indicator of the underlying behaviour of the series.

The trend value for a time point is revised as data for later times becomes available. I estimated the standard error of trend estimates at the end of the series (end trend) and for the same points when twelve further months of data are available (mid trend). Revisions of the trend (or trend movement) were defined as the difference between the mid and end values of the trend (or trend movement). The size
of the revision depends on the shape of the true series and on the sampling error in the estimated series. The mean squared trend revision for a series of unbiased estimates is the sum of two components: the mean squared trend revision that would have occurred even with no sampling error, and the variance of the estimate of revision. Thus the standard error of the revision is a measure of the sampling error component of the mean squared trend revision (see Bell 1999).

Seasonally adjusted figures are similarly subject to revisions. I present standard errors for level and movement of seasonally adjusted estimates at the end of the series. Standard errors for later revisions of these estimates were very similar.

The delete-a-group jackknife technique was used to produce estimates of standard error for the various trend and seasonally adjusted estimates. This technique requires producing replicate versions of the estimates. Unfortunately, the study provided replicate values for the time series only for time points from July 1993 to January 1999. Each of these replicate time series were supplemented by the previous 9 years of historical data so as to have sufficient data to apply the X11 package. Because the replicate seasonally adjusted and trend series are based on the same values before July 1993 the jackknife estimate of SE will tend to underestimate the true SE slightly, especially for times early in the series. To minimise this effect the measures of change in sampling error were averaged over months from January 1995 on only (and only up to January 1998, so that the 12 months to January 1999 can be used for estimating revisions).

Table 1
Standard error as percentage of standard error of current GR estimator

|  | $A K$ | BL | B1 | MR2 | MF |
| :---: | :---: | :---: | :---: | :---: | :---: |
| Employed persons: |  |  |  |  |  |
| level | 93 | 92 | 89 | 82 | 83 |
| movement | 95 | 95 | 89 | 66 | 69 |
| quarterly average | 93 | 92 | 89 | 85 | 85 |
| movement of quarterly average | 84 | 82 | 80 | 63 | 64 |
| seasonally adjusted | 94 | 92 | 90 | 87 | 88 |
| movement of seasonally adjusted | 96 | 95 | 91 | 68 | 71 |
| trend at end | 93 | 91 | 89 | 88 | 88 |
| movement of trend at end | 86 | 84 | 82 | 65 | 67 |
| revision of trend | 88 | 85 | 83 | 66 | 68 |
| revision of movement of trend | 89 | 86 | 84 | 67 | 69 |
| Unemployed persons: |  |  |  |  |  |
| level | 100 | 99 | 95 | 96 | 94 |
| movement | 101 | 101 | 95 | 87 | 86 |
| quarterly average | 100 | 99 | 95 | 100 | 98 |
| movement of quarterly average | 97 | 95 | 91 | 92 | 90 |
| seasonally adjusted | 100 | 99 | 95 | 96 | 95 |
| movement of seasonally adjusted | 102 | 102 | 95 | 87 | 86 |
| trend at end | 99 | 98 | 95 | 99 | 97 |
| movement of trend at end | 97 | 95 | 92 | 93 | 91 |
| revision of trend | 97 | 95 | 91 | 91 | 89 |
| revision of movement of trend | 97 | 95 | 92 | 92 | 90 |

Table 1 gives these average standard errors for various seasonally adjusted and trend measures, relative to those available from the current GR estimator, for both employed and unemployed persons. Also in the table are corresponding figures for level, movement, quarterly average and movement of quarterly average, as presented in graphs 5 to 8 .

I would argue that for many purposes the most important indicators are those that give the underlying direction of the series at the current end, i.e., movement of quarterly average, and movement of trend. A reduced standard error for these items makes the underlying direction of the series at the end clearer, even for users who rely on visual inspection or on some smoothing process other than the

X11 trend. This in tum improves the ability to detect turning points in the underlying series.

For movement of trend the B1 estimator achieves an $18 \%$ reduction in standard error for employed persons and an $8 \%$ reduction for unemployed persons. For the MR2 these reductions are $35 \%$ and $7 \%$ respectively. The composite estimators also reduce the contribution of sampling error to revisions in the trend series.

### 6.6 Summary

This paper presents a variant of the BLUE estimator, the B1 estimator, which applies the generalised regression technique to a composite estimate based on a window of seven months of data. On Australian data, the B1 has lower sampling error than the traditional BLUE or $A K$ estimators for a variety of measures including seasonally adjusted and trend estimates. The paper also evaluated a "modified regression" composite estimator MR2 proposed by A.C. Singh and a variant of this proposed by W. Fuller. These estimators gave considerably lower sampling errors than the B1 estimator for a number of measures, especially those based on employed persons.

The evaluation of a composite estimator will depend on many factors other than the sampling errors. The B1 estimator has the disadvantage that tabulations require weighted aggregation of seven months of data, whereas the modified regression estimators provide weights for a single month's data. On the other hand, the modified regression estimators may be biased if persons reporting in two successive months (the matched sample) are not representative of other persons (such as people moving house). Introducing the modified regression estimators would also induce a larger change in estimate and in seasonality than introducing the B1 estimator.

## ACKNOWLEDGEMENTS

The author wishes to thank the referees for their very helpful input. This work was supported by the Australian Bureau of Statistics. Views expressed in this paper are those of the author and do not necessarily represent those of the Australian Bureau of Statistics. Where quoted are used, they should be attributed clearly to the author.

## REFERENCES

BAILAR, B.A. (1975). The effect of rotation group bias on estimates from panel surveys. Journal of American Statistical Association, 70, 23-29.
BELL, P.A. (1998). Using state space models and composite estimation to measure the effects of telephone interviewing on labour force estimates. Working Papers in Econometrics and Applied Statistics, Catalogue no. 1351.0, no. 98/2, ABS, Canberra.

BELL, P.A. (1999). The impact of sample rotation pattems and composite estimation on survey outcomes. Working Papers in Econometrics and Applied Statistics, Catalogue no. 1351.0, no. 99/1, ABS, Canberra.

BELL, P.A., and CAROLAN, A. (1998). Trend estimation for small areas from a continuing survey with controlled sample overlap. Working Papers in Econometrics and Applied Statistics, Catalogue no. 1351.0, no. 98/1, ABS, Canberra.

FULLER, W.A. (1990). Analysis of repeated surveys. Survey Methodology, 16, 167-180.
FULLER, W.A. (1999). Canadian Regression Composite Estimation. Unpublished manuscript.

GURNEY, M., and DALY, J.F. (1965). A multivariate approach to estimation in periodic sample surveys. Proceedings of the Survey Research Methods Section, American Statistical Association, 247257.

JESSEN, R.J. (1942). Statistical investigation of a farm survey for obtaining farm facts. Iowa Agricultural station research Bulletin, 304.

KOTT, P.S. (1998). Using the delete-a-group jackknife variance estimator in practice. Proceedings of the Survey Research Methods Section, American Statistical Association, 483-486.

LENT, J., MILLER, S. and CANTWELL, P. (1994). Composite weights for the Current Population Surveys. Proceedings of the Survey Research Methods Section, American Statistical Association, 867-872.

LENT, J., MILLER, S. and CANTWELL, P. (1996). Effect of composite weights on some estimates from the Current Population Survey. Proceedings of the Survey Research Methods Section, American Statistical Association, 130-139.
RAO, C.R. (1973). Linear Statistical Inference and its Applications. Second edition, New York: John Wiley and Sons.

SINGH, A.C., and MERKOURIS, P. (1995). Composite estimation by modified regression for repeated surveys. Proceedings of the Survey Research Methods Section, American Statistical Association, 420-425.
SINGH, A.C. (1996). Combining information in survey sampling by modified regression. Proceedings of the Survey Research Methods Section, American Statistical Association, 120-129.
SINGH, A.C., KENNEDY B., WU S. and BRISEBOIS F. (1997). Composite estimation for the Canadian Labour Force Survey. Proceedings of the Survey Research Methods Section, American Statistical Association, 300-305.
SHISKIN, J., YOUNG, A. and MUSGRAVE, J. (1967). The X-11 variant of Census Method II Seasonal Adjustment, Bureau of the Census, U.S. Department of Commerce, Technical Paper 15.
YANSANEH, I.S., and FULLER, W.A. 1998. Optimal recursive estimation for repeated surveys. Survey Methodology, 24, 31-40.

# Regression Composite Estimation for the Canadian Labour Force Survey: Evaluation and Implementation 

JACK GAMBINO, BRIAN KENNEDY and MANGALA P. SINGH ${ }^{\mathbf{1}}$


#### Abstract

The Canadian Labour Force Survey (LFS) is a monthly survey with a complex rotating panel design. After extensive studies, including the investigation of a number of alternative methods for exploiting the sample overlap to improve the quality of estimates, the LFS has chosen a composite estimation method which achieves this goal while satisfying practical constraints. In addition, for variables where there is a substantial gain in efficiency, the new time series tend to make more sense from a subject-matter perspective. This makes it easier to explain LFS estimates to users and the media. Because of the reduced variance under composite estimation, for some variables it is now possible to publish monthly estimates where only three-month moving averages were published in the past. In addition, a greater number of series can be successfully seasonally adjusted.


KEY WORDS: Rotating panel survey; Estimation system; Weighting; Change estimate; Level estimate.

## 1. INTRODUCTION

### 1.1 Why Composite Estimation?

The Canadian Labour Force Survey (LFS) is a monthly survey of 54,000 households selected using a stratified multistage design. Households stay in the sample for six consecutive months, thus five-sixths of the sample is common between consecutive months. Each month, the members of a selected household are asked questions about their labour force status, earnings, and so on. In the LFS estimation system used prior to 2000 , initial design weights were modified using regression to produce final weights that respect age-sex and geographical (subprovincial region) population control totals. Each record then had a unique final weight that is used for all tabulations.

The estimation system used data from the current month only. No attempt was made to exploit the fact that the common sample can be used to improve estimates. However, characteristics such as employment by industry are highly correlated over time and unemployment is moderately correlated over time, thus there is potential for efficiency gains. Because of these gains, surveys similar to the LFS, such as the United States Current Population Survey (CPS), have used composite estimation to improve their estimates for many years. However, the LFS did not introduce composite estimation until January 2000.

In the early 1980s (see Kumar and Lee 1983), the CPS approach to composite estimation was studied for possible implementation in the LFS. Although the results showed that there were efficiency gains for Employed and, to a lesser extent, for Unemployed, it was felt that these gains were outweighed by the negative aspects of the method. These include the fact that the optimal parameters for Employed and Unemployed are quite different, which would have forced a trade-off between, on the one hand,
using a compromise set of parameters, thereby diluting the efficiency gains, and, on the other hand, having variables that do not add up to totals (e.g., Employed plus Unemployed would not equal Labour Force, unless one of the three is obtained as a residual). Another factor that worked against this form of composite estimation was that it was not compatible with the existing weighting, estimation and dissemination systems used by the LFS - the introduction of composite estimation would have required a complete overhaul of these systems.

Traditionally, the key estimates produced by the Labour Force Survey were monthly unemployment rates. However, with the increasing emphasis on estimates of employment level and on estimates of change in recent years, the need to find ways to make use of the common sample also increased since these estimates would benefit significantly. In the mid-1990s, therefore, interest in composite estimation was revived at Statistics Canada, and a regressionbased method that fit in well with the existing LFS estimation system was developed. This method is described in Singh, Kennedy, Wu and Brisebois (1997) with a more up to date version included in Singh, Kennedy and Wu (2001). The new methodology allows for a choice of methods, depending on one's objectives. If the primary interest is in estimates of level, then one can use leveldriven predictors in the procedure. If change is most important, then change-driven predictors can be used. One can go one step further and include both types of predictor in the procedure. However, in the latter case, the number of independent variables in the regression becomes large, which can lead to distortion of the final sample weights.

Preliminary results based on the new method using change-driven predictors and others using level-driven predictors were discussed at meetings of Statistics Canada's Advisory Committee on Statistical Methods. The method

[^4]addressed the problems with traditional composite estimators and showed substantial gains in efficiency. It was noted, however, that the estimator using change- driven predictors may lead to a drift in level estimates over time in some extreme situations. Also, it was decided, based on the committee's recommendation, that both estimates of level and of change should be given importance in the choice of predictors. After the exchange of technical notes between Wayne Fuller, J.N.K. Rao and Statistics Canada staff, a method suggested by Fuller, that combines the changedriven and level-driven approaches without the constraints associated with including both sets of predictors in the regression was adopted (see Fuller and Rao 2001). The solution is remarkably straightforward: take a linear combination of the level and change predictors: $X=(1-\alpha) X_{L}{ }^{+}$ $\alpha X_{C}$, and use it as the predictor. The change- and leveldriven predictors are now special cases. Furthermore, one can choose $\alpha$ to reflect the relative importance one wishes to give to level versus change.

The present paper describes the new composite estimator in section 2. An extensive evaluation of this estimator was carried out using actual LFS data for a number of characteristics over a long period of time. The results of these studies are summarized in section 3. Unlike traditional composite estimators, the regression based composite estimator requires that the matching of the sample between two consecutive months be done at the individual record level. This creates some interesting situations where one has to deal with nonrespondents and in scope and out of scope individuals between two consecutive months in such away that the quality of estimates of change is not compromised. Section 4 discusses the imputation procedure developed to deal with various situations that arise when dealing with incomplete data for two consecutive months. Finally, the success of this new composite estimator is judged not only on its statistical efficiency but its stability over time and its cost effectiveness, while achieving the following objectives: (i) minimizing changes to the old estimation system, (ii) producing a unique weight for each sample unit (iii) respecting age-sex and geography control totals and (iv) producing consistent estimates (in the sense that, e.g., Employed + Unemployed = Labour Force and Labour Force + Not In Labour Force = Population 15+). These objectives are discussed at various points in the paper, but especially in section 3.

## 2. THE REGRESSION COMPOSITE ESTIMATOR

Surveys such as the United States Current Population Survey have exploited their sample overlap by using $K$ composite or $A K$-composite estimators. Initially, the CPS used the $K$-composite estimator

$$
y_{t}^{\prime}=(1-K) y_{t}+K\left(y_{t-1}^{\prime}+\text { change }_{t-1, t}\right)
$$

with $K=1 / 2$ for time $t$, where change ${ }_{t-1, t}$ denotes an estimate of change based on the common, or matched, sample. This was later replaced by the $A K$-composite estimator

$$
\begin{aligned}
y_{t}^{\prime}= & (1-K) y_{t}+K\left(y_{t-1}^{\prime}+\text { change }_{t-1, t}\right) \\
& +A(\text { unmatched }- \text { matched })
\end{aligned}
$$

with $A=0.2$ and $K=0.4$ (see Cantwell and Ernst 1992). The optimal values of $A$ and $K$ depend on the variable of interest, and using different values for different variables poses problems of consistency (in the sense that parts do not add up to totals) in this approach. This prompted us to look for alternative approaches that satisfy the objectives mentioned at the end of the previous section.

It should be noted that we describe the new approach here at the person level, but in practice, person-level information is aggregated to the household level, and householdlevel records are then used by the estimation system.
.To use regression for weighting in the old LFS estimation system, a regression matrix $X$ is formed. Each person in the sample corresponds to a row of $X$. Each column of $X$ corresponds to a control total; e.g., column $c$ may be Male $20-24$, and the value in row $i$, column $c$ will equal 1 if person $i$ is a male between the ages of 20 and 24 , and 0 otherwise (similarly for columns comesponding to geographical areas). For further details on the estimation methods used by the Labour Force Survey, see Gambino, Singh, Dufour, Kennedy and Lindeyer (1998).

To exploit the/ sample that is common between months, the $X$ matrix is augmented by columns whose elements are defined in such a way that when this month's final weights are applied to the elements of each new column, the total is a composite estimate from the previous month, i.e., last month's composite estimate is used as a control total (strictly speaking, the control total is based on weights that reflect the current month's population). As we noted in the introduction, there are several ways to define the new columns, depending on one's objectives. We present below only the alternatives that were proposed for implementation.

A typical new column will correspond to employment in some industry, say agriculture. If one is primarily interested in estimates of level, the following way of forming columns produces good results. Let $M$ and $U$ denote the matched (common) and unmatched (birth) sample, respectively. For person $i$, and times $t-1$ and $t$, let $y_{i, t-1}$ and $y_{i, \text {, }}$ be indicator variables which equal 1 whenever the person was employed in agriculture. Then let

$$
x_{i}^{(L)}= \begin{cases}\bar{y}_{t-1}^{\prime} & \text { if } i \in U \\ y_{i, t-1} & \text { if } i \in M,\end{cases}
$$

where $\bar{y}_{t-1}^{\prime}$ is last month's composite estimate of the proportion of people employed in agriculture; in practice, we use $\bar{y}_{t-1}^{\prime}=\hat{Y}_{t-1}^{\prime} / P_{15+}$, where $P_{15+}$ denotes the population aged 15 and over. The corresponding control total is last month's estimate of the number of people employed in
agriculture, i.e., $\hat{Y}_{t-1}^{\prime}$. Thus the end result is that the final weighted sum of the elements of the new column will equal last month's estimate. This is almost the same as forcing this month's weights, applied to last month's values for the common sample, to reproduce last month's estimate of employment in agriculture (after adjusting by 5/6). We have used the superscript $L$ as a reminder that the goal here is to improve estimates of level.

If interest lies primarily in estimates of change, the following way of forming new columns of $X$ produces good results:

$$
x_{i}^{(C)}= \begin{cases}y_{i, t} & \text { if } i \in U \\ y_{i, t}+R\left(y_{i, t-1}-y_{i, t}\right) & \text { if } i \in M,\end{cases}
$$

where $R$ is a ratio that adjusts for the fact that five-sixths of the sample between months is common. The value $R=\Sigma_{\text {all }} w_{i} / \Sigma_{\text {matched }} w_{i}$ is used in the production system. For convenience, we used $R=6 / 5$ during development since, in practice, the difference between the two is small because procedures to balance the weights by rotation group are used (e.g., nonresponse adjustment is done separately by rotation group). As before, the corresponding control total is last month's estimate of the number of people employed in agriculture. Applying the final weights to the elements of this column of the $X$ matrix and summing produces the equality

$$
\hat{Y}_{t-1}^{\prime}=\hat{Y}_{t}^{\prime}-\hat{\Delta}_{t-1, t}^{M, f},
$$

or, in words, last month's estimate equals this month's estimate minus an estimate $\hat{\Delta}$ of $Y_{t}-Y_{t-1}$ based on the common sample. We use the superscript $f$ in $\hat{\Delta}$ as a reminder that the estimate of change is based on the final weights following composite estimation. In terms of the "pre-composite" weights, it is easy to show in the univariate case that

$$
\hat{Y}_{t}^{\prime}=(1-b) \hat{Y}_{t}+b\left(\hat{Y}_{t-1}^{\prime}+\hat{\Delta}_{t-1, t}^{M}\right)
$$

where $b$ is the regression coefficient and $\hat{\Delta}$ is the estimate of change based on the original weights. The more general case where auxiliary variables are present is given by Fuller and Rao (2001, equation 2.3).

Earlier results have shown that using the $L$ controls produces better estimates of level for the variables added to the $X$ matrix as controls. Similarly, adding $C$ controls produces good estimates of change for the variables that are added. Singh, et al. $(1997,2001)$ present efficiency gains for $C$-based estimates of level and change and refer to earlier results on $L$-based estimates.

Early in the development, an estimation system that used only the $C$-based controls was considered. However, there was some concern expressed about an estimation system based solely on change-driven controls since estimates of level are also very important (for example, they play a key role in the federal government's Employment Insurance program). These concems are summarized in Fuller and Rao (2001).

In principle, we can add both $L$ and $C$ controls to the regression, but this would result in a large number of columns in the $X$ matrix, which has undesirable consequences such as an increased number of extreme final weights, including negative weights. To avoid this, we would have to limit the number of industries included in the estimator. Wayne Fuller (see Fuller and Rao 2001) proposed an alternative which allows us to include the industries of greatest interest while allowing us to compromise between improving estimates of level and improving estimates of change. Fuller's alternative is to take a linear combination of the $L$ column and the $C$ column for an industry and use it as the new column in the $X$ matrix, i.e., use

$$
x_{i}=(1-\alpha) x_{i}^{(L)}+\alpha x_{i}^{(C)}
$$

The original level- and change-driven variables are special cases of Fuller's compromise.
Choice of $\alpha$ : Fuller and Rao (2001) showed that, based on some reasonable assumptions, values of $\alpha$ such as 0.65 and 0.75 produce reasonable estimates of both level and change. The actual choice of $\alpha$ depends on the variable of interest (specifically, its correlation over time) and on the relative importance of level versus change.

Our studies (see Appendix 1) showed that for the two most important variables, employed and unemployed, the best choices of $\alpha$ for estimates of level are 0.39 and 0.24 , respectively. The corresponding values for estimates of change are 0.99 and 0.81 , respectively. Clearly, there is a need to compromise between the goals of improving estimates of level and estimates of change.

To decide which values of $\alpha$ to study, we obtained compromise values of $\alpha$ by averaging the level-driven and change-driven values for each variable, i.e., we obtained approximately 0.7 and 0.52 for employed and unemployed, respectively. Results based on the values $\alpha=1$ and $\alpha=0.75$ had already been produced, so we added results for $\alpha=0.67$ and $\alpha=0.6$. Based on the results discussed below, which show that there are no substantial differences in the results for the three values $0.6,0.67$, and 0.75 , we chose to implement the value $\alpha=2 / 3$ in the production system.

## 3. FEATURES, PROPERTIES AND RESULTS

We present a summary of some of the features and properties of the regression composite estimator. Some graphical and numerical results are presented in section 3.1 below.

Systems implementation. An important advantage of the estimator is that it can be implemented within the old LFS estimation system in a straightforward manner since, essentially, one needs to augment the regression matrix, as described above. This was an important factor in our initiative to study and finally introduce composite estimation as
otherwise it would have cost a great deal more to change the system.

Weighting. Unlike the $A-K$ estimator, where weighting to satisfy population control totals and composite estimation are separate steps, weighting for the regression composite estimator is done in one step, i.e., simultaneously with weighting to satisfy the age-sex and geographical controls. For illustration, the way the regression matrix would be augmented when elements $x_{i}^{(C)}$ defined in section 2 are added is shown in Appendix 3. Adding the elements $x_{i}=(1-\alpha) x_{i}^{(L)}+\alpha x_{i}^{(C)}$ is similar. This not only preserves the consistency mentioned next but also retains the benefits of the controls applied to the usual regression estimator, i.e., the age-sex and geographic controls in our case.

Consistency. Because weighting for age-sex and geographical controls is done at the same time as weighting for the composite estimate controls, consistencies are preserved. In particular, parts add up to totals; e.g., Employed + Unemployed = Labour Force. In other approaches to composite estimation, consistency is achieved by other means which require either a separate step or a compromise of some kind.

Efficiency gains. For the variables that are added as control totals, there are substantial gains in efficiency for both estimates of level and of change. For $\alpha=1$, the gains for estimates of change can be dramatic; by choosing a smaller value of $\alpha$ we gain more for estimates of level while reducing the magnitude of the gains for estimates of change. Some results for the case $\alpha=2 / 3$ are given in section 3.1.

Seasonal adjustment. The time series of employment by various industries are scrutinized by both internal and external users of the Labour Force Survey. One important consequence of the gain in efficiency is that several of these series which could not be seasonally adjusted in the past can now be seasonally adjusted. In other words, composite estimation increases the signal-to-noise ratio sufficiently that seasonal adjustment becomes effective. A related consequence of composite estimation that is popular with users is that several estimates that were published as threemonth moving averages are now published as monthly estimates.

Systematic differences between composite and usual level estimates. In theory, the expectations, taken over all possible samples, for both the usual and composite estimators should be the same, making them both unbiased or almost unbiased. One would therefore expect that the estimates of level obtained using the two estimators would criss-cross each other over time. In practice, however, this does not happen. This is due to the fact that, when actual survey conditions are taken into account, the composite estimator and the usual estimator do not have the same expected value; for example, see Bailar (1975) and Kumar and Lee (1983) for results on the $K$ - and $A K$-composite estimator, respectively. Kumar and Lee show this by deriving explicit expressions for the expected value of the
usual estimator and the $A K$-composite estimator. The matched and unmatched samples differ because of differences in nonresponse rates and the mode of data collection (e.g., personal versus telephone interviewing, centralized versus decentralized interviewing). In practice, the units in the "birth" sample have a higher nonresponse rate, and the missing households tend to be smaller and have higher employment rates than the responding ones. Since the usual estimator and the composite estimator give different weights to the matched and unmatched sample, they will have different expected values. Thus time series for the two estimators can display systematic differences. In practice, these differences are usually swamped by sampling variation, but they become evident for more precise series such as Employed for big provinces like Ontario and for Canada. Our results for Employed are consistent with those described by Bailar (1975) for the U.S. Current Population Survey, i.e., the composite estimates for Employed tend to be smaller than the usual estimates. For Unemployed in Ontario, the difference between the two types of estimates tends to be much smaller.

Ways of reducing systematic differences between estimates from different rotation groups are currently being investigated. In particular, the possibility of introducing a weight adjustment for the number of households of different sizes by rotation group is being studied as a way of adjusting for the fact that small households are underrepresented in the birth rotation. This would benefit both the composite estimators and the usual regression estimator, and would probably reduce the gap between them.

### 3.1 Empirical Results

Employment and unemployment at the provincial level. Graph 1 shows total employment at the province level from 1987 to 1998 for Ontario. The time series for the composite estimation series for the four values of $\alpha$, i.e., for $0.6,0.67,0.75$ and 1 behave similarly. In these graphs, it is clear that there is a change in level for this series under composite estimation - the estimated number of employed persons is lower. The seasonally adjusted versions of the Ontario employment series based on the usual estimator and on the composite estimator for $\alpha=0.67$ are shown in Graph 2.

Graph 3 compares the usual estimates of Ontario unemployed to the regression composite estimate for $\alpha=0.67$. The effect of composite estimation on this variable is clearly less pronounced than on employmentrelated variables.

Graph 4 compares year-to-year changes in Ontario employment for the two estimators. Each point in the series is the difference between employment in year $y$, month $m$ and year $y-1$, month $m$. For example, the first point is January 1988 employment minus January 1987 employment. The composite estimation series is clearly smoother, especially in the second half of the twelve-year period.





Employment by subprovincial region. Graph 5 compares the usual estimate of employment with the composite estimate with $\alpha=0.67$ for an economic region in Ontario. The results for other subprovincial regions are similar. The behaviour of the usual and composite estimate series are very similar, thus, the effect of composite estimation is

neither beneficial nor harmful. For special tabulations, the LFS estimation system has the flexibility to allow the user to add controls at the economic region level if needed. There is already a control for the total population in each economic region.

Employment by industry, and seasonal adjustment. The composite estimates were compared to the usual regression estimate for sixteen industries. Graph 6A-6D show the results for two of them in Ontario. Though not included in these graphs, once again, the four values of $\alpha$ result in composite estimation series that generally behave similarly, although sometimes the series for $\alpha=1$ departs from the others. The composite estimation series tend to be less volatile than the regression series. This is particularly noticeable for the seasonally adjusted Trade series which we have included here because it illustrates the most extreme case. For this series, the behaviour of the original regression estimates in the first few years, in both the seasonally adjusted and unadjusted series, is difficult to explain from a subject-matter viewpoint. The behaviour of the Manufacturing series is more typical of the remaining fourteen industries.

Comparing the seasonally adjusted (Graph 6D) and unadjusted• (Graph 6C) series for Trade, we see that seasonal adjustment has had relatively little effect on the regression series, but has changed the composite series significantly. This is a manifestation of the ability of composite estimation to increase the signal-to-noise ratio sufficiently to make seasonal adjustment effective.

The seasonal adjustment program used by the Labour Force Survey computes a variety of measures that are used as indicators of the effectiveness of seasonal adjustment. Some of these measures are presented in Appendix 2. These show that, for Ontario employment in the twelve-year
period 1987-1998, composite estimation increases the number of industries that can be successfully seasonally adjusted. Results for other provinces and for Canada as a whole are similar.

A measure of stability. For several important data series, instead of monthly estimates, three-month moving averages were published in the past. This was due to the high sampling variability associated with these series, leading to unacceptable volatility in the monthly series. Of particular interest are province-level estimates by industry and by class of worker. It had been anticipated that the composite estimates for these series would demonstrate more stability, allowing the publication of monthly estimates instead of three-month averages. A measure of stability, the index of volatility, is computed as follows. For each industry, the month-to-month change in employment is calculated from seasonally adjusted estimates. Then the difference between consecutive change estimates is computed. The absolute value of this "change in the change" is expressed as a percentage of the corresponding monthly total estimate. These percentages are then averaged over the entire year. Large values of this measure occur when a series has many consecutive movements in opposite directions, indicating volatility.

The volatility index was computed for sixteen industries. Graphs 7A and 7B for two of these industries, Ontario Manufacturing and Trade, are included here, comparing the usual estimator, the three-month moving average of the usual estimator and the montly composite estimator with





Graph 6. Selected Employment by Industry


Graph 7. Index of Volatility
$\alpha=0.67$. For Manufacturing, the average indeces for the usual, composite and moving average estimates are 2.4, 1.8 and 0.60 , respectively. For Trade, the corresponding values are 2.4, 1.9 and 0.55 . For all industries, the volatility of the composite estimates typically falls between that of the usual monthly and three-month average estimates. Occasionally, for isolated years, the composite estimates are less volatile than the three-month averages or more volatile than the usual monthly estimates, but generally the volatility of the composite estimates is between that of the usual monthly estimates and that of the three-month moving averages. We also note that when the usual monthly estimates exhibit extreme volatility, the composite series tend to be more stable. The monthly regression estimates compete with the composite estimates only when the volatility index is low for both of them.

With the introduction of composite estimation, three-month moving averages were dropped in favour of the more desirable monthly estimates for industry series.

Variance estimates. For variables that are added as control totals, such as employment by industry, there can be substantial gains in efficiency at the province level, where efficiency is defined as $\operatorname{Var}(g r e g) / V a r$ (composite). For most industries, gains of 10 to 20 percent are typical, but they can be as a high as 40 percent. A 40 percent efficiency gain corresponds, for example, to reducing a 15 percent coefficient of variation to 12.7 percent and a 10 percent coefficient of variation to 8.5 percent. For province-level employment and unemployment estimates, the efficiency gains are more modest, typically in the five to ten percent range. For estimates of month-to-month change, the efficiency gains for controlled variables are bigger, usually more than double the gains for estimates of level.

For variables that are not controlled, there is little or no effect of composite estimation on efficiency unless the variable is highly correlated with a controlled variable. For example, at the province level, Employed Males shows a gain in efficiency because it is correlated with total employed, which is controlled. On the other hand, employment by subprovincial economic region shows neither gains nor losses.

## 4. TREATMENT OF MISSING DATA

By definition, the $x_{i}$ variables involve data from the current and previous month. This leads to complications when, for a given person in the common sample, data is available only for one month. This may occur due to nonresponse in either month or when a move or change in scope has taken place between the two months. The different cases that may occur are represented in the following diagram, where R denotes a response, X denotes a nonresponse and O denotes a unit that is out of scope.

|  | A | B | - | C | D |
| :--- | :---: | :---: | :---: | :---: | :---: |
| Month $t$ | XXX... | RRR... | RRR... | RRR... | OOO... |
| Month $t-1$ | RRR... | XXX... | RRR... | OOO... | RRR... |

In all these cases, namely A, B, C, and D, the objective is to find a solution such that $\sum_{i \in S} w_{i t} x_{i t}$ is still an estimator of $Y_{t-1}$. We set the following two objectives for handling the situation of missing data from either month of the common sample:
i) retain as many valid responses as possible, i.e., the option of removing a unit from the estimation process is rejected
ii) develop an imputation method that does not understate the estimate of change in any significant way.
In the case of nonresponse, there are two situations: Case A, where a household responded last month but not this month, and Case B which is the reverse situation. In the following, $i$ denotes a person in an affected household.
Case A: Replace $y_{i t}$ by $\hat{y}_{i t}$. This can be achieved in a number of ways. A simple approach is to replace $y_{i t}$ by the corresponding response from the previous month, i.e., $y_{i, t-1}$. During the early stages of the study, this approach was used but rejected later as it can bias (understate) the estimate of change significantly. For the LFS estimation system, it was decided to use the previous month's known demographic and employment characteristics of persons to
form imputation classes and then use hot deck imputation (i.e., current month's data) to obtain $\hat{\boldsymbol{y}}_{\text {ir }}$ An alternative would be to use a mean of some sort.
Case B: The procedure is analogous, i.e., when last month's value is missing, then imputation classes are formed using data from month $t$ and the donor is found using data from responding units in month $t-1$.

In the case where unit $i$ has moved or changed scope, the following situations may arise.
Case C: Suppose that unit $i$ was out of scope at time $t-1$ but is in scope at time $t$ (e.g., a person who just turned 15, or a newly arrived immigrant). Then unit $i$ should contribute 0 at time $t-1$ and $y_{i t}$ at time $t$. Hence we let $x_{i t}=0$ since $\Sigma w_{i t} x_{i t}$ should estimate $Y_{t-1}$.
Case D: Conversely, suppose that unit $i$ was in scope and is now out of scope. This includes, e.g., people who left the country, joined the military or died. Such units should be dropped since the target population is the in-scope population at time $t$ (and the ultimate goal is to estimate $Y_{t}$ ). Since we sample dwellings but collect data for individuals within those dwellings, two other situations arise due to movement of persons in and out of the sampled dwellings.
Case i): Suppose that unit $i$ was in the population at both times but in a sampled dwelling only at time $t$ (i.e., a person who moved from a non-sampled dwelling to a sampled dwelling). Then his/her status at time $t-1$ is unknown, i.e., $y_{i, r-1}$ is unknown. For all such cases, as in the nonresponse case, we can impute a value $\hat{y}_{i, t-1}$ for $y_{i, t-1}$ either from a donor in the sample or by a sample mean. The LFS uses hot deck imputation.
Case ii): Finally, consider the case where unit $i$ was in the sample at time $t-1$ but moved to a non-sampled dwelling at time $t$. Since the LFS sample is a sample of dwellings and not a sample of people, this unit should simply be dropped when computing estimates of $Y_{i}$.

## 5. CONCLUSION

The composite estimator described in this document meets all the objectives that were set at the beginning of this project and summarized in the introduction. It produces estimates of level and change that are more efficient than the estimates produced by the usual regression estimator while satisfying all operational and consistency constraints. The impact of the composite estimator with the value $\alpha=2 / 3$ on the many time series produced by the Labour Force Survey is generally moderate. When the impact is substantial, as in the Ontario Trade series, for example, the new series tend to make more sense from a subject-matter expert's perspective. This type of improvement in the series makes it easier to explain LFS estimates to users and the media.

The composite estimates have other features that users find very desirable. Because of the reduced variance under composite estimation, it is possible to publish monthly
estimates in many cases where only three-month moving averages were published in the past. In addition, a greater number of series can be successfully seasonally adjusted.

Implementation of composite estimation for the LFS is an important first step. Studies to improve the treatment of nonsampling errors are ongoing, and their results can be incorporated into the weighting and estimation system at any time. The system has the great advantage that it is very flexible. For example, the value of $\alpha$ can be changed easily, hence a comparison of a broad range of $\alpha$ values for a number of important variables is planned. This may lead to a system in which different $\alpha$ values are used for different control variables, while still having a unique final weight per record.

## ACKNOWLEDGEMENTS

We would like to thank Avi Singh and Statistics Canada's Advisory Committee on Statistical Methods for their contributions to this project. We are also grateful to the many people whose comments on earlier versions of this paper improved it greatly.

## APPENDIX 1

Relationship between $\alpha, \rho$ and ( $A, K$ ). Kumar and Lee (1983) found optimal values of $A$ and $K$ in $A K$-composite estimation for estimates of level and change as a function of the correlation coefficient $\rho$. We derived an approximate relationship between the $A$ and $K$ values, $\rho$ and $\alpha$. This result was then used to find good values of $\alpha$ for several variables. These are presented in Tables 1 and 2 for estimates of level and change, respectively. The $A$ and $K$ values in the tables are the optimal ones for the corresponding value of $\rho$. The values of $\alpha$ in the tables are consistent with those obtained by Wayne Fuller based on an AR(1) model (personal communication). The value of $\alpha$ for Labour Force in Table 2 exceeds one because of the approximation.

Table 1

| $\alpha$ Values for Several Variables - Level |  |  |  |  |
| ---: | :--- | :--- | :--- | :--- |
| Variable | $\rho$ | $A$ | $K$ | $\alpha$ |
| Employed | 0.852 | 0.49 | 0.8 | 0.385 |
| Unemployed | 0.58 | 0.38 | 0.5 | 0.242 |
| Labour Force | 0.843 | 0.48 | 0.8 | 0.403 |
| E.P. Agriculture | 0.955 | 0.38 | 0.8 | 0.448 |

Table 2
$\alpha$ Values for Several Variables - Change

| Variable | $\rho$ | $A$ | $K$ | $\alpha$ |
| ---: | :--- | :--- | :--- | :--- |
| Employed | 0.852 | 0.1 | 0.9 | 0.995 |
| Unemployed | 0.58 | 0.2 | 0.6 | 0.806 |
| Labour Force | 0.843 | 0.1 | 0.9 | 1.009 |
| E.P. Agriculture | 0.955 | 0 | 0.9 | 0.959 |

## APPENDIX 2:

Seasonal adjustment measures for Ontario employment by industry

|  | F Value |  |  |  | M7 |  |  | SMOOTH |  |
| :--- | ---: | ---: | ---: | ---: | ---: | ---: | ---: | ---: | :---: |
| Industry | greg | $\alpha=0.60$ | $\alpha=0.75$ | greg | $\alpha=0.60$ | $\alpha=0.75$ | greg | $\alpha=0.60$ |  |
| Agriculture | 87.76 | 120.18 | 112.7 | 0.27 | 0.23 | 0.24 | 37.94 | 45.36 |  |
| Forestry | 21.34 | 24.58 | 23.22 | 0.5 | 0.52 | 0.57 | 21.76 | 26.78 |  |
| Utilities | 4.29 | 3.48 | 6.8 | 1.1 | 1.25 | 0.82 | 15.39 | 15.52 |  |
| Construction | 128.3 | 275.06 | 246.93 | 0.26 | 0.16 | 0.17 | 41.68 | 57.5 |  |
| Manufacturing | 38.22 | 55.6 | 69.21 | 0.37 | 0.3 | 0.3 | 29.02 | 31.94 |  |
| Trade | 9.93 | 15.12 | 20.35 | 0.8 | 0.68 | 0.53 | 25.13 | 34.92 |  |
| Transportation | 9.16 | 8.64 | 9.69 | 0.94 | 0.75 | 0.7 | 15.36 | 23.33 |  |
| Finance | 6.49 | 8.94 | 8.84 | 1.22 | 0.76 | 0.77 | 13.45 | 19.67 |  |
| Professional | 5.3 | 12.91 | 9.81 | 1.03 | 0.72 | 0.76 | 12.45 | 19.52 |  |
| Management | 14.72 | 24.98 | 20.35 | 0.67 | 0.52 | 0.52 | 16.2 | 22.17 |  |
| Education | 67.37 | 219.62 | 214.37 | 0.33 | 0.16 | 0.19 | 53.25 | 66.47 |  |
| Health Care | 8.78 | 10.73 | 8.48 | 0.8 | 0.66 | 0.75 | 16.09 | 19.92 |  |
| Information | 21.13 | 52.31 | 62.94 | 0.66 | 0.36 | 0.35 | 24.29 | 33.46 |  |
| Accommodations | 44.85 | 75.37 | 78.03 | 0.36 | 0.34 | 0.3 | 31.89 | 44.29 |  |
| Other Services | 2.61 | 13.17 | 12 | 1.41 | 0.75 | 0.81 | 18.58 | 26.27 |  |

## Description of Measures

F-value: F -value for the test performed within the X11-ARIMA program to check for the presence of stable seasonality. The higher the F , the more significant is the presence of stable seasonality.

M7: Measure that combines the test for stable and moving seasonality. Generally, when M7 is greater than 1 , there is no identifiable seasonality present in the series; therefore, the series should not be adjusted.

SMOOTH: Percentage difference between the standard deviation of the month-to-month changes in the original series and the standard deviation of the month-to-month changes in the seasonally adjusted series. The larger this value, the more smoothing was obtained through the seasonal adjustment process.

## APPENDIX 3:

Implementing Regression Composite Estimation within the LFS Estimation Framework: Illustrated Using the Change-driven Approach

## Original $X$ matrix



Modified $X$ matrix for composite estimation when $x_{i}^{(C)}$ are added


For birth units, set a, b, c, . . to indicate this month's status (e.g., $a=1$ if employed, 0 otherwise). For matched units, do the following:
$a=e_{t}+\left(e_{t-1}-e_{1}\right) \times 6 / 5$ where $e=1$ if person is employed, $e=0$ otherwise
$\mathrm{d}=\mathrm{ag}_{\mathrm{t}}+\left(\mathrm{ag}_{-1} . \mathrm{ag}_{\mathrm{g}}\right) \times 6 / 5$ where $\mathrm{ag}=1$ if person is employed in agriculture, $\mathrm{ag}=0$ otherwise

## Examples:

(i) Suppose Person 2 was employed in agriculture both last month and this month. Then $e_{t-1}=e_{t}=1$ and $\mathrm{ag}_{\mathrm{t}-1}=\mathrm{ag}=1$, hence $\mathrm{c}=1-0=1$ and $\mathrm{d}=1-0=1$.
(ii) Suppose Person 2 was employed in agriculture last month and in mining this month. Then $e_{t .1}=e_{t}=1$, $\mathrm{ag}_{\mathrm{it}}=1$ and $\mathrm{ag}_{\mathrm{g}}=0$ hence $\mathrm{c}=1-0=1$ and $\mathrm{d}=0+$ $(1-0) * 6 / 5=1.2$.
(iii) Suppose Person 2 was employed in mining last month and in agriculture this month. Then $\mathrm{e}_{t \cdot 1}=\mathrm{e}_{\mathrm{t}}=1$, $\mathrm{ag}_{\mathrm{t}-1}=0$ and $\mathrm{ag}_{\mathrm{t}}=1$ hence $\mathrm{c}=1-0=1$ and $\mathrm{d}=1+$ $(0-1)^{*} 6 / 5=-0.2$.

## REFERENCES

BAILAR, B.A. (1975). The effects of rotation group bias on estimates from panel surveys. Journal of the American Statistical Association, 70, 23-30.

CANTWELL, P.J., and ERNST, L.R. (1992). New developments in composite estimation for the Current Population Survey. Proceedings: Symposium 92, Design and Analysis of Longitudinal Surveys, Statistics Canada, 121-130.
FULLER, W.A., and RAO, J.N.K. (2001). A regression composite estimator with application to the Canadian Labour Force Survey. Survey Methodology, 27, 45-51.
GAMBINO, J.G., SINGH, M.P., DUFOUR, J., KENNEDY, B. and LINDEYER, J. (1998). Methodology of the Canadian Labour Force Survey. Statistics Canada, Catalogue number 71-526.
KUMAR, S., and LEE, H. (1983). Evaluation of composite estimation for the Canadian Labour Force Survey. Survey Methodology, 9, 178-201.
SINGH, A.C., KENNEDY, B. and WU, S. (2001). Regression composite estimation for the Canadian Labour Force Survey with a rotating panel design. Survey Methodology, 27, 33-44.
SINGH, A.C., KENNEDY, B., WU, S. and BRISEBOIS, F. (1997). Composite estimation for the Canadian Labour Force Survey. Proceedings of the Survey Research Methods Section, American Statistical Association, 300-305.

# Variance Estimation After Imputation 

JAE-KWANG KIM ${ }^{\mathbf{1}}$


#### Abstract

Imputation is commonly used to compensate for item nonresponse. Variance estimation after imputation has generated considerable discussion and several variance estimators have been proposed. We propose a variance estimator based on a pseudo data set used only for variance estimation. Standard complete data variance estimators applied to the pseudo data set lead to consistent estimators for linear estimators under various imputation methods, including without-replacement hot deck imputation and with-replacement hot deck imputation. The asymptotic equivalence of the proposed method and the adjusted jackknife method of Rao and Sitter (1995) is illustrated. The proposed method is directly applicable to variance estimation for two-phase sampling.


KEY WORDS: Two-phase sampling; Item nonresponse; Deterministic imputation; Random imputation.

## 1. INTRODUCTION

Imputation, inserting values for missing items, is commonly used for handling missing survey data. An advantage of imputation is its convenience. That is, we can apply standard complete data programs for computing point estimates to the imputed data set. Rubin (1996), Fay (1996), and Rao (1996) reviewed various issues on imputation.

All imputation methods use some type of model. After designating a model, we can use either deterministic imputation or random imputation based on the model. Under random imputation, missing values are imputed by the use of some form of probability sampling. We call this additional random mechanism the imputation mechanism. On the other hand, deterministic imputation does not introduce an additional random mechanism. When the set of respondents is viewed as a random sample from the original sample, the selection mechanism of the respondents is called the response mechanism. The response mechanism is often regarded as the second phase of sampling. See Sämdal and Swensson (1987) for details.

With a suitable imputation model and method, the bias due to nonresponse can be greatly reduced relative to using only the observed data. However, it is well known that a variance estimator which uses the imputed data as if it were observed data is inconsistent.

Various methods have been proposed for variance estimation after imputation. Rubin and Schenker (1986) and Rubin (1987) advocate multiple imputation. Multiple imputation creates multiple data sets and calculates the complete data statistics for each imputed data set. The variance estimator is calculated by combining two terms, the withindataset variance term and the between-dataset variance term. Multiple imputation applies standard variance estimators to each data set to compute within-dataset variance terms and applies the standard point estimators to compute
a between-imputed-dataset variance term. This method requires the imputation method to be proper. That is, the imputation should satisfy conditions 1-3 in Rubin (1987, pages 118-119). These conditions are not always easy to achieve. (For example, see Fay 1992). Even the multiple imputation methods described in Schafer (1997) are not shown to be proper in the sense of Rubin. As noted by Rao (1996), some commonly used imputation methods, including hot deck imputation and regression imputation, are not proper.

Rao and Shao (1992) and Rao and Sitter (1995) proposed an adjusted jackknife variance estimator. The suggested procedure is applicable to a number of imputation methods and sample designs. The actual calculation using standard complete data software is not easy because special computations are performed to adjust the imputed values for each pseudo replicate. Also, Särndal (1992) proposed a variance estimation method that explicitly uses the model considered for imputation.

Essentially, Rubin's method generates several pseudo data sets for variance estimation and applies the standard variance estimators to each data set to compute the withindataset variance terms, while Rao's method and Särndal's method apply a special variance estimator to the imputed data set. In this paper, a method to create a single pseudo data set for variance estimation is proposed. In section 2, the new method is introduced in a two-phase sampling setup. In section 3, we illustrate extensions of the suggested method to the random imputation method. In section 4, we extend the suggested method to complex sampling designs. In section 5, comparisons are made with the adjusted jackknife variance estimator. In section 6, a limited simulation study is presented. Some concluding remarks are made in section 7. Outlines of some proofs are given in the appendix.

[^5]
## 2. A VARIANCE ESTIMATION METHOD

We outline a variance estimation procedure applicable for two-phase samples and for imputed samples. The procedure requires a separate data set for variance estimation in addition to the tabulation data set. To introduce the procedure and to illustrate the concepts, consider a two-phase sample. Let the second phase be a simple random sample of size $r$ selected from the first phase, which is a simple random sample of size $n$ selected from an infinite population. Let the regression estimator of the mean of a characteristic $y$ be

$$
\begin{equation*}
\hat{\rho}_{y}=\bar{y}_{2}+\left(\bar{x}_{1}-\bar{x}_{2}\right) \hat{\beta}, \tag{1}
\end{equation*}
$$

where

$$
\begin{aligned}
\left(\bar{y}_{2}, \bar{x}_{2}\right) & =r^{-1} \sum_{i=1}^{r}\left(y_{i}, x_{i}\right), \\
\bar{x}_{1} & =n^{-1} \sum_{i=1}^{n} x_{i}, \\
\hat{\beta} & =\left[\sum_{i=1}^{r}\left(x_{i}-\bar{x}_{2}\right)^{2}\right]^{-1} \sum_{i=1}^{r}\left(x_{i}-\bar{x}_{2}\right)\left(y_{i}-\bar{y}_{2}\right)
\end{aligned}
$$

and the second phase units are indexed from one to $r$. It is well known (e.g., Cochran 1977, equation 12.51) that the variance of the regression estimator is, approximately,

$$
\begin{equation*}
V\left\{\hat{\mu}_{y}\right\}=\left[n^{-1} \rho^{2}+r^{-1}\left(1-\rho^{2}\right)\right] \sigma_{y}^{2}, \tag{2}
\end{equation*}
$$

where $\rho$ is the population correlation between $y$ and $x$ and $\sigma_{y}^{2}$, is the population variance of $y$. An estimator of the variance is, by classical regression theory,

$$
\begin{align*}
\hat{V}\left\{\hat{\mu}_{y}\right\}= & n^{-1}(n-1)^{-1} \sum_{i=1}^{n}\left(\hat{y}_{i}-\bar{y}_{y}\right)^{2} \\
& +r^{-1}(r-2)^{-1} \sum_{i=1}^{r}\left(y_{i}-\hat{y}_{i}\right)^{2} \tag{3}
\end{align*}
$$

where $\hat{y}_{i}=\bar{y}_{2}+\left(x_{i}-\bar{x}_{2}\right) \hat{\beta}$ for $i=1,2, \ldots, n$, and $\bar{y}_{I}=n^{-1} \sum_{i=1}^{n} \hat{y}_{i}$. Observe that $\bar{y}_{I}$ is an alternative way of writing $\rho_{y}$ in (1).

Let

$$
\begin{equation*}
c_{r}=\left[n(n-1) r^{-1}(r-2)^{-1}\right]^{1 / 2} \tag{4}
\end{equation*}
$$

and

$$
y_{i}^{*}= \begin{cases}\hat{y}_{i}, & i=r+1, r+2, \ldots, n  \tag{5}\\ \hat{y}_{i}+c_{r}\left(y_{i}-\hat{y}_{i}\right), & i=1,2, \ldots, r .\end{cases}
$$

Then,

$$
\begin{equation*}
\hat{V}\left\{\hat{\mu}_{y}\right\}=n^{-1}(n-1)^{-1} \sum_{i=1}^{n}\left(y_{i}^{*}-\bar{y}_{I}\right)^{2} \tag{6}
\end{equation*}
$$

where $\bar{y}_{I}$ is the mean of the $y_{i}^{*}$, as well as the mean of the $\hat{y}_{i}$, because the sum of $y_{i}-\hat{y}_{i}$ is zero. Equation (6) is the operational form of the suggested estimator. The variance estimation data set contains the pseudo observation $y_{i}^{*}$.

To the extent that the model for imputation matches that of two-phase sampling, equation (6) is applicable to an imputed data set. For example, if we assume that missing data are missing at random and use regression to impute the missing value with $\hat{y}_{i}$, then equation (6) is immediately applicable. Of course, regression imputation or two-phase sampling can use a vector $x$.

## 3. EXTENSIONS TO RANDOM IMPUTATION

A moderate extension of the method described in section 2 enables us to estimate the variance of a sample mean using random imputation. In fact, alternative approaches are possible.

As one approach, assume that the imputation model is the regression model

$$
\begin{equation*}
y_{i}=\mathbf{x}_{i} \beta+e_{i} \tag{7}
\end{equation*}
$$

where the first element of every. $\mathbf{x}_{i}$ is equal to 1 and the $e_{i}$ are uncorrelated ( $0, \sigma_{e}^{2}$ ) random variables.

Assume the model is estimated and that the imputed values are

$$
\begin{equation*}
\ddot{y}_{i}=\hat{y}_{i}+\ddot{e}_{i}, \quad i=r+1, r+2, \ldots, n \tag{8}
\end{equation*}
$$

where $\hat{y}_{i}=\mathbf{x}_{i} \hat{\beta}$ with $\hat{\beta}=\left(\sum_{i=1}^{r} \mathbf{x}_{i}^{\prime} \mathbf{x}_{i}\right)^{-1} \sum_{i=1}^{r} \mathbf{x}_{i}^{\prime} y_{i}$ and $\ddot{e}_{i}$ is chosen at random from the set $\hat{\mathbf{e}}_{r}=$ $\left\{\hat{e}_{i}=y_{i}-\hat{y}_{i} ; i=1,2, \ldots, r\right\}$. The estimator of the mean of $y$ is

$$
\begin{equation*}
\hat{\mu}_{y}=n^{-1} \sum_{i=1}^{n} \ddot{y}_{i} \tag{9}
\end{equation*}
$$

where $\ddot{y}_{i}=y_{i}$ if $i=1,2, \ldots, r$.
If the $\ddot{e}_{i}$ are chosen with replacement with equal probability from the set $\hat{\mathbf{e}}_{r}$, then the variance $\hat{\mu}_{y}$ is, approximately,

$$
\begin{equation*}
V\left\{\hat{\mu}_{y}\right\}=\left[n^{-1} R^{2}+\left(r^{-1}+n^{-2} m\right)\left(1-R^{2}\right)\right] \sigma_{y}^{2}, \tag{10}
\end{equation*}
$$

where $m=n-r$ and $R^{2}$ is the squared multiple correlation coefficient between $y$ and $\mathbf{x}$. The increase in variance due to using random imputation with $\ddot{e}_{i}$, rather than using $\ddot{e}_{i}=0$, is $n^{-2} m\left(1-R^{2}\right) \sigma_{y}^{2}$.
Therefore, an estimator of the variance of the imputed sample mean is given by (6) where the $c_{r}$ of (4) is

$$
\begin{equation*}
c_{l}=\left[n(n-1)\left(r^{-1}+n^{-2} m\right)(r-p)^{-1}\right]^{1 / 2}, \tag{11}
\end{equation*}
$$

and $p$ is the dimension of $\beta$. We have

$$
\begin{align*}
\hat{V}\left\{\hat{\mu}_{y}\right\} & =n^{-1}(n-1)^{-1} \sum_{i=1}^{n}\left(\hat{y}_{i}-\bar{y}_{l}\right)^{2} \\
& +\left(r^{-1}+n^{-2} m\right)(r-p)^{-1} \sum_{i=1}^{r}\left(y_{i}-\hat{y}_{i}\right)^{2} \tag{12}
\end{align*}
$$

where $\bar{y}_{I}=\sum_{i=1}^{n} \hat{y}_{i}$. The estimator of the variance using $c_{I}$ of equation (11) is an estimator of the unconditional variance, the average over all possible imputed sample. Derivations of (10) and (12) are given in Appendix A.

To consider an alternative variance estimation approach, we assume that a random selection procedure is used for imputation but place no restriction on the procedure, other than that the probabilities of selection are inversely proportional to the probability that the $y$-value responds. In addition, we record the number of times an $\hat{\boldsymbol{e}}$ value is used as a donor in the imputation.
Let

$$
y_{i}^{*}= \begin{cases}\hat{y}_{i} & i=r+1, r+2, \ldots, n  \tag{13}\\ \hat{y}_{i}+c_{r}\left(y_{i}-\hat{y}_{i}\right) & i=1,2, \ldots, r\end{cases}
$$

with

$$
\begin{equation*}
c_{r}=\left[n^{-1}(n-1) r(r-p)^{-1}\right]^{1 / 2}\left(1+d_{i}\right) \tag{14}
\end{equation*}
$$

where $d_{i}$ is the number of times $\hat{e}_{i}$ is used as a donor. The term $\left[n^{-1}(n-1) r(r-p)^{-1}\right]^{1 / 2}$ is used to adjust for the effect of estimating $p$ regression parameters. Then, the variance estimator (6) can be written as

$$
\begin{align*}
& \hat{V}\left\{\hat{\mu}_{y}\right\}=n^{-1}(n-1)^{-1} \sum_{i=1}^{n}\left(\hat{y}_{i}-\bar{y}_{I}\right)^{2} \\
& \quad+n^{-2} r(r-p)^{-1} \sum_{i=1}^{r}\left(1+d_{i}\right)^{2}\left(y_{i}-\hat{y}_{i}\right)^{2} \tag{15}
\end{align*}
$$

If the imputation method is simple random sampling with replacement, then, conditional on the sample and the respondents,

$$
\begin{equation*}
E_{I}\left\{\left(1+d_{i}\right)^{2}\right\}=\left(\frac{n}{r}\right)^{2}+\frac{m}{r}\left(1-\frac{1}{r}\right) \tag{16}
\end{equation*}
$$

where the notation $I$ is used here to denote the expectation with respect to the imputation mechanism generated by random imputation. The equality in (16) establishes the equivalence of (12) to (15) under with-replacement selection. It is shown in Appendix B that $\hat{V}\left\{\hat{\mu}_{y}\right\}$ in (15) is also a valid estimator when donors are selected without replacement. Since the proposed variance estimation method is the conditional variance given the realized imputed sample, it has wide applicability.

## 4. COMPLEX SAMPLING DESIGNS

### 4.1 Deterministic Imputation

The suggested method is applicable to complex sampling designs as well as to simple random sampling. Assume that the full sample estimator of the mean of $y$ can be written as $\bar{y}=\sum_{i=1}^{n} w_{i} y_{i}$, where $w_{i}$ is the sampling weight of unit $i$ in the sample. Assume that $\sum_{i=1}^{n} w_{i}=1$.

If the first $r$ elements are observed and the remaining $n-r$ elements are missing, then the estimator of the mean of $y$ under regression imputation is

$$
\begin{equation*}
\bar{y}_{I}=\sum_{i=1}^{r} w_{i} y_{i}+\sum_{i=r+1}^{n} w_{i} \hat{y}_{i} \tag{17}
\end{equation*}
$$

where

$$
\begin{aligned}
& \hat{y}_{i}=\mathbf{x}_{i} \hat{\beta} \\
& \hat{\beta}=\left[\sum_{i=1}^{r} w_{i}^{*} \mathbf{x}_{i}^{\prime} \mathbf{x}_{i}\right]^{-1} \sum_{i=1}^{r} w_{i}^{*} \mathbf{x}_{i}^{\prime} y_{i} .
\end{aligned}
$$

Here $w_{i}^{*}$ is the sampling weight of unit $i$ in the secondphase sample and is defined by

$$
\begin{aligned}
& w_{i}^{*}=[\operatorname{Pr}(i \text { is in the second phase sample } \mid i \text { is in the } \\
& \text { first phase sample })]^{-1} w_{i} .
\end{aligned}
$$

Also, $\sum_{i=1}^{r} w_{i}^{*}=1$. If we assume that the second phase sample is a random sample of size $r$ from the $n$ first phase sample, then $w_{i}^{*}=n r^{-1} w_{i}$. Under certain conditions we can write the estimator in (17) as

$$
\begin{equation*}
\bar{y}_{l}=\sum_{i=1}^{n} w_{i} \hat{y}_{i} \tag{18}
\end{equation*}
$$

The representation (18) holds if $\left(w_{i}^{*}\right)^{-1} w_{i}$ is in the column space of the matrix $X=\left(\mathbf{x}_{1}^{\prime}, \ldots, \mathbf{x}_{r_{*}^{\prime}}^{\prime}\right)^{\prime}$ because then we have $\sum_{i=1}^{r} w_{i}\left(y_{i}-\hat{y}_{i}\right)=0$ from $\sum_{i=1}^{r} w_{i}^{*} x_{i}^{\prime}\left(y_{i}-\hat{y}_{i}\right)=0$.

We assume a sequence of samples and finite populations such as that described in Fuller (1998). Define $\bar{x}_{1}=\sum_{i=1}^{n} w_{i} x_{i}$ and $\left(\bar{x}_{2}, \bar{y}_{2}\right)=\sum_{i=1}^{r} w_{i}^{*}\left(\mathbf{x}_{i}, y_{i}\right)$. We also adopt the same assumptions as in Fuller (1998). That is

$$
\begin{equation*}
E\left(\overline{\mathrm{x}}_{1}, \overline{\mathrm{x}}_{2}, \bar{y}_{2}\right)=\left(\mu_{x}, \mu_{x}, \mu_{y}\right) \tag{19}
\end{equation*}
$$

and

$$
\begin{equation*}
V\left\{(\hat{\beta}-\beta)^{\prime}, \overline{\mathbf{x}}_{1}, \overline{\mathbf{x}}_{2}, \bar{y}_{2}\right\}=O\left(n^{-1}\right) \tag{20}
\end{equation*}
$$

where $\left(\mu_{x}, \mu_{y}\right)=N^{-1} \sum_{i=1}^{N}\left(\mathbf{x}_{i}, y_{i}\right)$ and $\beta=\left(\sum_{i=1}^{N} \mathbf{x}_{i}^{\prime} \mathbf{x}_{i}\right)^{-1}$ $\sum_{i=1}^{N} \mathbf{x}_{i}^{\prime} y_{i}$.

For $i=1,2, \ldots, N$, define

$$
a_{i}= \begin{cases}1 & \text { if unit } i \text { responds when sampled } \\ 0 & \text { otherwise }\end{cases}
$$

and $\mathbf{a}=\left(a_{1}, a_{2}, \ldots, a_{N}\right)$. The extended definition of $a_{i}$ is discussed by Fay (1991) and used in Shao and Steel (1999). Now, let

$$
\begin{equation*}
\bar{y}_{n}=\sum_{i=1}^{n} w_{i} \tilde{y}_{i}^{*} \tag{21}
\end{equation*}
$$

where

$$
\begin{equation*}
\tilde{y}_{i}^{*}=\tilde{y}_{i}+a_{i} w_{i}^{-1} w_{i}^{*}\left(y_{i}-\tilde{y}_{i}\right) \tag{22}
\end{equation*}
$$

with $\tilde{y}_{i}=x_{i} \beta$. Then, we have $\bar{y}_{I}=\bar{y}_{I l}+\left(\overline{\mathbf{x}}_{1}-\bar{x}_{2}\right)(\hat{\beta}-\beta)$. By (19) and (20), we have $\bar{y}_{I}=\bar{y}_{I l}+O_{p}\left(n^{-1}\right)$ and $V\left(\bar{y}_{I}-\bar{Y}_{N}\right)=V\left(\bar{y}_{l l}-\bar{Y}_{N}\right)+o\left(n^{-1}\right)$. Now,

$$
\begin{equation*}
\left.V\left(\bar{y}_{H}-\bar{Y}_{N}\right)=V\left[E\left(\bar{y}_{H}-\bar{Y}_{N} \mid \mathbf{a}\right)\right]+E\left[V\left(\bar{y}_{H}-\bar{Y}_{N}\right) \mid \mathbf{a}\right)\right] . \tag{23}
\end{equation*}
$$

The first term on the right side of (23) is 0 because $E\left(\bar{y}_{I l}-\bar{Y}_{N} \mid \mathbf{a}\right)=0$ under model (7). To estimate the second term in (23), note that conditional on a, $\bar{y}_{I I}$ is a linear estimator. Hence, the standard variance estimation method applied to the pseudo data set $\tilde{\mathbf{Y}}^{*} \equiv\left\{\tilde{\boldsymbol{y}}_{i}^{*} ; i=1,2, \ldots, n\right\}$ will unbiasedly estimate the variance of $\bar{y}_{l l}=\Sigma_{i=1}^{n} w_{i} \tilde{y}_{i}^{*}$. Since the set $\tilde{\mathbf{Y}}^{*}$ is not observable, we can use the set $\mathbf{Y}^{*} \equiv\left\{y_{i}^{*} ; i=1,2, \ldots, n\right\}$, where

$$
\begin{equation*}
y_{i}^{*}=\hat{y}_{i}+a_{i} w_{i}^{-1} w_{i}^{*}\left(y_{i}-\hat{y}_{i}\right) \tag{24}
\end{equation*}
$$

to get a consistent variance estimator.
To illustrate that the set $\mathbf{Y}^{*}$ can be used to approximate the variance estimator, assume that the full sample variance estimator of $\bar{y}$ can be written as

$$
\hat{V}=\sum_{i=1}^{L} c_{i}\left(\bar{y}^{(i)}-\bar{y}\right)^{2}
$$

where $L$ is the number of replications, $c_{i}$ is the $i$-th replication factor, and $\bar{y}^{(i)}=\Sigma_{j=1}^{n} w_{j} M_{j}^{(i)} y_{j}$ is the $i$-th replicate of $\bar{y}$. The term $M_{j}^{(i)}$ is the replication multiplier applied to the weight of unit $j$ at the $i$-th replication. For example, under simple random sampling, the jackknife multiplier is

$$
M_{j}^{(i)}= \begin{cases}(n-1)^{-1} n & \text { if } i \neq j \\ 0 & \text { if } i=j .\end{cases}
$$

Assume that the replicate variance estimator $\hat{V}$ is applied to the set $\mathbf{Y}^{*}$ to get

$$
\hat{V}^{*}=\sum_{i=1}^{L} c_{i}\left(\bar{y}_{I}^{*(i)}-\bar{y}_{I}\right)^{2}
$$

where $\bar{y}_{j}^{*(i)}=\Sigma_{j=1}^{n} w_{j} M_{j}^{(i)} y_{j}^{*}$ with $y_{j}^{*}$ being defined in (24). Then, we have

$$
\begin{equation*}
\bar{y}_{I}^{*(i)}-\bar{y}_{I}=\bar{y}_{l l}^{*(i)}-\bar{y}_{l l}+\left(\overline{\mathbf{x}}_{1}^{(i)}-\overline{\mathbf{x}}_{2}^{(i)}-\overline{\mathbf{x}}_{1}+\overline{\mathbf{x}}_{2}\right)(\hat{\beta}-\boldsymbol{\beta}) \tag{25}
\end{equation*}
$$

where

$$
\left(\overline{\mathbf{x}}_{1}^{(i)}, \overline{\mathbf{x}}_{2}^{(i)}\right)=\sum_{j=1}^{n} w_{j} M_{j}^{(i)}\left(\mathbf{x}_{j}, a_{j} w_{j}^{-1} w_{j}^{*} \mathbf{x}_{j}\right) .
$$

It is shown in Appendix C that

$$
\begin{equation*}
\hat{V}^{*}=\sum_{i=1}^{L} c_{i}\left(\bar{y}_{I l}^{*(i)}-\bar{y}_{I l}\right)^{2}+o_{p}\left(n^{-1}\right) . \tag{26}
\end{equation*}
$$

Therefore, the standard jackknife variance estimator applied to the pseudo data set $\mathbf{Y}^{*}$ can be used to approximate the
standard jackknife variance estimator applied to the pseudo data set $\tilde{\mathbf{Y}}^{*}$.

### 4.2 Random Imputation

The arguments for variance estimation with random imputation are quite similar to those for deterministic imputation described in the previous subsection. First, define the imputation indicator function

$$
d_{i j}= \begin{cases}1 & \text { if unit } i \text { is used as donor for unit } j  \tag{27}\\ 0 & \text { otherwise. }\end{cases}
$$

Then, the estimator of the mean of $y$ using random imputation is

$$
\begin{equation*}
\bar{y}_{I}=\sum_{i=1}^{n} w_{i} y_{i}^{*} \tag{28}
\end{equation*}
$$

where

$$
\begin{equation*}
\bar{y}_{i}^{*}=\hat{y}_{i}+a_{i}\left(1+d_{i}\right)\left(y_{i}-\hat{y}_{i}\right) \tag{29}
\end{equation*}
$$

and

$$
\begin{equation*}
d_{i}=\sum_{j=1}^{n}\left(1-a_{j}\right) d_{i j} w_{i}^{-1} w_{j} . \tag{30}
\end{equation*}
$$

If the original sample weights are the same, then $d_{i}$ is the number of times that unit $i$ is used as a donor. We assume that

$$
\begin{equation*}
E\left[a_{i}\left(1+d_{i}\right) \mid F_{1}\right]=1 \tag{31}
\end{equation*}
$$

where $F_{1}=\left\{\left(i, \mathbf{x}_{i}, y_{i}\right) ; i=1,2, \ldots, n\right\}$. The expectation in (31) is with respect to the joint distribution of the response mechanism and the imputation mechanism. Then, we have

$$
E\left(\bar{y}_{1} \mid F_{1}\right) \doteq \bar{y} .
$$

If we assume equal response probability, then, by (31), the probability of selection of donors should be proportional to the weights. This is the Rao and Shao (1992) setup for random imputation.

Now, let

$$
\begin{equation*}
\bar{y}_{l l}=\sum_{i=1}^{n} w_{i}\left[\tilde{y}_{i}+a_{i}\left(1+d_{i}\right)\left(y_{i}-\tilde{y}_{i}\right)\right] \tag{32}
\end{equation*}
$$

where $\tilde{y}_{i}=\mathbf{x}_{i} \boldsymbol{\beta}$. Then, we also have $\bar{y}_{I}=$ $\left(\overline{\mathbf{x}}_{d}-\overline{\mathbf{x}}_{1}\right)(\hat{\beta}-\beta) \bar{y}_{11}+$ where $\overline{\mathbf{x}}_{d}=\sum_{i=1}^{n} w_{i} a_{i}\left(1+d_{i}\right) \mathbf{x}_{i}$. By the assumption (31), we have $E\left(\overline{\mathbf{x}}_{d}-\overline{\mathbf{x}}_{1} \mid F_{1}\right)=0$. Under mild conditions, $\overline{\mathbf{x}}_{d}-\overline{\mathbf{x}}_{1}=O_{p}\left(n^{-1 / 2}\right)$ and $\bar{y}_{I}=\bar{y}_{l l}+O_{p}\left(n^{-1}\right)$. Now,

$$
V\left(\bar{y}_{I}-\bar{Y}_{N}\right)=V\left[E\left(\bar{y}_{I}-\bar{Y}_{N} \mid \mathbf{a}, \mathbf{d}\right)\right]+E\left[V\left(\bar{y}_{I}-\bar{Y}_{N} \mid \mathbf{a}, \mathbf{d}\right)\right]
$$

where $\mathbf{d}=\left(d_{1}, d_{2}, \ldots, d_{N}\right)$. Conditional on a and $\mathbf{d}$, the estimator $\bar{y}_{I I}$ is a linear estimator. Hence, the pseudo data

$$
\begin{equation*}
y_{i}^{*}=\hat{y}_{i}+a_{i}\left(1+d_{i}\right)\left(y_{i}-\hat{y}_{i}\right) \tag{33}
\end{equation*}
$$

can be used to estimate the variance of $\bar{y}_{I}$.

## 5. COMPARISONS WITH ADJUSTED JACKKNIFE METHOD

Rao and Sitter (1995) proposed an adjusted jackknife variance estimator for the ratio imputation problem. Under the setup described in section 4 , the ratio imputed estimator of $\mu_{y}$ is

$$
\hat{\mu}_{I}=\sum_{i=1}^{n} w_{i}\left[a_{i} y_{i}+\left(1-a_{i}\right) \hat{y}_{i}\right]
$$

with $\hat{y}_{i}=x_{i} \hat{R}$ and $\hat{R}=\left(\sum_{i=1}^{n} w_{i} a_{i} x_{i}\right)^{-1} \Sigma_{i=1}^{n} w_{i} a_{i} y_{i}$. The Rao and Sitter (1995) variance estimator is

$$
\begin{equation*}
V_{a}=\sum_{i=1}^{L} c_{i}\left(\hat{\mu}_{I}^{(i)}-\hat{\mu}_{I}\right)^{2} \tag{34}
\end{equation*}
$$

where the adjusted jackknife replicate at the $i$-th replication is

$$
\begin{equation*}
\hat{\mu}_{I}^{(i)}=\sum_{j=1}^{n} w_{j} M_{j}^{(i)} y_{j}^{*(i)} \tag{35}
\end{equation*}
$$

where

$$
y_{j}^{*(i)}= \begin{cases}x_{j} \hat{R}^{(i)} & \text { if } a_{i}=1  \tag{36}\\ x_{j} \hat{R} & \text { if } a_{i}=0\end{cases}
$$

with $\quad \hat{R}^{(i)}=\left(\Sigma_{j=1}^{n} w_{j} M_{j}^{(i)} a_{j} x_{j}\right)^{-1} \Sigma_{j=1}^{n} w_{j} M_{j}^{(i)} a_{j} y_{j}$. The adjusted values (36) in the Rao and Sitter (1995) method can also be regarded as pseudo data for variance estimation. Note that the calculation of the pseudo data (36) requires recalculation of $\hat{R}^{(i)}$ for each $i$ with $a_{i}=1$.

We modify the calculation of the pseudo values $y_{i}^{*}$ in (5) to

$$
y_{i}^{*}= \begin{cases}\hat{y}_{i} & \text { if } a_{i}=0  \tag{37}\\ \hat{y}_{i}+c_{r}\left(\frac{\bar{x}_{1}}{\bar{x}_{2}}\right)\left(y_{i}-\hat{y}_{i}\right) & \text { if } a_{i}=1,\end{cases}
$$

where $\bar{x}_{2}=\sum_{i=1}^{n} w_{i} r^{-1} n a_{i} x_{i}, \bar{x}_{1}=n^{-1} \sum_{i=1}^{n} w_{i} x_{i}$ and $c_{r} \doteq r^{-1} n$. The term ( $\bar{x}_{1} / \bar{x}_{2}$ ) is inserted to improve the conditional properties of $V_{J}$ given the first phase sample. The resulting variance estimator is approximately equivalent to the adjusted jackknife variance estimator (34). To see this, note that the adjusted values (35) can be written in the form

$$
\hat{\mu}_{I}^{(i)}=\left(\sum_{j=1}^{n} w_{j} M_{j}^{(i)} x_{j}\right) \frac{\sum_{j=1}^{n} w_{j} M_{j}^{(i)} a_{j} y_{j}}{\sum_{j=1}^{n} w_{j} M_{j}^{(i)} a_{j} x_{j}}=: \hat{Z}^{(i)} \frac{\hat{S}^{(i)}}{\hat{T}^{(i)}},
$$

where $A=: B$ denotes that we define $B$ to be $A$. Also, define $\hat{Z}=\sum_{j=1}^{n} w_{j} x_{j}, \hat{S}=\sum_{j=1}^{n} w_{j} a_{j} y_{j}$, and $\hat{T}=\sum_{j=1}^{n} w_{j} a_{j} x_{j}$.

Then by the first order Taylor expansion,

$$
\begin{align*}
\hat{Z}^{(i)} \frac{\hat{S}^{(i)}}{\hat{T}^{(i)}}= & \hat{Z} \frac{\hat{S}}{\hat{T}}+\left(\hat{Z}^{(i)}-\hat{Z}\right) \frac{\hat{S}}{\hat{T}} \\
& +\left(\hat{S}^{(i)}-\hat{S}\right) \frac{\hat{Z}}{\hat{T}}-\left(\hat{T}^{(i)}-\hat{T}\right) \frac{\hat{Z} \hat{S}}{\hat{T}^{2}} \\
= & {\left[\hat{Z}^{(i)} \frac{\hat{S}}{\hat{T}}+\frac{\hat{Z}}{\hat{T}}\left(\hat{S}^{(i)}-\hat{T}^{(i)} \frac{\hat{S}}{\hat{T}}\right)\right] } \tag{38}
\end{align*}
$$

Note that the right side of (38) is exactly equal to

$$
\sum_{j=1}^{n} w_{j} M^{(i)}\left[\frac{\hat{S}}{\hat{T}}+\frac{\hat{\underline{L}}}{\hat{T}} a_{j}\left(y_{j}-\frac{\hat{S}}{\hat{T}}\right)\right] .
$$

Thus, the pseudo data for variance estimation can be written as

$$
y_{i}^{*}=\frac{\hat{S}}{\hat{T}}+\frac{\hat{\hat{z}}}{\hat{T}} a_{i}\left(y_{i}-\frac{\hat{S}}{\hat{T}}\right),
$$

which reduces to (37). Hence, the proposed method is exactly a first order Taylor linearization of the Rao and Sitter method in the case of ratio imputation. Therefore, we can expect our proposed method to have the same asymptotic properties as the Rao and Sitter method up to the order of $n^{-1}$.

The variance estimation method using the pseudo data set calculated by (37) is easy to implement because we can directly use existing software, which is more difficult with the Rao and Shao (1992) or Rao and Sitter (1995) method. Furthermore, if we calculate the pseudo data by (13), then the data set works for without-replacement hot deck imputation as well as for with-replacement hot deck imputation.

## 6. A SIMULATION STUDY

The preceding theory was tested in a simulation study using an artificial, finite population, from which repeated samples were drawn. The population has $L=32$ strata, $N_{h}$ clusters in stratum $h$, and 20 ultimate units in each cluster. The values of the population parameters were chosen to correspond to real populations encountered in the U.S. National Assessment of Educational Progress Study (Hansen and Tepping 1985) and are listed in Table 1. The finite population units are

$$
y_{h i j}=y_{h i}+e_{h i j},
$$

where

$$
y_{h i} \stackrel{\mathrm{iid}}{\sim} N\left(\mu_{h}, \sigma_{h}^{2}\right), h=1,2, \ldots, L, i=1,2, \ldots, N_{h},
$$

and

$$
e_{h i j} \stackrel{\mathrm{iid}}{\sim} N\left(0, \frac{1-\rho}{\rho} \sigma_{h}^{2}\right), j=1,2, \ldots, 20 .
$$

Shao, Chen and Chen (1998) also used the same population in their simulation study. The value of the intra-cluster correlation $\rho$ considered in the simulation is $\rho=0.3$. Simulations with other values of $\rho$ produced similar results and are not listed here for brevity.

Table 1
Parameters of the Finite Population for Simulation

| $h$ | $N_{h}$ | $\mu_{h}$ | $\sigma_{h}$ | $h$ | $N_{h}$ | $\mu_{h}$ | $\sigma_{h}$ |
| :---: | :---: | ---: | :---: | :---: | :---: | :---: | :---: |
| 1 | 13 | 100.0 | 20.0 | 2 | 16 | 95.0 | 19.0 |
| 3 | 20 | 90.0 | 18.0 | 4 | 25 | 98.0 | 19.6 |
| 5 | 25 | 93.0 | 18.6 | 6 | 25 | 98.0 | 19.6 |
| 7 | 25 | 96.0 | 19.2 | 8 | 28 | 94.0 | 18.8 |
| 9 | 28 | 92.0 | 18.4 | 10 | 28 | 96.0 | 19.2 |
| 11 | 31 | 94.0 | 18.8 | 12 | 31 | 9.0 | 18.4 |
| 13 | 31 | 90.0 | 18.0 | 14 | 31 | 96.0 | 19.2 |
| 15 | 31 | 94.0 | 18.8 | 16 | 31 | 92.0 | 18.4 |
| 17 | 31 | 90.0 | 18.0 | 18 | 31 | 88.0 | 17.6 |
| 19 | 31 | 86.0 | 17.2 | 20 | 34 | 84.0 | 16.8 |
| 21 | 34 | 82.0 | 16.4 | 22 | 34 | 80.0 | 16.0 |
| 23 | 34 | 90.0 | 18.0 | 24 | 37 | 85.0 | 17.0 |
| 25 | 37 | 80.0 | 16.0 | 26 | 37 | 90.0 | 18.0 |
| 27 | 37 | 85.0 | 17.0 | 28 | 39 | 80.0 | 16.0 |
| 29 | 39 | 75.0 | 15.0 | 30 | 42 | 75.0 | 15.0 |
| 31 | 42 | 75.0 | 15.0 | 32 | 42 | 75.0 | 15.0 |

We consider a stratified cluster sampling design, where $n_{h}=2$ clusters are selected with replacement from stratum $h$ with equal probability and all of the ultimate units in the selected clusters are in the sample. The sampling fraction is $6.4 \%$. For each sampled unit $y_{h i j}$, a response indicator variable $a_{h i j}$ is generated from

$$
a_{h j} \stackrel{\text { iid }}{\sim} \text { Bemoulli }(p) \text {, }
$$

and that $a_{h i j}$ is independent of $y_{h i j}$. The value of $p$ considered in the simulation are $p=0.9,0.8,0.7,0.6$, and 0.5 .

A set of 5,000 samples were selected using the same sampling design. In each of the selected samples, three imputation methods are considered;
[M1] With-replacement weighted hot deck imputation considered by Rao and Shao (1992), where a missing value is imputed by a value randomly selected from the respondents with replacement with probability proportional to the survey weights.
[M2] . Without-replacement weighted hot deck imputation, which is the same as [M1] expect that the selection was performed using a withoutreplacement sample. The without-replacement selection of donors is carried out systematically using the method described by Hansen, Hurwitz, and Madow (1953, page 343) from the respondents sorted by random order.
[M3] Overall mean imputation, where the weighted mean of the respondents in the sample is imputed.

Hence, all the imputation methods use a single imputation cell that collapses all the strata.

In each imputed data set we computed three variance estimators $\hat{V}_{n}$, naive variance estimator treating the imputed data as if it were observed data, $\hat{V}_{a}$, the adjusted jackknife variance estimator of Rao and Shao (1992) for [M1] and [M2] and of Rao and Sitter (1995) for [M3], and $\hat{V}^{*}$, the jackknife variance estimator based on the pseudo data. The pseudo data set is constructed by (29) for [M1] and [M2] and by (24) for [M3]. The complete sample variance estimator used a standard jackknife for stratified cluster sampling, in which a cluster is deleted for each replication. Note that the standard jackknife is a consistent estimator of the variance under the model with nonzero intracluster correlation. Thus, the standard jackknife method based on the pseudo data can be applicable to the data set considered. The point estimators of the population mean are unbiased under the three different imputation schemes and are not listed here.

Table 2 presents the relative bias of the three variance estimators, the standard error of the relative bias of the variance estimators, and the sample correlation coefficient between the Rao's adjusted jackknife variance estimator and the new variance estimator based on the 5,000 samples. The relative bias of $\hat{V}$ as an estimator of the variance of $\bar{y}_{I}$ is calculated by $\left[\operatorname{Var}_{\mathrm{B}}\left(\bar{y}_{I}\right)\right]^{-1}\left[E_{B}(\hat{V})-\operatorname{Var}_{\mathrm{B}}\left(\bar{y}_{I}\right)\right]$, where the subscript $B$ denotes the distribution generated by the Monte Carlo simulation. The correlation coefficients of the two variance estimators are computed to give a measure the relative linearity behavior of the two variance estimators.

Table 2
Relative Bias of the Variance Estimator, Standard Error of the Relative Bias, and Sample Correlation Coefficient Between the Rao's Variance Estimator and the New Variance Estimator Based on 5,000 Samples

| Response <br> Rate $(p)$ | Imputation <br> Method | Rel. Bias $\times 100($ S.E. $\times 100)$ |  | Corr. |  |
| :---: | :---: | :---: | :---: | :---: | :---: |
| Naive | Rao | New | Coeff. $r$ |  |  |
| 0.9 | M1 | $-17.40(2.02)$ | $1.61(2.03)$ | $1.70(2.04)$ | 0.967 |
|  | M2 | $-17.50(2.00)$ | $1.41(2.01)$ | $0.81(2.03)$ | 0.974 |
|  | M3 | $-18.03(2.03)$ | $1.16(2.05)$ | $1.15(2.04)$ | 1.000 |
|  | M1 | $-34.45(2.01)$ | $0.65(2.03)$ | $0.49(2.05)$ | 0.939 |
| 0.8 | M2 | $-32.89(2.01)$ | $2.49(2.04)$ | $0.19(2.03)$ | 0.947 |
|  | M3 | $-34.96(2.01)$ | $1.59(2.03)$ | $1.59(2.03)$ | 1.000 |
|  | M1 | $-48.96(2.01)$ | $0.21(1.99)$ | $0.41(2.04)$ | 0.912 |
|  | M2 | $-44.76(2.02)$ | $5.31(2.05)$ | $0.76(2.05)$ | 0.920 |
| 0.7 | M3 | $-50.21(2.02)$ | $1.53(2.05)$ | $1.52(2.04)$ | 1.000 |
|  |  |  |  |  |  |
|  | M1 | $-59.80(2.02)$ | $1.58(2.05)$ | $1.27(2.06)$ | 0.892 |
| 0.6 | M2 | $-54.86(2.03)$ | $7.10(2.07)$ | $-0.75(2.07)$ | 0.899 |
|  | M3 | $-64.11(2.00)$ | $-0.35(2.04)$ | $-0.35(2.01)$ | 1.000 |
|  | M1 | $-69.75(1.99)$ | $0.84(2.03)$ | $1.12(2.03)$ | 0.873 |
| 0.5 | M2 | $-59.90(2.01)$ | $15.07(2.07)$ | $2.27(2.06)$ | 0.872 |
|  | M3 | $-74.44(1.97)$ | $1.99(2.00)$ | $1.98(2.00)$ | 1.000 |
|  |  |  |  |  |  |

Table 2 supports our theory in the following ways.

1. As is well known, the naive variance estimator seriously underestimates the true variance. The adjusted jackknife variance estimator performs well for [M1] and [M3], but not for [M2]. The theory for the adjusted jackknife method assumes that hot deck imputations are done using the with-replacement selection which is not used in [M2]. As the response rate decreases in Table 2, the relative bias of the adjusted jackknife becomes larger.
2. The new method based on the pseudo data performs well even for the without-replacement imputation [M2]. As was discussed at the end of section 3, a single formula (29) can be used as the pseudo data for a large class of imputation methods.
3. As is observed in the correlation coefficients, the behaviors of the adjusted jackknife variance estimator and the proposed variance estimator are very similar for mean imputation [M3]. This is because the two variance estimators are asymptotically equivalent, as discussed in section 5.

## 7. CONCLUDING REMARKS

We have described methods of making pseudo data to be used for variance estimation. Generally speaking, the pseudo data can be described as

$$
y_{i}^{*}= \begin{cases}\hat{y}_{i} & i=r+1, r+2, \ldots, n  \tag{39}\\ \hat{y}_{i}+c_{i} g_{i}\left(y_{i}-\hat{y}_{i}\right) & i=1,2, \ldots, r,\end{cases}
$$

where $\hat{y}_{i}$ is the predicted value of $y_{i}$ under the model used for imputation. If $c_{i} g_{i}=1$, then the variance estimator treats the imputed values as observations. A suitable choice of $c_{i} g_{i}>1$ leads to a consistent variance estimator. If the imputation method is deterministic and the respondents are regarded as a random sample from the original sample, then $c_{i} \doteq r^{-1} n>1$. For a two-phase sampling with a complex design, $c_{i}=w_{i}^{-1} w_{i}^{*}$, where $w_{i}$ is the sampling weight of the unit $i$ for the first-phase sample and $w_{i}^{*}$ is the sampling weight of the unit $i$ for the second-phase sample.

The $g_{i}$ in (39) is the adjustement made to improve the conditional properties given the auxiliary variable $x$. For ratio imputation,

$$
g_{i}=\left(\bar{x}_{2}\right)^{-1} \bar{x}_{1}
$$

where $\bar{x}_{2}=\sum_{i=1}^{r} w_{i}^{*} x_{i}$ and $\bar{x}_{1}=\sum_{i=1}^{n} w_{i} x_{i}$. For regression imputation with scalar $x$,

$$
g_{i}=1+\left(\bar{x}_{1}-\bar{x}_{2}\right)\left\{\sum_{k=1}^{r} w_{k}^{*}\left(x_{k}-\bar{x}_{2}\right)^{2}\right\}^{-1}\left(x_{i}-\bar{x}_{2}\right) .
$$

In either case, we have

$$
\sum_{i=1}^{r} w_{i}^{*} g_{i} x_{i}=\bar{x}_{1} .
$$

While this paper was under review, Shao and Steel (1999) also provided similar methods in the case of deterministic imputation. Our method is more general in the sense that we also considered random imputation and introduced $c_{i}$ term to improve finite sample properties.

## ACKNOWLEDGEMENTS

The author thanks his thesis adviser Wayne A. Fuller for valuable discussions. The author also thanks Pamela Abbitt, F. Jay Breidt, Lou Rizzo, Richard Valliant, and the referees for helpful comments, which greatly improved the paper. Most of this work was done while the author was a graduate student at Iowa State University and was funded in part by cooperative agreement 68-3A75-43 between the USDA Natural Resources Conservation Service and Iowa State University and by Cooperative Agreement 43-3AEU-380088 between Iowa State University, the National Agricultural Statistics Service and the U.S. Bureau of Census.

## APPENDIX

## A. Proof of Equation (10) and (12)

The estimator $\hat{\mu}_{y}$ in (9) can be written as

$$
\begin{equation*}
\hat{\mu}_{y}=n^{-1} \sum_{i=1}^{n} \hat{y}_{i}+n^{-1} \sum_{i=1}^{r}\left(1+d_{i}\right) \hat{e}_{i} \tag{A.1}
\end{equation*}
$$

where $d_{i}$ is the number of times that unit $i$ is used as a donor. Under the equal probability and with-replacement imputation mechanism, we have

$$
E_{l}\left(d_{i}\right)=r^{-1} m
$$

and

$$
\operatorname{Cov}_{I}\left(d_{i}, d_{j}\right)= \begin{cases}r^{-1} m\left(1-r^{-1}\right) & \text { if } i=j \\ -r^{-2} m & \text { if } i \neq j\end{cases}
$$

where the subscript $I$ denotes the variation due to the imputation mechanism. It follows that $E_{I}\left(\hat{\mu}_{y}\right)=n^{-1} \sum_{i=1}^{n} \hat{y}_{i}$ and $V_{I}\left(\hat{\mu}_{y}\right)=n^{-2} r^{-1} m \Sigma_{i=1}^{r} \hat{e}_{i}^{2}$. Hence,

$$
\begin{equation*}
V\left(\hat{\mu}_{y}\right) \doteq V\left(n^{-1} \sum_{i=1}^{n} \hat{y}_{i}\right)+E\left(n^{-2} r^{-1} m \sum_{i=1}^{r} \hat{e}_{i}^{2} \cdot\right) \tag{A.2}
\end{equation*}
$$

Now, by an similar argument similar to the one leading to (2), we have

$$
\begin{equation*}
\operatorname{Var}\left(n^{-1} \sum_{i=1}^{n} \hat{y}_{i}\right)=\left[n^{-1} R^{2}+r^{-1}\left(1-R^{2}\right)\right] \sigma_{y}^{2} \tag{A.3}
\end{equation*}
$$

Since $\hat{y}_{i}-\bar{y}_{f}=\left(\mathbf{x}_{i}-\overline{\mathbf{x}}_{i}\right) \beta+o_{p}(1)$, we apply classical regression theory to get

$$
\begin{equation*}
E\left[(r-p)^{-1} \sum_{i=1}^{r} \hat{e}_{i}^{2}\right]=\left(1-R^{2}\right) \sigma_{y}^{2}, \tag{A.4}
\end{equation*}
$$

and

$$
\begin{equation*}
E\left[(n-1)^{-1} \sum_{i=1}^{n}\left(\hat{y}_{i}-\bar{y}_{I}\right)^{2}\right]=R^{2} \sigma_{y}^{2} . \tag{A.5}
\end{equation*}
$$

Therefore, (10) is proved and the estimator in (12) is consistent for the variance in (10).

## B. Validity of (15) Under the Without-Replacement Imputation Mechanism

We assume that $m=k r+t$ where $k$ and $t$ are nonnegative integers and $t<r$. Let the estimator of the mean of $y$ have the form (A.1). Let the imputation be performed such that $t$ of the respondents are used $k+1$ times for imputation and $r-t$ units are used $k$ times for imputation. The $t$ of the respondents that are used $k+1$ times are chosen by simple random sampling without replacement. Then,

$$
E_{I}\left(d_{i}\right)=k+r^{-1} t=r^{-1} m
$$

and

$$
\operatorname{Cov}_{I}\left(d_{i}, d_{j}\right)= \begin{cases}r^{-1} t\left(1-r^{-1} t\right) & \text { if } i=j \\ -r^{-2} t & \text { if } i \neq j .\end{cases}
$$

So, by similar arguments as in the proof of (A.2), we have

$$
\begin{equation*}
V\left(\hat{\mu}_{y}\right) \doteq V\left(\bar{y}_{I}\right)+E\left(n^{-2} r^{-1} t \sum_{i=1}^{\dot{r}} \hat{e}_{i}^{2}\right) . \tag{B.1}
\end{equation*}
$$

Hence, using (A.3) and (A.4), we have

$$
\begin{equation*}
V\left\{\hat{\mu}_{y}\right\}=\left[n^{-1} R^{2}+\left(r^{-1}+n^{-2} t\right)\left(1-R^{2}\right)\right] \sigma_{y}^{2} . \tag{B.2}
\end{equation*}
$$

Now, conditional on the realized sample and the respondents, we have

$$
E_{I}\left\{\left(1+d_{i}\right)^{2}\right\}=\left(\frac{n}{r}\right)^{2}+\frac{t}{r}\left(1-\frac{t}{r}\right)
$$

so that $\hat{V}\left\{\mu_{y}\right\}$ in (15) satisfies

$$
\begin{aligned}
E_{l}\left(\hat{V}\left\{\mu_{y}\right\}\right) \doteq & n^{-1}(n-1)^{-1} \sum_{i=1}^{n}\left(\hat{y}_{i}-\bar{y}_{l}\right)^{2} \\
+ & {\left[r^{-1}+n^{-2} t\left(1-r^{-1} t\right)\right] } \\
& (r-p)^{-1} \sum_{i=1}^{r}\left(y_{i}-\hat{y}_{i}\right)^{2} .
\end{aligned}
$$

Therefore, using (A.4) and (A.5), we have the approximate unbiasedness of the $\hat{V}\left\{\mu_{y}\right\}$ under the without-replacement imputation mechanism.

## C. Proof of Equation (26)

First, define $R_{n}^{(i)}=\left(\mathbf{x}_{1}^{(i)}-\bar{x}_{2}^{(i)}\right)(\hat{\beta}-\beta)$ and $R_{n}=$ $\left(\overline{\mathbf{x}}_{1}-\overline{\mathbf{x}}_{2}\right)(\hat{\beta}-\boldsymbol{\beta})$. From the equality (25),

$$
\hat{V}^{*}=\sum_{i=1}^{L} c_{i}\left(\bar{y}_{I}^{*(i)}-\bar{y}_{I}\right)^{2}=A_{n}+B_{n}+2 C_{n}
$$

where $A_{n}=\sum_{i=1}^{L} c_{i}\left(\bar{y}_{n i}^{*(i)}-\bar{y}_{n}\right)^{2}, \quad B_{n}=\sum_{i=1}^{L} c_{i}\left(R_{n}^{(i)}-R_{n}\right)^{2}$, and $C_{n}=\Sigma_{i=1}^{L} c_{i}\left(\bar{y}_{n}^{*(i)}-\bar{y}_{H l}\right)\left(R_{n}^{(i)}-R_{n}\right)$. Hence, by the assumption (20), (26) follows because $A_{n}=O_{p}\left(n^{-1}\right)$, $B_{n}=o_{p}\left(n^{-1}\right)$, and $C_{n}=o_{p}\left(n^{-1}\right)$. The last property comes $B_{n}{ }_{p}^{p}{ }^{n}$ Cauchy-Schwartz inequality, $C_{n}^{2} \leq A_{n} B_{n}$.

## REFERENCES

COCHRAN, W.G. (1977). Sampling Techniques. New York: John Wiley and Sons.
FAY, R.E. (1991). A design-based perspective on missing data variance. Proceedings of the Bureau of the Census Annual Research conference, 429-440.
FAY, R.E. (1992). When are inferences from multiple imputation valid? Proceedings of the Section on Survey Research Methods, American Statistical Association, 227-232.
FAY, R.E. (1996). Alternative paradigms for the analysis of imputed survey data. Journal of the American Statistical Association, 91, 490-498.
FULLER, W.A. (1998). Replication variance estimation for twophase samples. Statistica Sinica, 8, 1153-1164.
HANSEN, M., HURWITZ, W.N. and MADOWS, W.G. (1953). Sample Survey Methods and Theory, Vol. I, New York: John Wiley and Sons.
HANSEN, M., and TEPPING, B.J. (1985). Estimation for Variance in NAEP. Unpublished memorandum, Westat, Washington, D.C.
RAO, J.N.K. (1996). On variance estimation with imputed survey data. Journal of the American Statistical Association, 91, 499506.

RAO, J.N.K., and SHAO, J. (1992). Jackknife variance estimation with survey data under hot deck imputation. Biometrika, 79, 811822.

RAO, J.N.K., and SITTER, R.R. (1995). Variance estimation under two-phase sampling with application to imputation for missing data. Biometrika, 82, 453-460.
RUBIN, D.B. (1987). Multiple Imputation for Nonresponse in Surveys. New York: John Wiley and Sons.
RUBIN, D.B. (1996). Multiple imputation after 18+ years. Journal of the American Statistical Association, 91, 473-489.
RUBIN, D.B., and SCHENKER, N. (1986). Multiple imputation for interval estimation from simple random samples with ignorable nonresponse. Journal of the American Statistical Association, 81, 366-374.
SÄRNDAL, C.-E. (1992). Methods for estimating the precision when imputation has been used. Survey Methodology, 18, 241-252.

SÄRNDAL, C.-E., and SWENSSON, B. (1987). A general view of estimation for two phases of selection with applications to twophase sampling and nonresponse. International Statistical Review, 55, 279-294.

SCHAFER, J.L. (1997). Analysis of Incomplete Multivariate Data. Chapman \& Hall.

SHAO, J., CHEN, Y. and CHEN, Y. (1998). Balanced repeated replication for stratified multistage survey data under imputation. Journal of the American Statistical Association, 93, 819-831.

SHAO, J., and STEEL, P. (1999). Variance estimation for survey data with composite imputation and nonnegligible sampling fraction. Journal of the American Statistical Association, 94, 254-265.

# A Multivariate Technique for Multiply Imputing Missing Values Using a Sequence of Regression Models 

TRIVELLORE E. RAGHUNATHAN, JAMES M. LEPKOWSKI, JOHN VAN HOEWYK and PETER SOLENBERGER ${ }^{1}$


#### Abstract

This article describes and evaluates a procedure for imputing missing values for a relatively complex data structure when the data are missing at random. The imputations are obtained by fitting a sequence of regression models and drawing values from the corresponding predictive distributions. The types of regression models used are linear, logistic, Poisson, generalized logit or a mixture of these depending on the type of variable being imputed. Two additional common features in the imputation process are incorporated: restriction to a relevant subpopulation for some variables and logical bounds or constraints for the imputed values. The restrictions involve subsetting the sample individuals that satisfy certain criteria while fitting the regression models. The bounds involve drawing values from a truncated predictive distribution. The development of this method was partly motivated by the analysis of two data sets which are used as illustrations. The sequential regression procedure is applied to perform multiple imputation analysis for the two applied problems. The sampling properties of inferences from multiply imputed data sets created using the sequential regression method are evaluated through simulated data sets.


KEY WORDS: Item nonresponse; Missing at random; Multiple imputation; Nonignorable missing mechanism; Regression; Sampling properties and simulations.

## 1. INTRODUCTION

Incomplete data is a pervasive problem faced by most applied researchers. Several methods have been, and continue to be, developed to draw inferences from data sets with missing values (Little and Rubin 1987). The multiple imputation framework suggested by Rubin (1978, 1987a, 1996) is an attractive option if a data set is to be used by multiple researchers with differing levels of statistical expertise. This approach involves imputing several plausible sets of missing values in the incomplete data set resulting in several completed data sets. Each completed data set is analyzed separately, say by fitting a particular regression model. The resulting inferences - point estimates and the covariance matrices - are then combined using the formula given in Rubin (1987a, Chap. 3) and refinements thereof (Li, Raghunathan and Rubin 1991; Li, Meng, Raghunathan and Rubin 1991; Meng and Rubin 1992; and Barnard 1995).

Imputation based approaches for handling missing data, in general, are quite useful in practice because once the missing values have been imputed, existing complete-data software can be used to analyze the data. Since software development for complete data analysis is keeping pace with the introduction of new statistical methods, applied researchers without knowledge of particular missing data techniques or resources to generate their own code for implementing new missing data procedures will be able to fit finely tuned substantive models for a specific problem at
hand. An added advantage of the multiple imputation approach is that by repeatedly applying the complete data software, one can obtain valid point and interval estimates under a fairly general set of conditions (Rubin 1987a). Several researchers (see, for example, the list of references in Rubin 1996) have applied this technique under a variety of settings and have demonstrated, through analysis of simulated and actual data sets, the appropriateness of this approach. Alternatives such as single imputation with an appropriate variance estimation procedure, for example, modified Jackknife Repeated Replication Technique (Rao and Shao 1992) also have this advantage. The imputation approach described in this paper can also be used to create single imputation with an alternative variance estimation procedure.

The development of imputation methods from varying perspectives has a long history (Madow, Nisselson, Olkin and Rubin 1983). A theoretically appealing framework for developing imputation methods is the Bayesian approach. This approach specifies an explicit model for variables with missing values, conditional on the fully observed variables and some unknown parameters, a prior distribution for the unknown parameters, and a model for the missing data mechanism, which does not need to be specified under an ignorable missing data mechanism (Rubin 1976). This explicit model then generates a posterior predictive distribution of the missing values conditional on the observed values. The imputations are draws from this posterior predictive distribution. Several computer programs and

[^6]algorithms are available for imputing missing values under multivariate normality (Rubin and Schafer 1990), the multivariate $t$ distribution (Liu 1995), and several variations of the general location model (Schafer 1997; Raghunathan and Grizzle 1995; and Raghunathan and Siscovick 1996). The latter model can handle the joint distribution of categorical and continuous variables and was first proposed by Olkin and Tate (1961), and used by Little and Schluchter (1985) explicitly for missing data problems. An important property of these approaches is that they are fully conditional on all the observed information. Several simulation studies (for example, Raghunathan and Grizzle 1995) indicate that the inferences drawn from such imputed data have desirable sampling properties.

Survey data sets often consist of large numbers of variables which have a variety of distributional forms. Typically, such data sets have hundreds of variables, some continuous, others counts, many dichotomous or polytomous, and even some semi-continuous or limited dependent variables. Moreover, the distributions of the continuous variables alone may involve normal, lognormal, and other distributions. Postulating a full Bayesian model can be very difficult in this situation. Furthermore, survey data commonly have two additional features that make the modeling process even more complex. First, certain restrictions are imperative. For example, the variable "Number of Years Since Quit Smoking" is defined only for former smokers; hence, the imputation process for this variable should be restricted only to former smokers. Restrictions also arise due to skip pattems in the questionnaire. For example, certain questions about income from a second job are asked only when the respondent indicates that he/she has a second job. The imputation of such variables has to be handled in a hierarchical manner.

Second, there are certain logical or consistency bounds for the missing values that must be incorporated in the imputation process. Such interrelationships among the variables make the model specification difficult. For instance, "Years of Smoking" is restricted to current or past smokers and the imputed values must be less than Age - $\boldsymbol{x}$ years, where $x$ may be chosen based on certain other characteristics, such as evidence of smoking as a teen-ager. For a former smoker, $x$ also includes years since smoking ceased. Another example of bounds is discussed in Heeringa, Little and Raghunathan (1997). They address imputation of bracketed response questions in which a respondent is unable or unwilling to provide an exact response (e.g., income and assets), but does define the bounds within which the imputed values must lie.

The goal of this paper is to propose and evaluate a general purpose multivariate imputation procedure that can handle a relatively complex data structure where explicit full multivariate models cannot be easily formulated but the imputed values for each individual are fully conditional on all the values observed for that individual. The approach is to consider imputation on a variable by variable basis but to
condition on all observed variables. The basic strategy creates imputations through a sequence of multiple regressions, varying the type of regression model by the type of variable being imputed. Covariates include all other variables observed or imputed for that individual. The imputations are defined as draws from the posterior predictive distribution specified by the regression model with a flat or non-informative prior distribution for the parameters in the regression model. The sequence of imputing missing values can be continued in a cyclical manner, each time overwriting previously drawn values, building interdependence among imputed values and exploiting the correlational structure among covariates. To generate multiple imputations, the same procedure can be applied with different random starting seeds or taking every $P^{\mathrm{W}}$ imputed set of values in the cycles mentioned above.

The variables in the data set are assumed to be of the following five types: (1) continuous, (2) binary, (3) categorical (polytomous with more than two categories), (4) counts and (5) mixed (a continuous variable with a non-zero probability mass at zero). Computationally, binary and categorical variables can be treated identically, but distinguishing them helps in conceptual understanding and in the description of the basic algorithm. We also assume that the population is essentially infinite, the sample is a simple random sample and the missing data mechanism is ignorable (Rubin 1976). The use of multiple imputation in a complex design setting has, as yet, not been fully investigated and is beyond the scope of the current paper.

In this paper we describe the sequential regression multivariate imputation (SRMI) approach in section 2 and evaluate two applications of the approach in sections 3 and 4. In the first application, it is difficult to postulate a joint multivariate distribution because of the complex systematic relationship between the variables and restrictions. In the second application, a general location model can be used to create multiple imputations (Olkin and Tate 1961; and Little and Schluchter 1985). Hence, we compare multiple imputation inferences resulting from the SRMI approach to those resulting from a joint multivariate model. The results of a simulation study investigating the sampling properties of imputed data inferences are presented in section 5, and a concluding discussion with directions for future research are given in section 6 .

## 2. IMPUTATION METHOD

For a sample of size $n$, let $X$ denote a $n \times p$ design or predictor matrix containing all the variables with no missing values. $X$ consists of continuous, binary, count or mixed variables, and appropriate dummy variables representing categorical variables. In addition, $\boldsymbol{X}$ may also consist of a column of ones to model an intercept parameter, offset variables, and certain design variables. Let $Y_{1}, Y_{2}, \ldots, Y_{k}$ denote $k$ variables with missing values, ordered, without
loss of generality, by the amount of missing values, from least to most. The pattern need not be monotone. (In a monotone pattern of missing data, $Y_{2}$ is observed only for a subset of subjects on whom $Y_{1}$ is observed, $Y_{3}$ is observed only for a subset of those on whom $Y_{2}$ is observed and so on.)

For model based imputations, the joint conditional density of $Y_{1}, Y_{2}, \ldots, Y_{k}$ given $X$ can be factored as

$$
\begin{align*}
& f\left(Y_{1}, Y_{2}, \ldots, Y_{k} \mid X, \theta_{1}, \theta_{2}, \ldots, \theta_{k}\right)= \\
& f_{1}\left(Y_{1} \mid X, \theta_{1}\right) f_{2}\left(Y_{2} \mid X, Y_{1}, \theta_{2}\right) \ldots \\
& \quad f_{k}\left(Y_{k} \mid X, Y_{1}, Y_{2}, \ldots, Y_{k-1}, \theta_{k}\right) \tag{1}
\end{align*}
$$

where $f_{j}, j=1,2, \ldots, k$ are the conditional density functions and $\theta_{j}$ is a vector of parameters in the conditional distribution (e.g., regression coefficients and dispersion parameters). In the sample survey context this can be viewed as a superpopulation model. We model each conditional density through an appropriate regression model with unknown parameters, $\theta_{j}$, and draw from the corresponding predictive distribution of the missing values given the observed values. We assume that the prior distribution for the parameters $\theta=\left(\theta_{1}, \theta_{2}, \ldots, \theta_{k}\right)$ is $\pi(\theta) \propto 1$ (diffuse relative to the likelihood). However, the method can easily be modified for specified proper prior distributions.

Each conditional regression is based on one of the following models:

1. A normal linear regression model on a suitable scale (for example, a Box-Cox power transformation may be used to achieve normality) if $Y_{j}$ is continuous;
2. A logistic regression model if $Y_{j}$ is binary;
3. A polytomous or generalized logit regression model if $Y_{j}$ categorical;
4. A Poisson loglinear model if $Y_{j}$ is a count variable; and
5. A two-stage model where zero-non zero status is imputed using logistic regression, and conditional on non-zero status, a normal linear regression model is used to impute non-zero values, if $Y_{j}$ is mixed.

Each imputation consists of c "rounds". Start round 1 by regressing the variable with the fewest number of missing values, $Y_{1}$ on $X$, imputing the missing values under the appropriate regression model. Assuming a flat prior for the regression coefficients, the imputations, for the missing values in $Y_{1}$ are the draws from the corresponding posterior predictive distribution (See Appendix A for a detailed discussion about drawing values for various regression models.) Then update $X$ by appending $Y_{1}$ appropriately (for example, dummy variables, if it is categorical) and move on to the next variable, $Y_{2}$, with the next fewest missing values. Repeat the imputation process using updated $X$ as predictors until all the variables have been imputed. That is, $Y_{1}$ is regressed on $U=X ; Y_{2}$ is regressed
on $U=\left(X, Y_{1}\right)$ where $Y_{1}$ has imputed values; $Y_{3}$ is regressed on $U=\left(X, Y_{1}, Y_{2}\right)$ where $Y_{1}$ and $Y_{2}$ have imputed values; and so on.

The imputation process is then repeated in rounds 2 through $c$, modifying the predictor set to include all $Y$ variables except the one used as the dependent variable. Thus, regress $Y_{1}$ on $X$ and $Y_{2}, Y_{3}, \ldots Y_{k}$; regress $Y_{2}$ on $X$ and $Y_{1}, Y_{3}, \ldots, Y_{k}$; and so on. Repeated cycles continue for a prespecified number of rounds, or until stable imputed values occur.

The procedure outlined above needs modification to incorporate restrictions and bounds. The restrictions are handled by fitting the models to an appropriate subset of individuals. For example, a Poisson regression model could be applied to impute any missing values for the variable "Number of Pregnancies." The imputation will be restricted to women in the sample. As a covariate, though, this variable may be treated differently when imputing subsequent variables. For instance, certain dummy variables may be created based on this variable, which hare then appended to the matrix $U$ before proceeding with the imputation of the next variable.

Consider another example, "Years Smoking Cigarettes," where the sample would be restricted to current or past smokers. If there is no evidence of smoking as a teenager, "Years Smoking Cigarettes" for a current smoker should satisfy the bound ( 0, Age -18). If there is some indication of smoking as a teenager then the range may be restricted to, say ( 0 , Age - 12). For a past smoker these ranges will be ( 0 , Age - 18 - YRSQUIT) and ( 0 , Age - 12 - YRSQUIT) respectively, where YRSQUIT is the years since the individual quit smoking. The appropriate regression model for this variable is a truncated version of the normal linear regression model (possibly on a transformed scale). The parameters, the regression coefficients and the residual variance need to be drawn from the corresponding posterior distributions. The imputations are then drawn from the corresponding truncated normal distribution conditional on the drawn value of the parameters.

It is difficult to draw values of parameters directly from their posterior distribution with truncated normal likelihoods. However, it can be easily computed for a given parameter value. The Sampling-Importance-Resampling (SIR) algorithm (Rubin 1987b, Raghunathan and Rubin 1988) can be used to draw from the actual posterior distribution. First, draw several trial parameter values from the posterior distribution without applying the bounds (untruncated normal linear regression model). Second, attach an importance ratio to each trial value, defined as the ratio of the actual posterior density with bounds to the trial density (the posterior density without bounds), both evaluated at the drawn value. Finally, resample a single parameter value with probability proportional to the importance ratios. This method requires careful monitoring of the distribution of importance ratios (Gelman, Carlin, Stern and Rubin 1995).

The bounds can also be applied to polytomous variables. For instance, suppose that a variable $Y$ can take one of $k$ values, but the observed data suggests that the missing value for a particular subject can either be $j$ or $l$. The contribution to the likelihood from this subject corresponds to the conditional binomial distribution. The draws in the multinomial step (see Appendix A) are made from the conditional distribution for these two categories. That is, the imputed value is $j$ with probabilities $s_{j} \cdot=P_{j} / /\left(P_{j}+P_{i}\right)$ and $l$ with probability $1-s_{j} \cdot$.

At the completion of the initial round of imputations, the first complete data set with no missing values is available. The factorization in Equation (1) defines a joint conditional distribution of $Y_{1}, Y_{2}, \ldots, Y_{k}$, given $X$. If the pattern of missing data is monotone, the imputations in the first round are approximate draws from the joint posterior predictive density of the missing values given the observed values. Note that the draws from the logistic, polytomous, and count variables are from large sample approximations of the posterior density of the regression coefficients. It is possible to improve upon these approximations by using, for example, the SIR algorithm or another rejection algorithm in each subsequent round.

When the pattern of missing data is not monotone, one can develop a Gibbs sampling algorithm (Geman and Geman 1984; Gelfand and Smith 1990) corresponding to Model (1). For example, conditional on the drawn values of the parameters $\theta_{2}, \theta_{3}, \ldots, \theta_{k}$ and the missing values drawn in the first round, the second round would draw values of $\theta_{1}$ from the appropriate conditional posterior density which is proportional to the first term in Equation (1). Next draw the missing values in $Y_{1}$ conditional on this drawn value of the parameter $\theta_{1}$, all other observed or imputed values for that subject and other parameters, $\theta_{2}, \theta_{3}, \ldots, \theta_{k}$ in the model. That is, the missing values in $Y_{j}$ at round $(t+1)$ need to be drawn from the conditional density,

$$
\begin{equation*}
f_{j}^{*}\left(Y_{j} \mid \theta_{1}^{(t+1)}, Y_{1}^{(t+1)}, \ldots, \theta_{j}^{(t+1)}, \theta_{j+1}^{(t)}, Y_{j+1}^{(t)}, \ldots, \theta_{k}^{(t)}, Y_{k}^{(t)}, X\right) \tag{2}
\end{equation*}
$$

computed based on the joint distribution in (1), where $Y_{l}^{(t)}$ is the imputed or observed values for variable $Y_{l}$ at round $t$. Though this is conceptually possible, it is difficult even to compute this density in most practical settings with restrictions, bounds, and the types of variables being considered.

Our proposal is to draw missing values in $Y_{j}$ at round $(t+1)$ from a predictive distribution corresponding to conditional density,

$$
\begin{equation*}
g_{j}\left(Y_{j} \mid Y_{1}^{(t+1)}, Y_{2}^{(t+1)}, \ldots, Y_{j-1}^{(t+1)}, Y_{j+1}^{(t)}, \ldots, Y_{k}^{(t)}, X, \varphi_{j}\right) \tag{3}
\end{equation*}
$$

where the conditional density $g_{j}$ is specified by one of the regression models described earlier that depends upon the variable type for $Y_{j}$, and $\varphi_{j}$ is the unknown regression parameters with diffuse prior. That is, the new imputed values for a variable are conditional on the previously imputed values of other variables, and the newly imputed values of variables that preceded the currently imputed variable. This proposal may be viewed as an approximation to an actual

Gibbs sampling where the conditional density (2) is approximated by the conditional density (3). Furthermore, this approximation can be improved by considering the SIR or some other rejection type algorithm if the conditional density in (2) can be computed up to a constant.

There are some other particular cases where this approximation is equivalent to drawing values from a posterior predictive distribution under a fully parametric model. For example, if all the variables are continuous and each conditional regression model is a normal linear regression model with constant variance, then the algorithm converges to a joint predictive distribution under a multivariate normal distribution with an improper prior for the mean and the covariance matrix.

It is theoretically possible that a sequence of draws based on densities in (3) may not converge to a stationary distribution, because these conditional densities may not be compatible with any multivariate joint conditional distribution of $Y_{1}, Y_{2}, \ldots, Y_{k}$ given $X$ (Gelman and Speed 1993). Our empirical investigations using several practical data sets have not identified, so far, any such anomalies. In several large data sets, we find the conditional densities (2) and (3) to be quite similar. As discussed in sections 4 and 5, the draws from this approach are comparable to those based on an explicit Bayesian model.

## 3. EFFECT OF SMOKING ON PRIMARY CARDIAC ARREST

In our first illustration, the SRMI approach is applied to a case-control study examining the relationship between cigarette smoking and the incidence of primary cardiac arrest (Siscovick, Raghunathan, King, Weinmann, Wicklund, Albright, Bovbjerg, Arbogast, Kushi, Cobb, Copass, Psaty, Retzlaff, Childs and Knopp 1995). In this study it is difficult to formulate an explicit model which captures the full complexity of the data. The case subjects were all King County, Washington residents who had out-of-hospital primary cardiac arrests between 1988 and 1994. The case subjects were identified through a review of paramedic incident reports. Control subjects were selected by random digit dialing from King County and matched to case subjects on gender and age (within seven years). To be eligible, subjects (case and control) were required to be between 25 and 74 years of age, married, and free of clinically-diagnosed heart disease or some other lifethreatening conditions such as cancer, liver disease, lung disease, or end-stage renal disease.

Because primary cardiac arrest has a case-fatality rate greater than $80 \%$, the eligibility criterion of marriage was included so that information regarding risk factor exposure (i.e., smoker status, years smoked) could be ascertained from surrogate respondents (i.e., spouses). Among control and surviving cases subjects, both subject and surrogate were interviewed to gather exposure data. The control and
the surviving cases subjects were interviewed mainly to study the reliability of measurements from their surrogates. Among the variables considered in this paper, there were practically no differences in the measurements obtained from the subjects and their surrogates for control or case subjects.

Table 1 gives the means, standard deviations, and percent missing values for key variables by case-control status. The exposure variables are indicator variables for Former Smoker ( $X_{1}$ ), Current Smoker ( $X_{2}$ ) and Years Smoked ( $X_{3}$ ). The confounding variables considered are Age, Body Mass Index (BMI) (BMI=Weight [in $\mathrm{Kg}] / \mathrm{Height}^{2}$ [in Meters]), and the binary variables Female and Education (High School Graduate). The substantive model of interest is the logistic regression model,

$$
\begin{aligned}
\log [\operatorname{Pr}(C=1) / \operatorname{Pr}(C=0)] & =\alpha_{0}+\alpha_{1} X_{1}+\alpha_{2} X_{2}+\alpha_{3} X_{1} X_{3} \\
& +\alpha_{4} X_{2} X_{3}+\alpha_{5} \text { Age }+\alpha_{6} \text { BMI } \\
& +\alpha_{7} \text { Female }+\alpha_{8} \text { Education }
\end{aligned}
$$

where $C$ is an indicator of cardiac arrest. Preliminary investigations indicated that linear terms for Age and BMI, are appropriate.

Table 1
Means and Proportions (in \%) for Key Variables and Percent Missing

| Variable | Control ( $n=551$ ) |  | Cases ( $n=347$ ) |  |
| :---: | :---: | :---: | :---: | :---: |
|  | \% Missing | Mean (SD) | \% Missing | Mean (SD) |
| Age | 0.0 | 58.4 (10.4) | 0.0 | 59.4 (9.9) |
| BMI | 8.2 | 25.8 (4.1) | 2.6 | 26.4 (4.6) |
| Years Smoked | 16.8 | 24.8 (14.7) | 5.4 | 31.7 (13.8) |
|  | Proportion |  |  | Proportion |
| Female | 0.0 | 23.2 | 0.0 | 19.9 |
| 2 High School | 0.0 | 76.8 | 0.0 | 61.9 |
| Smoking Status |  |  |  |  |
| Never Smoked | 0.0 | 47.2 | 0.0 | 27.3 |
| Former Smoker | 0.0 | 42.1 | 0.0 | 38.2 |
| Current Smoker | 0.0 | 10.7 | 0.0 | 34.5 |

There are no missing values for the variables Age, Female, Education, Smoking Status ( $X_{1}, X_{2}$ ), and C. Thus, for purposes of imputation, define $X=(1$, Age, Female Education, $X_{1}, X_{2}, C$ ). Log (BMI), having the fewest missing values, was regressed first on $X$ through a normal linear regression model. Residual diagnostics indicated a $\log$-transform improved the normality of residuals.

Next, Years Smoked was regressed on $U=(X, \log$ (BMI)). For this variable the sample was restricted to current and former smokers. Moreover, imputed values for Years Smoked were bounded by AGE-18, unless a respondent reported that they smoked in school (SCHSMK), and then they were bounded by AGE-12. For former smokers, imputed values were also bounded by how long ago the respondent had quit smoking (YRSQUIT). Thus, imputed values for former smokers who did not
smoke in school were bounded by AGE-18-YRSQUIT, while imputed values for former smokers that did smoke in school were bounded by AGE-12-YRSQUIT. Some subjects (5\%) had missing values on the two auxiliary items (SCHSMK, YRSQUTT) which were imputed prior to defining the upper bounds of Years Smoked. The inherent structure of this data set makes it difficult to develop explicitly a joint distribution of the variables with missing values conditional on the completed observed variables. SRMI is thus an appealing approach to handle for this type of data.

In imputing the missing values, we performed 1,000 rounds for each of 25 different starting random seeds resulting in $M=25$ imputations. The logistic regression model was fit to each imputed data set to obtain maximum likelihood estimates of the regression coefficients and asymptotic covariance matrices.

We used the standard multiple imputation variance formula (Rubin 1987a, Chap. 3) to compute the multiply imputed estimate of the regression coefficients and the covariance matrix. Briefly, suppose that $\hat{\alpha}^{(l)}$ is the estimate of the vector of regression coefficients $\alpha$ in the logistic model, and $V^{(I)}$ its covariance matrix, based on imputed data set $l$. The multiply imputed estimate of $\alpha$ is

$$
\hat{\alpha}_{\mathrm{MI}}=\sum_{l=1}^{M} \mathrm{\partial}^{(l)} / M
$$

and its covariance matrix is

$$
V_{\mathrm{MI}}=\sum_{l=1}^{M} V^{(l)} / M+\frac{M+1}{M} B_{M}
$$

where

$$
B_{M}=\sum_{l=1}^{M}\left(\hat{\alpha}^{(l)}-\hat{\alpha}_{\mathrm{MI}}\right)\left(\hat{\alpha}^{(l)}-\hat{a}_{\mathrm{MI}}\right)^{t} /(M-1)
$$

The number of imputations is larger than what is usually recommended. We performed 25 imputations with different random seeds to assess whether the Gibbs style rounds lead us to a region of the imputed values that is very different from the observed data. Graphical displays of the imputed and observed values indicated that none of the imputations in the 25,000 rounds were incompatible with the observed data distribution.

Table 2, the complete-case analysis, gives the point estimates and their standard errors based on subjects with all variables observed. A total of 103 subjects ( $11.5 \%$ ) had missing values in one or more predictors. A complete-case analysis, which is generally valid only when the data are missing completely at random was performed after deleting these 103 subjects (See Column 2, Table 2). Logistic regression analyses with a missing data indicator as the dependent variable and a number of completely observed variables as predictors indicated that the data are not missing completely at random. One may expect, therefore, that the complete case estimates and standard errors are biased.

Table 2
Point Estimates (Standard Errors) of Logistic Regression Coefficients for Model of Primary Cardiac Arrest for Complete Cases, SRMI Methods 1* and 2**

| Predictor Variables | Complete Case |  | SRMI |  |  |  |
| :--- | :---: | :---: | :---: | :---: | :---: | :---: |
|  | $(n=795)$ |  | Method 1 $(n=898)$ |  | Method 2 ( $n=898)$ |  |
|  | Estimate (SE) |  | Estimate (SE) |  | Estimate (SE) |  |
| Intercept | -2.922 | $(0.791)$ | -2.610 | $(0.757)$ | -2.348 | $(0.627)$ |
| Age | 0.015 | $(0.009)$ | 0.015 | $(0.009)$ | 0.014 | $(0.008)$ |
| Female | -0.007 | $(0.203)$ | -0.115 | $(0.189)$ | -0.119 | $(0.177)$ |
| Education | -0.448 | $(0.173)$ | -0.467 | $(0.166)$ | -0.444 | $(0.133)$ |
| BMI | 0.056 | $(0.018)$ | 0.049 | $(0.013)$ | 0.055 | $(0.009)$ |
| Current Smoker | 1.693 | $(0.569)$ | 2.001 | $(0.543)$ | 1.998 | $(0.448)$ |
| Former Smoker | 0.003 | $(0.284)$ | -0.029 | $(0.262)$ | -0.011 | $(0.223)$ |
| Current Smoker $\times$ Yrs Smoked | -0.003 | $(0.015)$ | -0.008 | $(0.013)$ | -0.005 | $(0.011)$ |
| Former Smoker $\times$ Yrs Smoked | 0.019 | $(0.009)$ | 0.014 | $(0.009)$ | 0.014 | $(0.009)$ |

* Method 1 - Imputation restricted to model variables
** Method 2 - Imputation includes model and auxiliary variables

Table 2, SRMI Method 1, gives estimates and their standard errors for SRMI using only the variables in the substantive model. These estimates are quite similar to the complete-case analysis estimates. The multiple imputation standard errors are smaller due to additional subjects with imputed data. There are modest changes in the relationship between smoking and primary cardiac arrest. The completecase analysis indicates a statistically significant relationship between years smoked and primary cardiac arrest for former smokers, while no such association is indicated in the analysis of multiply imputed data.

One of the advantages of the multiple imputation approach is that the imputation process can use additional variables not in the substantive analysis. Such situations arise when a common research database with many variables is used by different researchers, each using a subset of the variables. The imputation may be carried out for the entire database, where prediction for missing values in each variable borrows strength from all other variables in the data set. Such imputations have been shown to improve efficiency compared to those based only on variables in the particular substantive model (Raghunathan and Siscovick 1996).

Table 2, SRMI Method 2, provides multiple imputation estimates and their standard errors obtained when the entire data set was imputed using 50 additional variables. These included dietary indicators, physiological measures, socioeconomic status, and behavioural variables. The point estimates are modestly different for all the variables. The standard errors, though, are considerably smaller when compared to the multiple imputation approach using only variables in the substantive model (SRMI, Method 1). This is not surprising because many of the additional variables such as blood pressure, cholesterol counts, alcohol consumption, and physical activity were highly predictive of BMI and smoking related variables.

## 4. PARENTAL PSYCHOLOGICAL DISORDERS AND CHILD DEVELOPMENT

A second illustration examines the effects of parental psychological disorders on several measures of childhood development. Little and Schuchter (1985) analyzed the data using a general location model to obtain maximum likelihood estimates of the parameters of the joint distribution. This general location model was employed to create multiple imputations using Markov Chain Monte Carlo methods (Schafer 1997), producing fully Bayesian modelbased multiply imputed data sets. We also created multiple imputations using the SRMI procedure.

The study data consists of 69 families with two children each. Each family was classified into one of the three risk categories: (1) Normal Risk - no parental psychiatric disorders; (2) Moderate Risk - one parent diagnosed with a psychiatric illness or a chronic physical illness; and (3) High Risk - one parent diagnosed with schizophrenia or an affective mental disorder. There are three primary dependent variables of interest: $Y_{1 c}$, number of psychiatric symptoms (dichotomized as high/low) for child $c ; Y_{2 c}$, the standardized reading scores for child $c$; and $Y_{3 c}$, the standardized verbal comprehension score for child $c$.

We consider three models in investigating the impact of parental psychological disorders on childhood development. The first is a mixed effects logistic regression model:

$$
\operatorname{logit}\left[\operatorname{Pr}\left(Y_{1 i c}=1\right)\right]=\beta_{0}+\beta_{1} U_{1 i}+\beta_{2} U_{2 i}+\gamma_{j}
$$

where $Y_{1 i c}=1$ if child $c$ in family $i$ is classified as having a high number of symptoms and 0 otherwise; $U_{1 i}=1$ if family $i$ is classified as a moderate risk group and 0 otherwise; $U_{2 i}=1$ if family $i$ is classified as a high risk group and 0 otherwise; and $\gamma_{i}$ are random effects assumed to be identically and independently distributed normal random variables with mean 0 and variance $\varphi_{r}^{2}$. This
random effect accounts for intraclass correlation between the two children within the same family. With complete data, this model may be fit by maximizing the numerically integrated likelihood function of ( $\beta_{0}, \beta_{1}, \beta_{2}, \varphi_{\gamma}^{2}$ ) using the Newton-Raphson algorithm and the Gaussian quadrature method for the numerical integration of the likelihood function. These types of models can be easily fit with complete data, but are difficult to fit with missing data.

The second and third regression models relate the child's reading and verbal scores, respectively, to risk group after adjusting for the number of symptoms ( $Y_{1}$ ). An investigation of the residuals after a few preliminary rounds or reading and verbal score imputations indicated a $\log$ scale was appropriate. Thus, denoting $Y_{2 i c}$ and $Y_{3 i c}$ as the logarithm of the reading and verbal scores, respectively, for child $c$ in family $i$, we posited the following mixed effects regression model,

$$
Y_{2 i c}=\alpha_{0}+\alpha_{1} U_{1 i}+\alpha_{2} U_{2 i}+\alpha_{3} Y_{1 i c}+\delta_{i}+\varepsilon_{i c} .
$$

where $\delta_{i}$ and $\varepsilon_{i c}$ are mutually independent normal random variables with mean 0 and variances $\sigma_{\delta}^{2}$ and $\sigma_{\varepsilon}^{2}$ respectively. Again, with no missing data in the covariates, the maximum likelihood estimates of the unknown parameters can be readily obtained using, for example, the PROC MIXED procedure in SAS.

There were no missing values in the classification of the risk groups, and thus we defined $X=\left(1, U_{1}, U_{2}\right)$. The variables with missing values, $Y_{21}, Y_{22}, Y_{31}$ and $Y_{32}$ were imputed using normal linear regression, and the missing values in $Y_{11}$ and $Y_{12}$ were imputed using logistic regression. We created $M=25$ SRMIs, repeating the process through 1,000 rounds and 25 different seeds. The SRMI multiply imputed data sets were analyzed and combined using the methods described earlier. To compare these results with the multiply imputed inferences when the imputations are draws from the posterior predictive distribution under the general location model we created 25 imputations under a fully Bayesian model using software developed by Schafer (1997). The point estimates and
standard errors for the three models using SRMI and Bayes multiple imputation approaches are presented in Table 3. There are no real meaningful differences between the SRMI estimates and standard errors and those resulting from the Bayesian imputation. Children of parents in the high risk group are approximately $7.8[\exp (2.048)]$ times more likely to have a high number of symptoms than children with parents in the normal group under the SRMI. The $\mathbf{9 5 \%}$ confidence interval for this relative risk is (3.8, 16.0). For the moderate risk, group, the corresponding point and interval estimates are 3.7 and (1.8,7.8). These estimates may be contrasted with those obtained based on the complete-case analysis (not shown): $7.4(2.3,24.2)$ for the high risk group, and $3.5(1.0,11.9)$ for the moderate risk group (data not shown). Though the point estimates of the relative risks are similar, the complete-case confidence intervals are wider because they are based only on $60 \%$ of the observations.

Based on the estimated regression coefficients in Table 3 , one can infer, after adjusting, for the number of symptoms, that children in the moderate and high risk groups have lower reading scores, by about 11 points [exp (4.654)-exp(4.654-0.110)], when compared to the normal group. On the other hand, the complete-case analysis estimates a score of 16 points lower for children in the moderate risk group than their counterparts in the normal group, and children in the high risk group score about 19 points lower when compared to the normal group.

The SRMI analysis of verbal scores suggests that the children in the moderate and high risk groups score about 20 and 24 points lower, respectively, than their counterparts in the normal group. However, the complete-case analysis shows the moderate risk group scores lower by 36 points and the high risk group scores lower by about 39 points when compared to the normal group. Thus, the completecase estimates of the effects of parental psychological disorders on the child's reading and verbal scores are quite different than those obtained by the analysis of the multiply imputed data. This is not surprising because the data on reading and verbal scores are not missing completely at

Table 3
Point Estimates (Standard Errors) of Regression Coefficients for Three Models of Child Development Under SRMI and Bayesian Imputation

| Predictor Variables | Imp. Method |  | Dependent Variable |  |  |  |  |  |  |  |  |  |
| :--- | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
|  |  | Symptoms |  |  |  |  |  |  |  | Reading Score |  | Verbal Score |
| Intercept | SRMI | -0.678 | $(0.256)$ | 4.654 | $(0.013)$ | 4.873 | $(0.020)$ |  |  |  |  |  |
|  | Bayes | -0.688 | $(0.257)$ | 4.556 | $(0.013)$ | 4.991 | $(0.021)$ |  |  |  |  |  |
| High Risk Group | SRMI | 2.048 | $(0.356)$ | -0.109 | $(0.022)$ | -0.191 | $(0.032)$ |  |  |  |  |  |
|  | Bayes | 2.033 | $(0.350)$ | -0.108 | $(0.021)$ | -0.180 | $(0.033)$ |  |  |  |  |  |
| Moderate Risk Group | SRMI | 1.289 | $(0.366)$ | -0.110 | $(0.022)$ | -0.162 | $(0.033)$ |  |  |  |  |  |
|  | Bayes | 1.300 | $(0.360)$ | -0.109 | $(0.023)$ | -0.167 | $(0.035)$ |  |  |  |  |  |
| Symptoms | SRMI |  | - |  | 0.032 | $(0.022)$ | -0.083 | $(0.032)$ |  |  |  |  |
|  | Bayes |  | - |  | 0.031 | $(0.019)$ | -0.080 | $(0.030)$ |  |  |  |  |

random and are related to the risk group as well as the number of symptoms of the child.

## 5. SIMULATION STUDY

The analyses described in sections 3 and 4 indicate that sensible results can be obtained by applying the SRMI approach to handling missing values. Nevertheless, it is difficult to conclude based on such case studies whether or not the approach will result in valid inferences in routine applications. A simulation study was designed to investigate the repeated sampling properties of inferences from imputed data sets created with the SRMI approach. Complete data sets were generated from hypothetical populations, and elements deleted under an ignorable missing data mechanism. The deleted values were imputed and differences in summary statistics based on the imputed data sets and the before deletion or full data sets were assessed.

More formally, the strategy:
(1) generated a complete data set which did not agree perfectly with our multiple imputation strategy,
(2) estimated selected regression parameters,
(3) deleted certain values using an ignorable missing data mechanism,
(4) used SRMI to multiply impute the missing values, and
(5) obtained multiply imputed estimates for the regression parameters estimated in step 2.
The differences in the parameter are examined across several independent replications of this strategy.

A total of 2,500 complete data sets with three variables ( $U, Y_{1}, Y_{2}$ ) and sample size 100 were generated using the following models:

1. $\mathrm{U} \sim \operatorname{Normal}(0,1)$;
2. $\quad Y_{1} \sim$ Gamma with mean $\mu_{1}=\exp (U-1)$ and variance $\mu_{1}^{2} / 5$; and
3. $Y_{2} \sim$ Gamma with mean $\mu_{2}=\exp \left(-1+0.5 U+0.5 Y_{1}\right)$ and variance $\mu_{2}^{2} / 2$.

The model for $Y_{2}$ in step 3 is the primary regression model of interest with true regression coefficients $\beta_{0}=-1, \beta_{1}=\beta_{2}=0.5$, and dispersion parameter $\varphi^{2}=0.5$. For the complete data this model can be fixed using statistical software packages such as GLIM or Splus.

The deletion or missing data mechanisms were as follows:
(1) No missing values in $U$;
(2) the missing values in $Y_{1}$ depend on $U$ through a logistic function $\operatorname{logit}\left[\operatorname{Pr}\left(Y_{1}\right.\right.$ is missing $\left.)\right]=1.5+U$; and
(3) the missing values in $Y_{2}$ depend on $U$ and $Y_{1}$ through a logistic function $\operatorname{logit}\left[\operatorname{Pr}\left(Y_{2}\right.\right.$ is missing $\left.)\right]=1.5-$ $0.5 Y_{1}-0.5 U$.

These missing data mechanisms generated $22 \%$ missing data in $Y_{1}$ and $29 \%$ missing data in $Y_{2}$. The complete-case analysis would have only used $48 \%$ of the data.

Since SRMI allows us only to fit a normal linear regression model, the imputations were carried out as follows, Suppose that $Y_{1}$ has fewer missing values, and let $Z_{1}=\left(Y_{1}^{\lambda_{1}}-1\right) / \lambda_{1}$ be the Box-Cox transformation of the continuous variable. In the first round of imputations, assume that $Z_{1}$ has a normal distribution with mean $a_{0}+a_{1} U$ and variance $\sigma_{1}^{2}$, where $\lambda_{1}$ was estimated using the maximum likelihood approach, and that $Z_{2}=$ $\left(Y_{2}^{\lambda_{2}}-1\right) / \lambda_{2}$ has a normal distribution with mean $b_{0}+b_{1} U+b_{2} Z_{1}$ and variance $\sigma_{2}^{2}$, where $\lambda_{2}$ was estimated using maximum likelihood. In the subsequent rounds, $U$ and $Z_{2}$ are predictors for $Z_{1}$, and $U$ and $Z_{1}$ are predictors for $Z_{2}$. The estimation of a power transformation using maximum likelihood was automated while fitting each regression model.

For each of the 2,500 simulated data sets with missing values, a total 250 rounds with $M=5$ different random starts were created using SRMI. For each replicate, the resulting $M=5$ imputed data sets and the full data set (before deletion) were analyzed by fitting the Gamma model for $Y_{2}$ using maximum likelihood. The multiple imputation estimate was constructed as the average of the five imputed data estimates. To assess the differences in the point estimates we computed the standardized difference between the SRMI and full data estimates,

$$
\Delta(\beta)=\frac{100 \times \text { abs }(\text { SRMI estimate }- \text { Full Data Estimate })}{\text { SE(SRMI Estimate })}
$$

Table 4 gives the mean and standard deviation of $\Delta(\beta)$ for three regression coefficients $\beta_{0}, \beta_{1}$, and $\beta_{2}$ in the model. The SRMI estimates are typically within $8 \%$ of the full standard units. The actual coverage and the average length of the $95 \%$ SRMI confidence intervals were computed for the regression coefficients using the $t$ reference distribution described in Rubin (1987b). For each simulated data set and parameter, it was determined whether or not the true value (e.g., $\beta_{1}=0.5$ ) is contained within the corresponding interval. The proportion of intervals containing the true values were computed across the 2,500 replications and are provided in Table 4. For the full data sets, the actual coverage for $\beta_{1}$, for example, was $94.9 \%$ and for SRMI it was 95.4. In addition the average length of the confidence intervals were also computed. The average width of the full data confidence interval for $\beta_{1}$ was 0.91 and for SRMI the average length was 1.22 . That is, the SRMI data resulted in well calibrated intervals estimates.

The same simulation study was also used to compare the distributional properties of imputations from SRMI and a fully Bayesian method. For the model assumptions used to generate complete data, we developed a Markov Chain Monte-Carlo algorithm for drawing values from the actual posterior predictive distribution of the missing values given
the observed values. Each step of the draw used MetropolisHastings algorithm and required considerably more computational time than the SRMI method. Therefore, only the first 500 simulated data sets were used in this comparison. We computed two Kolmogrove-Smimoff (KS) statistics from each simulated data set: One comparing the imputations from the SRMI method and the actual hidden values and the other comparing the Bayesian imputations and the actual hidden values. There were no discernible differences in these two statistics across the 500 simulated data sets. A scatter plot of those 500 pairs of KS statistics showed a narrow scatter of points around a 45 degree line.

Table 4
Means and Standard Deviations for Standardized Differences Between SRMI Estimates and Full Data Estimates and Actual Coverage of Nominal 95\% Confidence Intervals

| Regression <br> Coefficient | Std. Difference | Confidence Coverage |  |  |
| :---: | :---: | :---: | :---: | :---: |
|  | Mean | SD | SRMI | Full Data |
| $\beta_{0}$ | 8.2 | 2.0 | 96.1 | 95.4 |
| $\beta_{1}$ | 8.8 | 1.7 | 95.4 | 94.9 |
| $\beta_{2}$ | 8.0 | 2.2 | 95.3 | 94.7 |

## 6. DISCUSSION

We have described and evaluated a sequential regression multivariate imputation procedure that can be used to impute missing values in a variety of complex data structures involving many types of variables, restrictions, and bounds. This procedure should be useful when the specification of a joint distribution of all the variables with missing values is difficult. A real advantage of the approach is its flexibility in handling each variable on a case by case basis. For instance, to preserve all the bivariate correlations, all the main effect terms must be included as regressors, and to preserve, say, three factor interactions all two factor interactions must be included as regressors in the imputation model. Implementation of this procedure only requires a good random number generator and fitting routines for a variety of multiple regression routines. A SAS based application implementing this approach can be downloaded from a web site (www.isr.umich.edu/ src/smp/ive).

In certain instances, one can modify the algorithm to reduce it to Gibbs sampling from the joint predictive distribution of the missing values given the observed values. However, the SRMI procedure will be more useful where an explicit model is difficult to formulate. In both the illustrations and the simulation, different random starts were used to monitor imputed values, an important aspect in many practical applications. This is a good practice when Gibbs sampling is used under an explicit Bayesian model (Gelman and Rubin 1992) and should be used when the sequential regression method discussed in this paper is used.

The simulation study described in section 5, though limited, is favorable as far as inferences based on the SRMI are concerned. The imputations from SRMI and Bayes model were comparable. The goal here, however, was to develop an imputation approach that is finely tuned on a variable by variable basis fully conditional on all the observed information, rather than an explicit joint multivariate distribution of all the variables. Furthermore, model sensitivity may be reduced by using a semiparametric regression model for each conditional regression. The Bayesian interpretation of the spline smoothing models (Silverman 1985) can be used to draw imputed values from the predictive distribution. Such modifications also deserve further investigation.

For some large data sets with many variables, the SRMI can be computationally intense. The algorithm can be modified to apply a variable selection method for each regression in each round. We compared the inferences with and without the variable selection on several large data sets such as the National Health Interview Survey and the National Medical Expenditure Survey using several hundred variables. The descriptive inferences as well as inferences based on linear and logistic regression models were very similar, still further detailed investigation is needed.

It is also possible to use the imputation approach discussed in this paper in conjunction with, for example, the Jackknife Repeated Replication (JRR) technique for variance estimation. Specifically, (1) re-impute, singly, the missing values in each jackknife replicate SRMI; (2) analyze the imputed replicate data set; and, finally, (3) combine the replicate estimates to obtain the point estimate and its covariance matrix. This approach is more computationally intensive than the multiple imputation approach. This integrated JRR imputation approach and several of its variations are currently under investigation.

Finally, it has been assumed that the data set arises from a simple random sample design. However, most surveys employ complex sample designs involving stratification, clustering, and weighting. Further work is needed to modify the sequential regression method to incorporate complex design features not reflected in the $X$ variables in expression (1). However, even if the imputation process ignores the complex design features, the analysis of completed data should be design based. Though this does not provide valid design-based inferences, it maintains the robustness underlying the design-based analysis to a certain degree. The integrated JRR imputation approach discussed above may have more appealing design-based properties in a complex design setting.

## ACKNOWLEDGEMENTS

The authors would like to thank the three referees for their careful reading of this article and their helpful suggestions. The research was partially supported by a NSF grant DMS-0803720.

## APPENDIX: REGRESSION MODELS AND IMPUTATIONS

Dropping the subscript indexing of the variables for brevity, the necessary steps for imputing each type of variable are as follows:
Continuous variable: For $Y$ (possibly transformed from the original scale for normality), a continuous variable, build a normal linear regression model, $Y=U \beta+e$, where $U$ is the most recently updated predictor matrix, $e$ has a multivariate normal distribution with mean zero and variance $\sigma^{2} I$, and $I$ is an identity matrix. Suppose that $\theta=(\beta, \log \sigma)$ has a uniform prior distribution over the appropriate dimensional real space. Fit this model based on the individuals for whom $Y$ is observed.

Let $B=\left(U^{t} U\right)^{-1} U^{t} Y$ be the estimated regression coefficient, SSE $=(Y-U B)^{t}(Y-U B)$ be the residual sum of squares and $\mathrm{df}=\operatorname{rows}(Y)-\operatorname{cols}(U)$ be the residual degrees of freedom, and $T$ be the Cholesky decomposition such that $T T^{t}=\left(U^{\prime} U\right)^{-1}$. The relevant posterior distributions can be derived easily (see, for example, Gelman, Carlin, Stern and Rubin 1995, Chap. 7), and the following steps then provide draws from the posterior predictive distribution of missing $Y$ values:

1. Generate a chi-square random deviate $u$ with $d f$ degrees of freedom and define $\sigma_{*}^{2}=\operatorname{SSE} / u$.
2. Generate a vector $z=\left(z_{1}, z_{2}, \ldots, z_{p}\right)$ of dimension $p=$ rows $(B)$ of random normal deviates and define $\beta_{*}=B+\sigma_{*} T z$.
3. Let $U_{\text {miss }}$ denote the $U$-matrix for those with missing $Y$ values. The imputed values are $Y_{*}=U_{\text {miss }} \beta_{*}+\sigma_{*} v$, where $v$ is an independent vector of dimension rows ( $U_{\text {miss }}$ ) of random normal deviates.

Binary Variable: When $Y$ is a binary variable, fit a logistic regression model relating $Y$ to $U$ (most recently updated), $\operatorname{logit}[\operatorname{Pr}(Y=1 \mid U)]=U \beta$, using individuals with observed $Y$. The imputed values for $Y$ are created through the following steps:

1. Let $B$ denote the maximum likelihood estimates of $\beta$ and $V$ its asymptotic covariance matrix (negative inverse of the observed Fisher information matrix). Let $T$ be the Cholesky decomposition of $V$ (that is, $T T^{t}=V$ ). Generate a vector $z$ of random normal deviates of dimension rows $(B)$. Define $\beta_{*}=B+T z$.
2. Let $U_{\text {miss }}$ denote the portion of $U$ for which $Y$ is missing. Define $P_{*}=\left[1+\exp \left(-U_{\text {miss }} \beta_{*}\right)\right]^{-1}$. Generate a vector $u$, of dimension rows ( $U_{\text {miss }}$ ) of uniform random numbers between 0 and 1 . Impute 1 if a particular component of $u$ is less than or equal to the corresponding component of $P$, and impute 0 otherwise.
This approach results only in approximate draws from the posterior predictive distribution of the missing values as
the draws of the parameter $\beta$ are from the asymptotic approximation of its actual posterior distribution. It is possible to draw from the actual distribution by modifying Step 1 using, for example, Sampling-Importance-Resampling (Rubin 1987b).
Mixed Variable: For $Y$, a mixed variable (that is, $Y$ either takes the value zero or a continuous value), model the zero values by a $0-1$ indicator to distinguish between 0 and nonzero values, and then model a normally distributed variable for the continuous portion of the distribution conditional on the indicator variable being equal to 1 . That is, use a two stage approach: impute a one or zero using the logistic approach described above; and then restricting the sample to those with non-zero values, use the continuous variable approach described above to impute a continuous value to replace the just imputed value of 1 .
Count Variable: For $Y$, a count variable, fit a Poisson regression model $Y \sim$ Poisson ( $\lambda$ ) where $\log \lambda=U \beta$. The imputations for missing values in $Y$ are created using the following steps:
3. Let $B$ denote the maximum likelihood estimate of $\beta, V$ its covariance matrix and $T$ the Cholesky decomposition of $V$. Generate a vector $z$ of random normal deviates of dimension rows ( $B$ ) and define $\beta_{*}=B+T z$.
4. Let $U_{\text {miss }}$ denote the portion of $U$ for which $Y$ is missing. Define $\lambda_{.}=\exp \left(U_{\text {miss }} \beta_{*}\right)$. Generate independent Poisson random variables with means as the elements of $\lambda_{\text {. }}$.
Polytomous Variable: For $Y$ that can take $k$ values, $j=1,2, \ldots, k$, let $\pi_{j}=\operatorname{Pr}(Y=j \mid U)$. Fit a polytomous regression model relating $Y$ to $U$ where $\log =\left(\pi_{j} / \pi_{k}\right)=U \beta_{j}$ for $j=1,2, \ldots, k-1$. Under the restriction $\Sigma_{j}^{k} \pi_{j}=1$, it follows that $\pi_{k}=\left(1+\sum_{j}^{k-1} \exp \left(U \beta_{j}\right)\right)^{-1}$.

Let $B$ denote the maximum likelihood estimate of the regression coefficients $\left(\beta_{1}^{t}, \beta_{2}^{t}, \ldots, \beta_{k-1}^{t}\right), V$ be the asymptotic covariance matrix and $T$ its Cholesky decomposition.
The following steps create imputations:

1. Define $\beta_{*}=B+T z$ where $z$ is a vector of random normal deviates of dimension rows $(B)$.
2. Let $U_{\text {miss }}$ denote the rows of $U$ with missing $Y$ and let $P_{i}^{*}=\exp \left\{U_{\text {miss }} \beta_{i} \cdot\right\} /\left\{1+\sum_{i} \exp \left(U_{\text {miss }} \beta_{i}\right)\right\}$ where $\beta_{i}$. is the appropriate elements of $\beta$. where $i=1,2, \ldots, k-1$ and $P_{k}^{*}=1-\Sigma_{i} P_{i}^{*}$.
3. Let $R_{0}=0, R_{j}=\sum_{j}^{j} P_{i}^{*}$ and $R_{k}=1$ be the cumulative sums of the probabilities. To impute values generate random uniform number $u$ and take $j$ as the imputed category if $R_{j-1} \leq u \leq R_{j}$.
Again, the imputation of mixed, count and categorical variables are from approximate posterior predictive distributions because the corresponding parameters are drawn from their asymptotic normal approximate posterior distributions.

## REFERENCES

BARNARD, J. 1995. Cross-Match Procedures for Multiple Imputation Inference: Bayesian Theory and Frequentist Evaluation. Unpublished Doctoral Thesis, University of Chicago, Department of Statistics.

GELFAND, A.E., and SMITH, A.M.F. (1990). Sampling based approaches to calculating marginal densities. Joumal of American Statistical Association, 85, 398-409.
GELMAN, A., CARLIN, J., STERN, H. and RUBIN, D.B. (1995). Bayesian Data Analysis. London. Chapman and Hall.
GELMAN, A., and RUBIN, D.B. (1992). Inference from iterative simulation using multiple sequences (with discussion). Statistical Science, 7, 457-472.
GELMAN, A., and SPEED T.P. (1993). Characterizing a joint probability distribution by conditionals. Journal of Royal Statistical Society, B, 55, 185-188.
GEMAN, S., and GEMAN, D. (1984). Stochatic relaxation, Gibbs distribution, and the Bayesian restoration of images. IEEE Transactions on Pattern Analysis and Machine Intelligence, 6, 721-741.

HEERINGA, S.G., LITTLE, R.J.A. and RAGHUNATHAN, T.E. (1997). Imputation of Multivariate Data on Household Net Worth. University of Michigan, Ann Arbor, Michigan.
LI, K.H., MENG, X.L., RAGHUNATHAN, T.E. and RUBIN, D.B. (1991). Significance levels from repeated $p$ values from multiplyimputed data. Statistica Sinica, 1, 65-92.
LI, K.H., RAGHUNATHAN, T.E. and RUBIN, D.B. (1991). Large sample significance levels from multiply imputed data using moment-based statistics and an F reference distribution. Journal of American Statistical Association, 86, 1065-1073.

LITTLE, R.J.A., and RUBIN, D.B. (1987). Statistical Analysis with Missing Data. New York: Wiley.
LITTLE, R.J.A., and SCHLUCHTER, M.D. (1985). Maximum likelihood estimation for mixed continuous and categorical data with missing values. Biometrika, 72, 497-512.
LIU, C. (1995). Missing data imputation using the multivariate $t$ distribution. Joumal of Multivariate Analysis, 53, 139-158.
MADOW, W.G., NISSELSON, H., OLKIN, I. and RUBIN, D.B. (1983). Incomplete Data in Sample Surveys. 1,2, and 3, New York, Academic Press.
MENG, X.L., and RUBIN, D.B. (1992). Performing likelihood ratio tests with multiply imputed data sets. Biometrika, 79, 103-111.
OLKIN, I., and TATE, R.F. (1961). Multivariate correlation models with mixed discrete and continuous variables. Annals of Mathematical Statistics, 32, 448-465.

RAGHUNATHAN, T.E., and GRIZZLE, J.E. (1995). A split questionnaire survey design. Journal of American Statistical Association, 90, 54-63.

RAGHUNATHAN, T.E., and RUBIN, D.B. (1988). An application of Bayesian statistics using sampling/importance resampling to a deceptively simple problem in quality control. Data Quality Control: Theory and Pragmatics, (G.E. Liepins and V.R.R. Uppuluri, Eds). New York: Marcel Dekker.
RAGHUNATHAN, T.E., and SISCOVICK, D.S. (1996). A multiple imputation analysis of a case-control study of the risk of primary cardiac arrest among pharmacologically treated hypertensives. Applied Statistics, 45, 335-352.

RAO, J.N.K., and SHAO, J. (1992). Jackknife variance estimation with survey data under hot-deck imputation. Biometrika, 79, 811822.

RUBIN, D.B. (1976). Inference and missing data (with discussion). Biometrika, 63, 581-592.

RUBIN, D.B. (1978). Multiple imputation in sample surveys - A phenomenological Bayesian approach to nonresponse. Proceedings of the Survey Research Methods Section, American Statistical Association, 20-34.

RUBIN, D.B. (1987a). Multiple Imputation for Nonresponse in Surveys. New York: Wiley.
RUBIN, D.B. (1987b). The SIR-algorithm - A discussion of Tanner and Wong's. The calculation of posterior distributions by data augmentation. Journal of American Statistical Association.

RUBIN, D.B. (1996). Multiple imputation after $18+$ years. Joumal of American Statistical Association, 91, 473-489.

RUBIN, D.B., and SCHAFER, J.L. (1990). Efficiently creating multiple imputations for incomplete multivariate normal data. Proceeding of the Statistical Computing Section of the American Statistical Association, 83-88.
SCHAFER, J.L. (1997). Analysis of Incomplete Multivariate Data by Simulation. New York; Chapman and Hall.

SILVERMAN, B.W. (1985). Some aspects of the spline smoothing approach to nonparametric regression curve fitting. Journal of Royal Statistical Society, B, 47, 1-52.
SISCOVICK, D.S., RAGHUNATHAN, T.E., KING, I., WEINMANN, S., WICKLUND, K.G., ALBRIGHT, J., BOVBJERG, V., ARBOGAST, P., KUSHI, L., COBB, L., COPASS, M.K., PSATY, B.M., RETZLAFF, B., CHILDS, M. and KNOPP, R.H. (1995). Dietary intake and cell-membrane levels of long-chain $n-3$ polyunsaturated fatty acids and the risk of primary cardiac arrest. Journal of American Medical Association, 274, 1363-1367.

# A Better Understanding of Weight Transformation Through a Measure of Change 

JOHANE DUFOUR, FRANÇOIS GAGNON, YVES MORIN, MARTIN RENAUD and CARL-ERIK SÄRNDAL ${ }^{1}$


#### Abstract

The literature on longitudinal surveys of households offers several approaches for creating a set of final weights for use in data analysis. Most of these approaches depend on various procedures for modifying weights. Initial weights are often transformed into a set of intermediate weights in order to compensate for nonresponse, and then into a set of final weights, through poststratification, in order to adjust the sample. The literature includes a great deal of information about this approach but none of the studies has really looked closely at an approach for measuring the relative importance of these two steps in measuring the effectiveness of the numerous existing alternatives for creating intermediate weights. The objective of this paper is to study and measure the change (from the initial to the final weight) which results from the procedure used to modify weights. A breakdown of the final weights is proposed in order to evaluate the relative impact of the nonresponse adjustment, the correction for poststratification and the interaction between these two adjustments. This measure of change is used as a tool for comparing the effectiveness of the various methods for adjusting for nonresponse, in particular the methods relying on the formation of Response Homogeneity Groups. The measure of change is examined through a simulation study, which uses data from a Statistics Canada longitudinal survey, the Survey of Labour and Income Dynamics. The measure of change is also applied to data obtained from a second longitudinal survey, the National Longitudinal Survey of Children and Youth.


KEY WORDS: Nonresponse; Weighting; Calibration; Longitudinal survey; Measure of change.

## 1. INTRODUCTION

The literature contains many two-step approaches to transforming weights for household surveys. The first step involves an adjustment of the initial weights in order to compensate for nonresponse; the resulting weights are called intermediate weights. The second step produces the final weights through the process of poststratification, or more commonly through calibration (see Deville and Särndal 1992), in order to ensure that the final weights respect certain known population control totals. All of these weight modifications are designed to produce the "best possible set of final weights".

At Statistics Canada, longitudinal surveys of households also use this two-step approach in weighting, and the research work undertaken by the Agency leans in this direction. The U.S. Bureau of the Census "Survey of Income and Program Participation (SIPP)" (see Rizzo, Kalton and Brick 1996) also uses this type of approach.

Several methods are recommended in the literature for adjusting weights to compensate for nonresponse. Rizzo et al. (1996) compared the estimates obtained through several of these methods to estimates from independent sources. However, not many authors have done simulations or proposed tools for comparing the relative effectiveness of the methods in terms of their ability to reduce the nonresponse bias.

The main objective of this document is to study and measure the change (between initial and final weights) resulting from the adoption of a two-step procedure for modifying weights. Thus, a measure of change involving
four components is proposed in order to quantify the relative impact of the nonresponse adjustment, the correction for poststratification and the interaction between these two adjustments. The second objective is to use the measure of change to compare the effectiveness of the different nonresponse adjustment methods through a simulation study based on data from the Longitudinal Survey of Labour and Income Dynamics (SLD) and from the National Longitudinal Survey of Children and Youth (NLSCY). The longitudinal surveys are unique in that a great deal of information about respondents and nonrespondents to the latest wave is available from respondents to the previous waves. Thus, more complex methods can be used to adjust for nonresponse.

A general framework for the weighting of longitudinal surveys of households is presented in section 2 . Then, the measure of change which will be used to quantify the stages of transformation between the initial and the final weights is presented in section 3 . Section 4 addresses the nonresponse adjustment strategies contained in the literature. This is followed by sections 5 and 6 , which contain the results of the studies based on the SLID and NLSCY. The last section presents the conclusions of this study.

## 2. GENERAL FRAMEWORK FOR LONGITUDINAL WEIGHTING

In a longitudinal survey of households, individuals in the initial sample are followed over time, and are referred to as longitudinal individuals. This set of individuals is the one

[^7]which will be used in the studies presented in this document. They are referred to as the "reference unit". This section provides an overview of the steps followed in order to modify the initial weight for longitudinal individuals into a final weight.

### 2.1 Initial Weights

$U=\{1, \ldots, k, \ldots, N\}$ is a finite population. We are interested in variable $y$ (the variable of interest), whose value for the $k$-th unit is recorded as $y_{k}$. The objective is to estimate the total $Y=\Sigma_{U} y_{k}$. Let $w_{0 k}$ be the initial weight for all $k \in s$ units, where $s$ is the longitudinal sample. In the absence of nonresponse, the set of initial weights $\left\{w_{0 k}: k \in s\right\}$ yields the $\hat{Y}=\Sigma_{s} w_{0 k} y_{k}$ estimator for $Y$. In this case we assume that the $w_{0 k}$ are normalized in order to ensure that $\Sigma_{s} w_{0 k}=N$. Although $\hat{Y}$ is unbiased for $Y, \hat{Y}$ has the drawback of not incorporating any ancillary information in the form of known control totals for poststrata.

### 2.2 Nonresponse Adjustment and Intermediate Weights

Most surveys have to deal with nonresponse. Two approaches are often used to compensate for this: imputation and the correction of the initial weights of respondents through an adjustment factor. The latter is the one more commonly used in household surveys to compensate for total nonresponse, while imputation is often preferred when dealing with partial nonresponse. Total nonresponse reduces the size of the sample since the $y_{k}$ value is only available for $k \in r$, where $r \subset s$ is the set of the $m$ responding units. For this reduced set of data, the initial $w_{0 k}$ weights are, on average, too small and we have $\sum_{r} w_{0 k}<N$. The estimator $\hat{Y}^{\prime}=\sum_{r} w_{0 k} y_{Y}$ is not admissible since it systematically underestimates $Y$.

Weight adjustment is often chosen in order to compensate for total nonresponse in household surveys. A common method of adjusting weights involves constructing Response Homogeneity Groups (RHGs). These are designed so that each one is comprised of reference units having a similar probability of response. Then, within each RHG, an adjustment factor equal to the inverse of the RHG's response rate (weighted or not) is calculated. For each respondent unit $k$, the adjustment for nonresponse involves multiplying $w_{0 k}$ by the RHG's adjustment factor. This operation results in a set of intermediate weights $\left\{w_{1 r}: k \in r\right\}$, where $\Sigma_{r} w_{1 k}=N$. With these weights, we can construct the estimator $\hat{Y}^{\prime \prime}=\Sigma_{r} w_{1 k} y_{k}$, which eliminates the underestimation which is characteristic of $\hat{Y}^{\prime}=\Sigma_{r} w_{0 k} y_{k}$. As in the case of the set of initial weights, the main drawback with this set is that it fails to incorporate the ancillary information available for poststrata.

### 2.3 Poststratification and Final Weights

A widely-used practice in household surveys involves modifying the intermediate weights through poststratification, or, more commonly, through calibration, so that
the sum of the final weights on the set of respondents will correspond to the known population counts. Thus, postratification produces a set of final weights $\left\{w_{2 k}: k \in r\right\}$, which incorporates the ancillary information and which is also consistent with the control totals for the poststrata. In this case, the final weights in each poststratum $p$ confirm $\Sigma_{r_{p}} w_{2 k}=N_{p}$, where $N_{p}$ is the known element and $r_{p}$ is the set of respondent units in the $p$-th poststratum. It follows that $\Sigma_{r_{r}} w_{2 k}=N$. Demographic and geographic variables are frequently used to define poststrata. The choice of poststrata, which must be sufficiently large, is limited by the availability of control totals. Several methods may be used to calibrate the intermediate weights to the selected control totals.

## 3. MEASURE OF CHANGE FROM INITIAL TO FINAL WEIGHTS

In this section, a measure of the change between initial and final weights is presented so to better understand the effect of the weight modification procedure. The breakdown of this measure into four components makes it possible to quantify the effect of each of the weighting steps described in section 2. These components will be used in sections 5 and 6 in the comparison of various methods for adjusting weights for nonresponse.

If the initial weights are normalized so that $\Sigma_{s} w_{0 k}=N$, and if $r \subset s$, then the three sets of weights described in section 2 confirm the following relations:

$$
\sum_{r} w_{0 k}<N, \sum_{r} w_{1 k}=N, \sum_{r} w_{2 k}=N .
$$

Let

$$
\bar{w}_{01}=\frac{\sum_{r} w_{1 k}}{\sum_{r} w_{0 k}} \text { and } \bar{w}_{02}=\frac{\sum_{r} w_{2 k}}{\sum_{r} w_{0 k}}
$$

The ratio $\bar{w}_{01}$ measures the average change in the intermediate weight set in relation to the initial weight set. As total nonresponse becomes more pronounced, $\bar{w}_{01}$ shifts farther away from the value of 1 , which is only obtained in the absence of nonresponse. The ratio $\bar{w}_{02}$ represents the average change in the set of final weights in relation to the set of initial weights.

The $\bar{w}_{01}$ and $\bar{w}_{02}$ ratios measure the average change in weight. To measure an individual change in weight, we define, for every $k \in r, r_{01 k}=w_{1 k} /\left(w_{0 k} \bar{w}_{01}\right)$, and $r_{02 k}=w_{2 k} /\left(w_{0 k} \bar{w}_{02}\right)$. These quantities vary around 1 . More specifically, their weighted averages equal 1:

$$
\frac{\sum_{r} w_{0 k} r_{01 k}}{\sum_{r} w_{0 k}}=\frac{\sum_{r} w_{0 k} r_{02 k}}{\sum_{r} w_{0 k}}=1
$$

The $r_{01 k}$ and $r_{02 k}$ quantities will be useful for measuring individual weight changes.

The total weight change, from the set of initial to final weights, going through the set of intermediate weights, can be calculated by a measure of change, also called distance. Here, $D$ is the following measure of change:

$$
D=\frac{\sum_{r} w_{0 k}\left(\frac{w_{2 k}}{w_{0 k}}-1\right)^{2}}{\sum_{r} w_{0 k}}
$$

In fact, $D$ is a weighted average of the following individual weight change factors:

$$
\left(\frac{w_{2 k}}{w_{0 k}}-1\right)^{2}=\left(\frac{w_{2 k}}{w_{1 k}} \frac{w_{1 k}}{w_{0 k}}-1\right)^{2}
$$

The measure of change $D$ breaks down into four components, as set out in the following equation:

$$
D=R_{01}+R_{12}+R_{\mathrm{int}}+G
$$

where:

$$
\begin{aligned}
R_{01} & =\bar{w}_{02}^{2} \frac{\sum_{r} w_{0 k}\left(r_{01 k}-1\right)^{2}}{\sum_{r} w_{0 k}}, \\
R_{12} & =\bar{w}_{02}^{2} \frac{\sum_{r} w_{0 k}\left(r_{02 k}-r_{01 k}\right)^{2}}{\sum_{r} w_{0 k}}, \\
R_{\mathrm{int}} & =2 \bar{w}_{02}^{2} \frac{\sum_{r} w_{0 k}\left(r_{01 k}-1\right)\left(r_{02 k}-r_{01 k}\right)}{\sum_{r} w_{0 k}} \text { and } \\
G & =\left(\bar{w}_{02}-1\right)^{2} .
\end{aligned}
$$

It should be noted that the measure of change $D$ is always positive, equality being at zero when the two following conditions are met:
(i) absence of nonresponse ( $r=s$ and $w_{1 k}=w_{0 k}$ for all $k$ ),
(ii) absence of poststratification effect on the intermediate weights ( $w_{2 k}=w_{1 k}$ for all $k$ ).

A high nonresponse rate would tend to increase the value of the measure of change $D$ since in such a case, $w_{1 k}$ is generally much larger than $w_{0 k}$.
$R_{01}$ measures the individual weight changes which result from going from the initial to the intermediate set. Later, we will see that the component $R_{01}$ is somehow associated with the quality of the nonresponse model and that a large $R_{01}$ value is preferable. $R_{12}$ measures the individual weight changes which result from going from the intermediate to
the final set. $R_{\text {int }}$ measures the interaction between the two types of change and $G$ measures the change in average weight between the initial and final sets.

In addition to its interpretation as a distance, the measure of change $D$ can also be interpreted as a mean square error of changes $w_{2 k} / w_{0 k}$ in relation to 1 , and in relation to the distribution defined by all the $w_{0 k}$. From this perspective, the component $G$ corresponds to the bias squared (or the square of the difference between the $\bar{w}_{02}$ average of $w_{2 k} / w_{0 k}$ and 1 ), while the sum of the other three components corresponds to the variance. In the simplest case, where a nonresponse adjustment is calculated using a single RHG, and where no postratification is applied, we have $w_{0 k}=N / n$ for all $k \in s$ (in the case of a size $n$ simple random selection) and $w_{1 k}=w_{2 k}=N / m$ for all $k \in r$, (where the nonresponse adjustment factor is $n / m$, i.e., the inverse of the response rate). We then have $D=G=\{(n / m)-1\}^{2}$ and $R_{01}=R_{12}=R_{\text {int }}=0$.

Some significant conclusions may be drawn from looking at the relative importance of $R_{01}, R_{12}$ and $R_{\text {int }}$. If $R_{01}$ is high at the same time that $R_{12}$ is not very high, the survey is one in which the nonresponse adjustment creates significant individual changes in weights, while poststratification only results in a slight change in individual weights. However, when $R_{12}$ is high, poststratification brings about very large individual changes. The results presented in sections 5 and 6 will show that $R_{01}$ can be used to compare the effectiveness of various nonresponse adjustment methods. As well, the sign of $R_{\text {int }}$ indicates whether the two types of individual change are moving in the same direction ( $R_{\text {int }}>0$ ) or in opposite directions ( $R_{\text {int }}<0$ ). In reality, we expect $R_{\text {int }}$ to be very small, if not negligible.

## 4. NONRESPONSE ADJUSTMENT STRATEGIES

The literature contains several methods for adjusting weights (including the method described in section 2.2) to compensate for nonresponse. Another method, which is frequently used in longitudinal surveys, involves adjusting weights in accordance with the inverse of the predicted probability of response obtained through a logistic regression. We also find methods of adjustment based on calibration, which use marginal distributions of the initial sample or of the population. Singh, Wu and Boyer (1995) used this approach in order to derive a method of adjustment capable of producing coherent estimates in longitudinal surveys from one wave to the next. Deville (1998) recommended a method of correction for nonresponse by calibration or balanced sampling. For a review of nonresponse adjustment methods, refer to Kalton and Kasprzyk (1986), Platek, Singh and Tremblay (1978), Chapman, Bailey and Kasprzyk (1986) and to Little (1986). In this document, only methods relying on the creation of RHGs are considered.

### 4.1 Formation of RHGs

In most surveys, aside from a few stratification variables from the sample frame, very little information is available about non-respondents. Therefore, the choice of RHGs is very limited and the strata are often used as RHGs. In these cases, the assumption is that the probability of response is the same for all units in a given stratum. However, in longitudinal surveys, a great deal of information about respondents and non-respondents in the current wave is available from the responses provided in the previous waves. This information can then be used to create RHGs within which the assumption of a uniform response mechanism is plausible. This leads to a better nonresponse adjustment and, therefore, a reduction in the risk of introducing a nonresponse bias into the estimates.

### 4.1.1 Method for the Selection of Variables for the Formation of RHGs

By definition, an RHG is formed from a set of variables capable of predicting the propensity to respond. If the set of variables which is defined at the outset is too large, univariate tests may be used to isolate the most important variables to distinguish the characteristics of respondents from those of nonrespondents. With this set of important variables, a selection method may then be applied for retaining the best variables for explaining the propensity to respond. Two of the current variable selection methods are: the Logistic Regression Model (LR) and the Segmentation Model (SM).

### 4.1.1.1 Logistic Regression

Under the LR method, the combined use of the "fact of having responded to the survey or not" as a dependent variable, standardized weights and the "stepwise" procedure result in a list of the most significant dichotomic variables for explaining the propensity to respond. As a general rule, RHGs are created according to $2^{q}$ possible combinations, based on a set of $q$ explanatory variables used. The LR is often referred to as the symmetrical approach. However, if certain additional constraints are applied when the RHGs are created, this could reduce their numbers. For instance, we could require a minimum number of reference units ( $n$ ) and a response rate (RR) (weighted or not weighted) greater than a certain level in each of the RHGs. Kalton and Kasprzyk (1986) encourage the use of such constraints in order to avoid increasing the variance associated with extreme weights. However, these constraints may reduce the effectiveness of the nonresponse adjustment and result in an increase in the bias. When an RHG does not meet one of these constraints, it has to be combined with another RHG. The combination of RHGs continues until all of the RHGs meet the additional constraints imposed. This leads to $2^{q}-J$ valid combinations, where $J$ represents the reduction resulting from the combination of RHGs.

For instance, in Figure 1, $2^{q}=8$ RHGs are created on the basis of $q=3$ explanatory variables. The shaded boxes in Figure 1 represent the RHGs. An adjustment factor is calculated within each RHG and the weight $w_{0 k}$ of each reference unit is then adjusted, accordingly.

### 4.1.1.2 Segmentation Model

The SM method, which is referred to as nonsymmetrical, is based on the CHAID (Chi-square Automatic Interaction Detection) algorithm developed by Kass (1980). It divides the sample into sub-groups according to the response rate of the explanatory variables by using a Chi-square test. The segmentation process continues until a significant explanatory variable is found. The final sub-groups created through the SM become the RHGs, for which the nonresponse adjustments are calculated. As in the case of the LR, additional constraints may be imposed.

In Figure 1 we see that the SM method divided the sample into several RHGs based on the different explanatory variables. The RHGs are once again represented by the shaded boxes. The segmentation continues until it is no longer possible to find explanatory variables.

### 4.1.2 Nonresponse Adjustment Factor

Whether the RHGs are formed by relying on the LR or the SM, a uniform response mechanism is assumed within each RHG. Thus, the nonresponse adjustment factor is given by the inverse of the response rate (weighted by $w_{0 k}$ or not weighted) for the RHG.

## 5. EMPIRICAL STUDY BASED ON THE SURVEY OF LABOUR AND INCOME DYNAMICS (SLID)

Data from the SLID were used for an empirical study designed to compare the effectiveness of the LR and SM. The SLID is a longitudinal survey of households that started in 1993; one of its objectives is to provide information on the economic well-being of Canadian society (see Lavigne and Michaud 1998).

These two methods were tested through a simulation by analyzing some variables of interest and various domains. The components of the measure of change, the absolute and relative biases and the variances were studied.

### 5.1 Description of the Empirical Study

The first step in the empirical study was to estimate the probability of response to the first wave of the survey for each of the units in the longitudinal sample. Variables which could potentially explain the propensity to respond (based on a preliminary interview) were used to form a very large number of RHGs. All of the individuals in the sample were assigned to an RHG on the basis of the values of the explanatory variables. A probability of response was then estimated for each RHG on the basis of the weighted response rate. Then, only the respondents and their


Figure 1. Depiction of the Formation of RHGs by Method
probability of response were retained in the reference sample for the simulation. Nonresponse was then generated for the reference sample through Poisson sampling. This procedure, illustrated in Figure 2, was independently repeated 100 times, thus creating 100 sets of respondents and non-respondents. The average response rate for each repetition was around $90 \%$, which was the rate observed in the first wave of the SLID.

For each of the 100 repetitions, a nonresponse adjustment was done using the LR method to create the RHGs. Similarly, a nonresponse adjustment was done using the SM to create RHGs for each of the first 20 repetitions. With the SM approach, the number of repetitions was limited to 20 , given the stability of the results and since several manual interventions and the use of a specific software package (in our case: Knowledge Seeker - ANGOSS Software 1995) were required.

Several variants of the variable selection method were studied:
a) LR $\_i$, where $i$ represents, out of the 100 repetitions, the approximate average of the number of RHGs generated through the LR method. In this study, $i=4,16,40,60$. For instance, for LR_40, the $q=6$ most important explanatory variables for the propensity to respond were first identified. The RHGs were then formed using the $\left(2^{q}-J\right)$ valid combinations of these $q=6$ explanatory variables. The imposition of additional constraints ( $n>30$ and RR>50\%) in each RHG led to the re-grouping of some RHGs. On average, out of 100
repetitions, 24 RHGs had to be regrouped ( $J=24$ ) and a total $2^{q}-J=2^{6}-24=40$ RHGs were formed, hence the LR_40 designation. In the simulation study, LR $i$, where $i=4,16,40,60$ RHGs corresponds, respectively to $q=2,4,6,8$ explanatory variables.
b) $\mathrm{SM}_{\_} i$, where $i$ indicates the approximate average in the first 20 repetitions of the number of RHGs generated through the SM method. In this study, $i=16,25,40$. For example, for SM_16, one SM was used with a significance level $p$ of 0.0001 . After the imposition of the same additional constraints as for the LR, an average 16 RHGs were created. SM $i$, where $i=16,25,40$ RHGs corresponds, respectively, to the significance levels of $0.0001 ; 0.0005 ; 0.0025$. The higher the level used, the easier it is to identify the significant differences, which makes it possible to achieve a more detailed segmentation and, hence, a greater number of RHGs.
c) A method with a single RHG (1_RHG) was also used for comparison purposes. This method involves defining the entire sample as a single RHG for each of the 100 repetitions. It should be noted that this method is only effective if the response mechanism is uniform within the entire sample, which is rarely the case.
At first, the initial weights were normalized so that $\Sigma_{s} w_{0 k}=N$, in order to eliminate the effect of undercoverage and to better isolate the effect of nonresponse. Thus, $G$ will only measure the average change caused by the nonresponse adjusment.


Figure 2. Illustration of the Simulation Process

Once the initial weights were normalized, each set of final weights was then the result of a two step process: a nonresponse adjustment (based on one of the eight methods mentioned: 1_GRH, LR_ $i$, where $i=4,16,40,60$ and SM $_{i} i$, where $i=16,25,40$ ) and a same poststratification ( 14 age-sex groups by province).

### 5.2 Analysis of the Results of the Empirical Study

For each of the methods discussed in the previous section, the components of the measure of change $D$ were studied. Also, the average, absolute and relative nonresponse bias and the average variance of the estimates were analyzed.

### 5.2.1 Measure of Change (D)

Table 1 presents the average value of $D$ and its components for each of the $M$ repetitions (where $M=100$ for the LR and $M=20$ for the SM ) as well as the percentage contribution of each element to the average value of $D$. We observe, in the first place, that for the $1_{-}$GRH method, $R_{01}$
is nil since one single nonresponse adjustment was made to the set of respondents. Thus, $w_{1 k}=\alpha w_{0 k}$, where $\alpha$ is a constant, so $r_{01 k}=1$ for every $k \in r$ and $R_{01}=0$. We also observe that $D$ increases as the number of RHGs increases, irrespective of whether the LR or SM method is used. Thus, the more RHGs there are to compensate for nonresponse, the greater the total change to which the weights are subjected. In addition, the values of $D$ are higher for the SM than for the LR.

For the LR and the SM, the contribution of $R_{01}$ to the measure of change increases as the number of RHGs increases, since nonresponse is more readily targeted as the number of RHGs increases. Consequently, the nonresponse adjustment often becomes more important and, thereby, the weights vary more and more. In addition, the contribution of $R_{01}$ to the measure of change is much more important with the SM than with the LR. This indicates that the SM seems to be better at modeling nonresponse and isolating the specific trends of the LR.

Table 1
Average Value of $D$ on Repetitions, for each Component and their Contribution (as a \%) to the Measure of Change for each of the Eight Nonresponse Adjustment Methods

| Method | $D$ | $R_{01}$ <br> $\left(\times 10^{-3}\right)$ | $R_{00} / D$ <br> $(\%)$ | $R_{12}$ <br> $\left(\times 10^{-3}\right)$ | $R_{12} / D$ <br> $(\%)$ | $R_{\text {int }}$ <br> $\left(\times 10^{-5}\right)$ | $R_{\text {int }} D$ <br> $(\%)$ | $G$ <br> $\left(\times 10^{-2}\right)$ | $G / D$ <br> $(\%)$ |
| :--- | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| 1_RHG | 0.012135 | 0.00 | 0.00 | 1.17 | 9.66 | 0.00 | 0.00 | 1.11 | 90.34 |
| LR_4 | 0.012952 | 0.78 | 6.04 | 1.10 | 8.49 | 0.06 | 0.01 | 1.11 | 85.46 |
| LR_16 | 0.013809 | 1.66 | 11.97 | 1.00 | 7.31 | 3.76 | 0.54 | 1.11 | 80.19 |
| LR_40 | 0.014426 | 2.32 | 16.02 | 0.96 | 6.66 | 4.02 | 0.55 | 1.11 | 76.77 |
| LR_60 | 0.014948 | 2.85 | 19.00 | 0.95 | 6.35 | 3.75 | 0.49 | 1.11 | 74.15 |
| SM_16 | 0.015712 | 3.42 | 21.33 | 0.97 | 6.19 | 3.40 | 0.43 | 1.11 | 72.05 |
| SM_25 | 0.016713 | 4.44 | 26.02 | 0.95 | 5.73 | 2.95 | 0.36 | 1.11 | 67.89 |
| SM_40 | 0.018202 | 5.97 | 32.37 | 0.95 | 5.23 | 1.20 | 0.14 | 1.11 | 62.26 |

As for $R_{12}$, it is almost constant, regardless of which method and number of RHGs are used. However, despite the fact that it changes very little, its contribution to the measure of change diminishes as the number of RHGs increases. This is due to the fact that there is more variation in the weights with a nonresponse adjustment, and the modifications which poststratification creates in the weights are less and less important as the number of RHGs increases.

In the case of $R_{\text {in }}$, its value is negligible and its contribution to the measure of change is very small. This means that the interaction between the nonresponse adjustment and poststratification is practically nil.

Finally, $G$ remains constant, irrespective of which method and how many RHGs are used. As with $R_{12}$, the contribution of $G$ to the measure of change diminishes as the number of RHGs increases. A larger number of RHGs is better at targeting nonresponse, thereby causing more variations in the set of intermediate weights.

Since, with all of these methods, $G$ is constant, $R_{\text {int }}$ is close to zero and $R_{12}$ is nearly constant, it is clear that the variations in $D$ are mostly influenced by the variations in $R_{01}$.

Graph 1 shows the average contribution in percentage of $R_{01}$ and $R_{12}$ to the measure of change. For LR and SM, the contribution of $R_{01}$ increases with the number of RHGs while that of $R_{12}$ diminishes. Also, the contribution of $R_{01}$ is greater for SM than for LR, while that of $R_{12}$ is less for SM than for LR. The profile of the contribution of $R_{01}$ is the same as the profile of $D$ (Table 1). This confirms that the variations in the measure of change are mainly due to the variations in $R_{01}$.

Graph 2 shows the comparison between the LR and SM in terms of the average percentage contribution in percentage of $R_{01}$ to $D$. For a given number of RHGs, $R_{01}$ contributes to a larger percentage of $D$ through the SM method than through the LR method. This means that individual changes in the weights between the initial and intermediate sets are greater for SM than for LR.



### 5.2.2 Relative and Absolute Biases

The Relative Bias (RB) and the Absolute Bias (AB) were used to compare the performance of LR relative to SM in reducing the nonresponse bias:

$$
\mathrm{RB}_{\mathrm{i}}=100\left(\frac{\hat{Y}_{i}-Y}{Y}\right) \text { and } \mathrm{AB}_{\mathrm{i}}=\hat{Y}_{i}-Y ;
$$

where $\hat{Y}_{i}$ is the estimate of the variable of interest obtained for the $i$-th repetition, $i=1,2, \ldots, M, M=100$ for the LR, $M=20$ for the SM and $Y$ is the total for the variable of interest obtained from the reference sample.

The Average Relative Bias (ARB) and the Average Absolute Bias (AAB) are calculated by taking, respectively, the average of the RB and the AB for all repetitions:

$$
\mathrm{ARB}=\frac{1}{M} \sum_{i=1}^{M} \mathrm{RB}_{\mathrm{i}} \text { and } \mathrm{AAB}=\frac{1}{M} \sum_{i=1}^{M} \mathrm{AB}_{\mathrm{i}}
$$

where $M=100$ in the case of the LR and $M=20$ in the case of the SM.

For the 100 repetitions, national estimates were produced for the following three variables: "person living, or not, in a family whose revenue is less than the Low Income Cutoff (LICO)", "Individual Total Income (TD" and "Individual Wages and Salaries (WS)". The ARB for each estimate was calculated for the eight methods under study. Given the large sample size, the low nonresponse rate ( $10 \%$ ) and the fact that a large number of control totals was used for poststratification, the ARB is very small (see Table 2) for each of the methods used.

In Table 2 we see that, for each of the three variables, the ARB is more or less constant for the SM, irrespective of how many RHGs are used. Also, for the LR, the ARB for the TI and SW is more or less constant not withstanding the number RHGs used. On the other hand, for the LICO, the ARB for method LR_4 is much smaller than the ARB for the other three LR methods. This could be due to the fact that the LICO is a variable derived from several other variables, unlike the TI and the SW, which are observed variables. The ARB for the three variables for method 1_RHG is much larger than the ARB produced by the SM and the LR, except for the LICO, since in this case the ARB is more or less equivalent to the ARB of the LR. Thus, it appears that method 1_RHG does not perform as well as the SM and the LR. In the best case, it is more or less equivalent to LR. Unlike SM, we observe that the progression
of ARB is not strictly downwards for the LR, as the number of RHGs increases.

Despite the fact that the ARB is minimal for the variables studied for Canada, it can increase rapidly for small domains. In this study, other domains were also reviewed. Although some variances were observed in several of these cases, it seems that the ARB for the SM is generally smaller than the ARB for the LR and the method 1_RHG. A more detailed study of a larger number of interest and domain variables would be beneficial for corroborating these conclusions.

As previously indicated, the individual changes in the weights caused by the nonresponse adjustments are greater for the SM than for the LR (see Graph 2). This would suggest that the SM is more effective in reducing the nonresponse AB for a fixed number of RHGs. Graph 3 confirms this observation, showing that the AAB for the LICO is smaller through the SM than through the LR method.

### 5.2.3 Variance Estimates

Variance estimates were produced for the three variables of interest through the Jackknife method. For LICO (Graph 4), the average variance of estimates is approximately the same, regardless of the method used. However, there is a slight decrease when the number of RHGs increases, for both the LR and the SM. Also, based on the empirical study, average variance estimates for the SM are slightly smaller than for the LR. Therefore, the larger dispersion in the weight (a higher value for $D$ ) does not entail an increase in variance.

## 6. APPLICATION TO THE NATIONAL LONGITUDINAL SURVEY OF CHILDREN AND YOUTH (NLSCY) DATA

In this section, most of the analyses done with the help of the LR and SM in the empirical study with data from the SLID are reproduced with the information obtained from the NLSCY. Just like the SLID, the NLSCY is a longitudinal survey of households. It started in 1994 and is designed to collect information for analyzing policies and developing programs addressing critical factors affecting the development of children in Canada (see Michaud, Morin, Clermont and Laflamme 1998).

Table 2
ARB (as a \%) for Different Variables Based on the Methods - Canada

| Variable | STUDIED METHOD |  |  |  |  |  |  |  |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
|  | 1_RHG | LR_4 | LR_16 | LR_40 | LR_60 | SM_16 | SM_25 | SM_40 |
| LICO | 0.37 | 0.15 | 0.43 | 0.37 | 0.31 | 0.14 | 0.12 | 0.08 |
| TI | -0.32 | -0.09 | -0.06 | -0.05 | -0.06 | -0.006 | -0.005 | 0.002 |
| WS | -0.44 | -0.13 | -0.15 | -0.19 | -0.14 | -0.10 | -0.09 | -0.09 |




### 6.1 Description and Analysis of the Results of the Application

The following methods were used for this study: LR $i$, where $i=4,14,41,70$ with, respectively $q=2,4,6,8$ variables, and SM_i, where $i=19,36$ with significance levels of 0.001 and 0.005 , respectively. The same two constraints imposed for the SLD were re-applied when the RHGs were created. The same poststratification was used ( 22 age-sex groups by province) for each of the methods under study.

Unlike the empirical study based on the SLID, only the data collected in the first two waves of the NLSCY were used. There was no simulation and the initial weights were not normalized ( $\Sigma_{s} w_{0 k}=\hat{N}<N$ ). It should be noted that the undercoverage of the NLSCY is around $13 \%$ and its nonresponse is around $8 \%$.

The conclusions drawn from the results presented in Table 3 are similar to those obtained in the simulation
(Table 1). However, we observe that the relative contribution by $R_{01}$ to the measure of change is weaker for the NLSCY than for the SLID. This result indicates that the nonresponse adjustment of the SLID produces larger individual changes in the weights, thereby resulting in a larger contribution by $R_{01}$. Therefore, the nonresponse adjustment in the case of the NLSCY had no significant effect on the individual changes in the weights, contrary to what was observed in the case of the SLID.

The relative contribution by $R_{12}$ to the measure of change is higher for the NLSCY than for the SLID. This result indicates that the more refined poststratification of the NLSCY results in greater individual changes in the weights, which translates into a greater contribution of $R_{12}$. Therefore, the NLSCY benefits a great deal from poststratification, which is less important for the SLID.

Table 3
Value of $D$, for each Component, and of their Contribution (as a \%) to the Measure of Change for each of the Six Nonresponse Adjustment Methods

| Method | $D$ | $R_{01}$ | $R_{01} / D$ <br> $(\%)$ | $R_{12}$ | $R_{12} / D$ <br> $(\%)$ | $R_{\text {int }}$ <br> $\left(\times 10^{-4}\right)$ | $R_{\text {in }} / D$ <br> $(\%)$ | $G$ | $G / D$ <br> $(\%)$ |
| :--- | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| LR_4 | 0.1475 | 0.0052 | 3.51 | 0.0369 | 25.05 | -4.63 | -0.31 | 0.1058 | 71.76 |
| LR_14 | 0.1497 | 0.0075 | 5.00 | 0.0367 | 24.69 | -5.50 | -0.37 | 0.1058 | 70.68 |
| LR_41 | 0.1530 | 0.0112 | 7.29 | 0.0369 | 24.13 | -9.16 | -0.60 | 0.1058 | 69.18 |
| LR_70 | 0.1564 | 0.0144 | 9.21 | 0.0362 | 23.13 | -0.19 | -0.01 | 0.1058 | 67.67 |
| SM_19 | 0.1608 | 0.0187 | 11.63 | 0.0371 | 23.07 | -8.24 | -0.51 | 0.1058 | 65.81 |
| SM_36 | 0.1640 | 0.0220 | 13.41 | 0.0373 | 22.76 | -11.30 | -0.69 | 0.1058 | 64.52 |

With respect to $R_{\text {innt }}$, as with the SLID, its contribution to the measure of change is negligible. Contrary to the SLID, the sign of $R_{\text {int }}$ is negative, which means that the interaction between $R_{01}$ and $\mathrm{R}_{12}$ is negative.

With respect to $G$, as in the case of the SLDD, it is the key source of contribution to the measure of change. In the case of the NLSCY, $G$ not only includes the average change in weight resulting from the nonresponse adjustment, but also the average change in weight resulting from the correction for undercoverage through poststratification.

When all of these results are compared, it becomes evident that the two surveys are very similar since $R_{\text {int }} \approx 0$ and the sum of the contributions to the measure of change of $R_{01}$ and $R_{12}$ is around $35 \%$ in both cases. However, the NLSCY is also very different from the SLID since $R_{12}$ predominates in the former one, while $R_{01}$ predominates in the latter.

Just as with the SLID, $D$ increases with the number of RHGs and this measure is greater for the SM than for the LR. In fact, the value of $D$ is greater for the NLSCY than for the SLDD, mainly because of the NLSCY under-
coverage, which results in an increase in $G$ and, therefore, in $D$.

The average contribution of $R_{01}$ for the LR and the SM increases with the number of RHGs, whereas that of $R_{12}$ diminishes (Graph 5). The contribution of $R_{01}$ is also greater for the SM than for the LR, unlike the contribution of $R_{12}$, which is smaller for the SM than for the LR.

As was observed with the empirical study, the profile of the contribution of $R_{01}$ to the measure of change is the same as that of the measure itself. This shows that the variations in $D$ depend directly on $R_{01}$.

Graph 6 enables us to compare the LR and the SM, presenting the average contribution of $R_{01}$, to the measure of change for the methods with an essentially equivalent number of RHGs. As with the SLID, the results indicate that nonresponse seems to be better targeted with the SM than with the LR method.

Unlike the SLID simulation study, the bias was not evaluated since no external source of data was available for evaluation purposes.

Graph 5 : Average contribution of $R_{01}$ and $R_{12}$ to the measure of change ( $D$ ) for each method



## 7. CONCLUSION

This document highlights the fact that the choice of RHGs and method for defining them depends on the: i) availability of ancillary information, ii) need to reduce the nonresponse bias for all estimates, and iii) time and operational constraints. The empirical study, as well as the NLSCY data, showed that the SM method appears to be better than the LR one in reducing the nonresponse bias. The results also demonstrated that the proposed measure of change can be a very useful tool for comparing different weighting strategies.

In particular, it would appear that, as the value of $R_{01}$ increases, the reduction of the bias obtained from using RHGs increases. Given the difficulty in obtaining a reliable estimate of the nonresponse bias in a survey, the relationship identified between the size of $R_{01}$ and the decrease in the bias suggests that $R_{01}$ should be used as a tool for evaluating nonresponse adjustment methods. This requires that $R_{01}$ first be determined for different RHG sets. Then, the set with the highest $R_{01}$ value is likely to be more effective than the other alternatives in reducing the nonresponse bias for most of the variables of interest.

The measure of change presented could also be used to compare the different calibration strategies. In this case, the nonresponse adjustment could remain the same for all of the poststratification methods under study. A detailed study of the behaviour of $R_{12}$ could be done and would no doubt lead to certain conclusions, as this study did about $R_{01}$. This type of study would not necessarily have to be restricted to the longitudinal context but could quite readily be done with a cross-sectional study. Also, the measure of change could be useful in evaluating different nonresponse adjustment methods in cross-sectional surveys.

## ACKNOWLEDGMENTS

The authors would like to thank M. Hladky, M. Latouche, C. Nadeau and N. Tremblay for their important contributions to this project.

## REFERENCES

ANGOSS SOFTWARE (1995). Knowledge Seeker IV for Windows - User's Guide. ANGOSS Software International Limited.

CHAPMAN, D.W., BAILEY, L. and KASPRZYK, D. (1986). Nonresponse adjustment procedures at the U.S. Bureau of the Census. Survey Methodology, 12, 161-179.
DEVILLE, J.-C. (1998). La correction de la non-réponse par calage ou par échantillonnage équilibré. 1998 Proceedings of the Survey Methods Section of the Statistical Society of Canada, 103-110.
DEVILLE, J.-C., and SÄRNDAL, C.-E. (1992). Calibration estimators and generalized raking techniques in survey sampling. Journal of the American Statistical Association, 87, 418, 376-382.
KALTON, G., and KASPRZYK, D. (1986). The treatment of missing survey data. Survey Methodology, 12, 1-16.
KASS, G.V. (1980). An exploratory technique for investigating large quantities of categorial data. Applied Statistics, 29, 119-127.
LAVIGNE, M., and MICHAUD, S. (1998). General aspects of the Survey Labour and Income Dynamics. SLID research paper, Statistics Canada, catalogue number 98-05.
LITTLE, R.J.A. (1986). Survey nonresponse adjustments for estimates of means. International Statistical Review, 54, 137-139.

MICHAUD, S., MORIN, Y., CLERMONT, Y. and LAFLAMME, G. (1998). Issues in the design of a survey to measure child development: The Experience of the Canadian National Longitudinal Survey of Children and Youth. Statistics Canada, Internal document.
PLATEK, R., SINGH, M.P. and TREMBLAY, V. (1978). Adjustment for nonresponse in surveys. Survey Sampling and Measurement. N.K. Namboodiri, Ed. Academic Press, 157-174.

RIZZO, L., KALTON, G. and BRICK, M. (1996). A comparison of some weighting adjustment methods for panel nonresponse. Survey Methodology, 22, 43-53.
SINGH, A. C., WU, S. and BOYER, R. (1995). Longitudinal survey nonresponse adjustment by weight calibration for estimation of gross flows. Proceedings of the Section on Survey Research Methods of the American Statistical Association, 396-401.

# Sampling and Weighting a Survey of Homeless Persons: A French Example 

PASCAL ARDILLY and DAVID LE BLANC ${ }^{1}$


#### Abstract

In 2001, the INSEE conducted a survey to better understand the homeless population. Since there was no survey frame to allow direct access to homeless persons, the survey principle involved sampling the services they received and questioning the individuals who used those services. Weighting the individual input to the survey proved difficult because a single individual could receive several services within the designated reference period. This article shows how it is possible to apply the weight sharing method to resolve this problem. In this type of survey, a single variable can produce several parameters of interest corresponding to populations varying with time. A set of weights corresponds to each definition of parameters. The article focuses, in particular, on "an average day" and "an average week" weight calculation. Information is also provided on the use data to be collected and the nonresponse adjustment.


KEY WORDS: Weight sharing; Incomplete frame; Homeless persons.

## 1. INTRODUCTION

In 2001, INSEE conducted a survey to better understand the homeless population. This was the first representative survey of this type in France (A survey of this type was conducted in the United States in 1991 by Research Triangle Institute (RTI) in the Washington metropolitan area (RTI 1993)). The survey principle was to reach homeless persons through the services provided to them, specifically, overnight accommodation and meals. Obviously, a person could use one or more of the services of the survey frame during the reference period considered, which creates a problem when weighting the survey's individual data files. In this article, we will show how the weight sharing method can be applied to this problem. In this type of survey, unlike most traditional household surveys, a single variable can produce several parameters of interest corresponding to different population concepts: the ones used most often by practitioners are the " average day" and "average week" parameters. A set of weights corresponds to each definition of parameters. We will provide precise definitions of these concepts and will focus in particular on the practical calculation of the corresponding weights. The article is laid out as follows: we will begin by stating the objectives of the survey, identifying its reference population and describing its sample design. We will then introduce the parameters of interest and derive the estimators of these parameters using the weight sharing method. We will describe the practical application of "average day" and "average week" weight calculations. Lastly, we will discuss practical considerations related to the nonresponse adjustment.

## 2. "HOMELESS" SURVEY

### 2.1 Objectives of the Survey

The purpose of the survey conducted by the INSEE in February 2001 was to obtain a better understanding of the "homeless" population. This population is normally defined by default as all persons who do not have a fixed residence. It is a population that is not captured by traditional household surveys conducted by the Institut since such surveys have an accommodation survey frame. Since there was no sampling frame for this population, the survey principle involved reaching the target population through the services provided to persons in difficulty, specifically accommodation and meals. These service are provided at certain times that vary depending on their nature: meals are provided every day at noon and in the evening, while overnight accommodation is provided once a day.

This indirect sampling introduces two biases into the population initially targeted and the population actually surveyed. First, the entire target population is not surveyed: only those members who use the services in the survey field are potentially sampled. Second, the population actually surveyed contains individuals who do not belong to the population initially targeted to the extent that the services provided primarily for homeless persons are also used by persons who live in a regular household but who are in a vulnerable situation (this is especially true in the case of meals). Throughout this article, while keeping this distinction in mind, we will however sometimes use the expression "homeless" to designate the persons using the services in the survey field.

[^8]
### 2.2 Reference Population

The main feature of the services surveyed is that they are provided in specific locations; this location is accordingly called a centre. Several types of services correspond to a given centre. The statistical unit sampled, which we will call a service, will be defined as a quadruplet (service, day, time interval, person): it consists of a given type of service in a given centre, on a given day, in a given interval of time, to a given person. Of course, a person could receive several services on the same day, let alone in a given week or during the survey month.

The survey reference period covers one month (January 15 to February 15, 2001). The total number of days in the survey reference period is designated as $J$, denoted by the index $j$.

The geographic field of the survey is that of population centres with more than 20,000 inhabitants.

The services in the survey field are those that are provided by one of the two types of services retained meals and accommodation - when they are provided at least one day during the survey reference period.

The reference population, designated as $P(J)$, consists of persons who receive at least one service in the survey field during the reference period.

This population of interest depends fundamentally on the reference period. Its size increases with the length of that period, but "more slowly" than the time: in actual fact, certain people are found in the centres every day. In reality, the change in $P(J)$ in relation to $J$ is complex because there are two separate phenomena coming into play that would appear to have different characteristic times:

- at any given time, the "homeless population" only occasionally visits the centres in the frame: to claim to cover that population, it would be necessary to survey over a period of time that would ensure that all persons in this population had used the services at least once (this period is not known but it is acknowledged in France, "according to the experts", that the population not covered during one full month of winter is negligible).
- the "homeless" population is self-renewing over time. Year to year, there are no doubt numerous persons coming into and going out of this population, linked to demographic change or economic or structural changes in society (persons coming into and going out of vulnerable situations).

The question of how to determine $J$ ultimately comes down to knowing whether interest is mainly in a concept of homeless "at a given moment" ( $J$ is relatively short) or a concept of homeless over a long period of time ( $J$ relatively long). The approach adopted by the INSEE is a compromise between the two.

### 2.3 Sample Design

The survey's sample design has three stages: selection of population centres, selection of centres and time intervals, and selection of services.

### 2.3.1 Selection of Population Centres

The first stage of the sample design consists of selecting the population centres, based on a size criterion defined as a combination of the population of the population centres and the ability to provide services so that they could be identified in the records of associations and of the Ministère de la Santé. This first selection stage was carried out several months before the other two. This screening was necessary because the exhaustive census of the centres and the data related to them (type of service provided, average capacity, days open, ...) was then carried out in the selected population centres. This operation was done twice: a detailed survey the year before the data collection and an update just before the start of the data collection. This process produced a survey frame of centres. This frame has a fundamental role: persons who used only non-identified centres were not be sampled.

### 2.3.2 Selection of Centres, Days and Time Intervals

For practical reasons, it was not possible to survey all of the centres and to keep an interviewer on site at a given centre the entire day. Nor was it possible to interview everyone in a centre. It was therefore imperative to sample:

- centres in the selected population centres (index $c$ )
- survey days during the collection period (index $j$ )
- intervals of time during the survey days (index $t$ ).
- persons within one of the selections (centre, day, time interval).
For theoretical reasons, time intervals were defined in such a way that an individual could not receive two different services during a single time interval (for example, one of these time intervals was the period from 11:00 a.m. to 2:00 p.m.). It was not reasonably possible to measure the links to the survey frame unless the persons interviewed could easily identify in time and space the services they received during the survey period. In the case of centres offering meals, one time interval covered the noon meal and one time interval covered the evening meal. It was assumed that an individual could use only one centre during the time interval corresponding to the noon meal, otherwise it would be necessary to ask the individual if he had already received a meal somewhere else or if he had eaten twice in the same centre. It was also determined that the length of an interval ensuring use of only one service was also the length of time that an interviewer could reasonably be asked to remain on site interviewing (two to three hours maximum). (Note that daytime accommodation is not part of the services included in the survey field. This restriction of the field reflects two concerns. First, it would be very difficult to divide the day into time intervals of
three or four hours and to determine the links using this breakdown (the memory effort required of the person interviewed would be significant and did not seem reasonable to the survey's designers). Second, it is very difficult to predict the use of these services. We wanted to avoid having a team of interviewers go to a site and not be able to conduct any interviews because of lack of use.)

In actual fact, there is no fundamental difference between the sampling of the centres and the sampling of the periods of time: the relevant units to be considered are the triplets ( $c, j, t$ ) that correspond to the overlap between a centre, a day and a time interval. Some of the boxes in the "time" and "centres" cross-tabulation table can be eliminated automatically prior to the selection, either because the centre is closed during the time slot considered, or because there is clearly not enough use. (In the latter case, caution must be exercised with respect to the possible restriction of the field should it be found that persons use only this centre and only attend during this time slot. If the latter are atypical, biases will be introduced into the estimations.)

The selection method used was a random selection of the triplets (centres, days, time intervals) in proportion to the size of the centres obtained during the centre census. (In practice, in order to avoid difficulties with centre officials, time intervals were grouped together when a centre was sampled more than four times during the survey period.) Centres were stratified by type. (For accommodation services, centres were stratified by the criteria of men only/women only/mixed accommodation.) However, since this "precautionary" stratification does not apply directly to the observation units, it is useful only if the behaviour of the individuals differs significantly by the type of centre in which they are found.

### 2.3.3 Selection of Services

This last stage of the sample design consisted in completing the sampling of services, that is, in selecting individuals in a selected centre on a given day during a given time interval. The data collected during the census of the centres were not generally enough to constitute a survey frame of services. Some accommodation centres had lists: this was the more positive scenario where persons could be selected using these lists. However, at the majority of centres (for example, a soup kitchen), it was not even known how many people would show up in a given time interval: it was therefore not possible to develop a survey frame of services. Sampling of the services was done on an equal probabilities basis. As is traditional in multiple stage surveys, selecting a constant number of services (last stage) ensures constant probabilities of selection and thereby limits the risk of expanding sampling variances.

In practice, the selection method used varies from one type of centre to the next, depending on the topography of the sites; existing list, waiting list, arrivals spaced over time, population "grouped" in no order at a single site at the same time, etc. It also takes into consideration the
maximum number of interviews that can reasonably be done by the interviewer or interviewers during the survey's time interval, and the fact that it is not desirable to keep the sampled persons too long after the closing of a centre or after meal service has stopped because of the risk of increasing the nonresponse rate.

In all instances, a "counter" counts the number $N$ of services provided during the sampling period. This is crucial to determining the selection probability of the sampled services. At the same time, the counter carries out a standard systematic selection (ideally, the selection should be done by another person (or "sampler") to avoid measurement errors in the use: For budget reasons, it was not possible to resolve this problem) using the following method:

- in centres where a list was available, $n$ services were selected, $n$ being set before the survey;
- in centres without a list, services were selected with a fixed $f$ sampling ratio. $f$ is determined based on the number of expected services $\tilde{N}$ and the number of services that we wanted to sample $\tilde{n}$ in order to ensure equal selection probabilities. In these cases, the size of the sample was not known in advance.


## 3. PARAMETERS OF INTEREST

The quantities of interest are essentially totals or ratios. We want to estimate a total in relation to a variable $y$ defined for the population $P(J)$,

$$
\begin{equation*}
Y_{J}=\sum_{k \in P^{\prime}(J)} y_{k} . \tag{1}
\end{equation*}
$$

One specific example of these totals is the size of $P(J)$, $N_{J}=\operatorname{card}(P(J))=\Sigma_{k \in P(J)} 1$.

We also want to estimate the average of $y$ in the reference population,

$$
\begin{equation*}
\bar{Y}_{J}=\frac{Y_{J}}{N_{J}}=\frac{1}{N_{J}} \sum_{k \in P(J)} y_{k} \tag{2}
\end{equation*}
$$

For example, $y$ can be the nationality of the individual, the age at which he completed his education, or the number of centres that he visited the day of the interview.

We then have to distinguish between two types of variables:

- variables that are fixed during the survey reference period (such as, age at time of completion of education);
- variables that vary during the survey reference period ( $y_{k}=y_{k}(j)$ ). The number of centres visited on the day of the survey fall into this category.

We will begin with the variables that are fixed during the survey reference period. Section 6 looks briefly at those variables that change during that period.

## 4. ESTIMATION OF A TOTAL OR RATIO IN CASES WHERE THE VARIABLE OF INTEREST IS CONSTANT DURING THE SURVEY PERIOD

For the convenience of the discussion, we will not present explicitly all of the selection stages. Instead, we will use as an example a population centre sampled at the first selection stage.
We note:
$C$ : all centres in the population centre open at least one day during the survey period, denoted by index $c$
$\Pi_{c, j, t}$ : all services provided in centre $c$ on day $j$ during time interval $t$, denoted by index $i$.
$\Pi_{j, t}:$ all services provided in the population centre on day $j$ during time interval $t$.
$P_{c, j, t}:$ all persons who visit centre $c$ on day $j$ during time interval $t$, denoted by index $k$.
$P_{j, t}$ : all persons who visit a centre in the population centre on day $j$ during time interval $t$.

Based on the definition of the time intervals, we find that for each individual $k \varepsilon P_{j, \text {, }}$, there is one and only one service $i$. Thus, there is a one-to-one correspondence between $P_{j, t}$ and $\Pi_{j, r}$. In other words, for every couple ( $j, t$ ), the $P_{c, j, t}$ are separate. On the other hand, $P_{c, j, t}$ and $P_{c \cdot, j, t}$. can have a non-empty intersection, when $t \neq t^{*}$.

The population of interest is therefore written

$$
P(J)=\bigcup_{c, j, t} P_{c, j, t}=U\left(\underset{c \varepsilon C}{\amalg} P_{c, j, t}\right) .
$$

The central point of the reasoning consists in expressing the total of one variable of the population of individuals (which is our total of interest) as the total of another variable of the population of services (which are the sampled units), since estimation of the latter does not pose any particular problem. To obtain this result, we can use direct reasoning or apply the weight sharing method, either of which may seem more natural.

Using direct reasoning; we define the application $K$, which links to each service $i$ received during reference period $J$ in all of the centres in the survey frame the individual who received that service.

$$
K: \underset{i \rightarrow K(i)}{\{\text { services }\} \rightarrow\{\text { individuals }\} .}
$$

The population of interest $P(J)$ is represented by $K$ of $\Pi(J)$, all services provided during the reference period in
all centres in the survey field. For each $k \varepsilon P(J)$, we define $r_{k}(J)=\operatorname{card}\left(K^{-1}(k)\right)$, the number of services provided to individual $k$ during period $J$ in all centres in the survey field, which we will also call the "number of links".

This gives us the fundamental equation:

$$
\begin{equation*}
Y_{J}=\sum_{k \in P(J)} y_{k}=\sum_{i \in \Pi(J)} \frac{y_{K(i)}}{r_{K(i)}^{(J)}} \tag{3}
\end{equation*}
$$

Since variable $y$ takes the same value for all services $i$ "pointing" to individual $k$, such that $K(i)=k$, the right-hand side can be written

$$
\sum_{k \in P(J)}\left[\sum_{i \in \Pi(J): K(i)=k} \frac{y_{k}}{r_{k}(J)}\right]=\sum_{k \in P(J)} \frac{y_{k}}{r_{k}(J)}\left[\sum_{i \in \Pi(J) ; K(i)=k} 1\right]
$$

But the quantity in the square brackets is the number of services provided to individual $k$ during period $J$, or $r_{k}(J)$, which proves the equation.

We can then see $y_{K(i)}$ as attached to corresponding service $i$ and write $y_{i}$ in place of $y_{K(i)}$, and $r_{i}(J)$ in place of $r_{K(i)}^{(J)}$. By using $z_{i}=y_{i} / r_{i}(J), Z=\Sigma_{i \mathrm{En}(J)} z_{i}$, we get $Z=Y_{J}$.

Formula (3) is none other than the weight sharing formula. The above reasoning is actually the reasoning underlying this method. (Only the expressions change; the weight sharing method describes the links between the sampled population and the population of interest by a matrix rather than an application, a single unit of the sampled population being able to "point" to several units of the population of interest.) The principle of this latter method is set out in Appendix 1.

### 4.1 Estimation of a Total

Let us now assume that we have a sample $s_{\Pi}$ of services to which a set of weights is linked $\left(w_{i}\right)_{i s J_{n}}$. We assume these weights are unbiased (this is the inverse of the probabilities of inclusion of services in the sample). $s_{\Pi}$ implicitly defines a sample of individuals $s_{p p}$, which is actually all of the individuals who receive the sampled services. The weight sharing formula (see Appendix 1) ensures that the estimator

$$
\hat{Y}_{J}=\sum_{s_{p}} y_{k} \tilde{w}_{k}
$$

is unbiased, where we write for every $k \varepsilon s_{P}$ :

$$
\begin{equation*}
\tilde{w}_{k}=\frac{1}{r_{k}(J)} \sum_{s_{\Pi} ; K(i)=k} w_{i} . \tag{4}
\end{equation*}
$$

Formula (4) simply states that an individual's weight is equal to the sum of the weights of the services that were used to "catch him", divided by the number of links with the survey frame, $r_{k}(J)$. In this way, it is possible to work directly on the individuals sampled: for each individual $k$, we calculate the weight $\tilde{w}_{k}$, and we estimate the total $Y_{J}$ by $\hat{Y}_{J}$.

Figure 1 gives a fictitious sampling example. The service universe contains 13 services, provided to 8 persons. 6 services are sampled. The sample of individuals contains 5 persons, individual number 2 having been "caught" by two different services. Using formula (4), the weights of the individuals sampled will be equal to:

$$
\tilde{w}_{1}=w_{1}, \tilde{w}_{2}=\frac{1}{2}\left(w_{2}+w_{8}\right), \tilde{w}_{3}=w_{10}, \tilde{w}_{6}=w_{7}, \tilde{w}_{7}=\frac{1}{3} w_{9} .
$$



Figure 1. The arrows represent the links between the services and the individuals. The shaded services were sampled. They point to shaded individuals. Dotted lines represent the links reported by individual 7, which were not used to include the individual in the sample.

If the services all have the same weight equal to $13 / 6$ (for example, if the services had been selected by simple random sampling), the number of persons having used services during the survey is estimated by:

$$
\hat{Y}_{J}=\sum_{s_{p}} \tilde{w}_{k}=\frac{13}{6}\left[1+\frac{1}{2} \cdot 2+1+1+\frac{1}{3}\right]=\frac{169}{18} \approx 9.39 .
$$

In this case where the variable being considered does not vary during the survey period, identifying the persons using the services does not affect the estimator bias. Consider an individual "caught" by two different services with weights $w_{1}$ and $w_{2}$. In practice, this could produce two cases:

- it is determined that this is the same individual; the weighting associated with this individual will be equal to $\left(w_{1}+w_{2}\right) / r_{k(J)}$, and the expression corresponding to the individual in the estimator will be equal to $y_{k}\left(w_{1}+w_{2}\right) / r_{k(J)}$.
- it is not determined that the individual has already been interviewed: two different individuals are counted; the weights associated with these individuals will be equal to $w_{1} / r_{k(J)}$ and $w_{2} / r_{k(J)}$, and the expression corresponding to these two pseudo-individuals in the estimator will still be equal to $y_{k}\left(w_{1}+w_{2}\right) / r_{k(J)}$.

Of course, this presumes that the information provided by the same person surveyed in two different locations/on two different days is the same, which is far from given.

However, identifying individuals can be important in order to limit nonresponse (see section 7).

### 4.2 Estimation of a Ratio

Let us now suppose that we are interested in the estimation of the average $\bar{Y}_{J}$ (see Formula (2)). $\bar{Y}_{J}$ Can be estimated by the Hajek estimator,

$$
\hat{\bar{Y}}_{J}=\frac{\hat{Y}_{J}}{\hat{N}_{J}}
$$

where $\hat{N}_{J}=\Sigma_{k \varepsilon s_{p}} \tilde{w}_{k}$.

### 4.3 Variance Calculation

The variance of the estimators presented above is calculated in the classic manner provided that the reasoning is based on services. The calculation is still complex because it is a multi-stage design with unequal probabilities. To avoid underestimating the true variance, it is essential that all services be retained in cases where several sampled services point to a single individual.

### 4.4 Comparison with Other Estimating Methods

Having introduced "weight sharing" estimators, it is appropriate to consider an alternative estimating method where we will try to estimate directly the selection probabilities of individuals in the sample. (The weight sharing estimator is not a classic Horvitz-Thompson estimator : the weights of that estimator clearly depend on the complete service sample (see formula (4)). This method can appear more natural. However, we must make two comments:

- it is not reasonably possible to obtain the selection probabilities of physical persons without relying on the services that the individual receives, based on the information provided by the latter when visiting the various centres. Based on the previous expressions, we get:

$$
\operatorname{Prob}\left(k \varepsilon s_{P}\right)=\operatorname{Prob}\left(\bigcup_{i \in \Pi(J) ; K(i)=k} i\right)
$$

The Poincare formula enables us to express this probability from single, double, triple, etc probabilities of inclusion of services. Except for the single inclusion probabilities, these are complex probabilities derived as they are from selections of unequal and without replacement probabilities. We cannot therefore hope to obtain a calculable expression for $\operatorname{Prob}\left(k \varepsilon s_{p}\right.$ ). In contrast, the weight sharing method is very simple to apply:

- in a more structured manner, a problem comes from the fact that the selection probabilities of unsampled services are not known in advance because of the multi-stage sample. At the earlier stages, the selection probabilities depend on the previous selection. In our case, we do not know the use of the centres that are not surveyed. To obtain the selection probability of an individual, we must know the inclusion probabilities of all services that the individual receives. On the other hand, one of the strengths of the weight sharing method is that the weights of units obtained indirectly (in this case individuals) can only depend on the weights of units sampled directly (services). Lavallee (1995) points out this advantage of the method.


## 5. ESTIMATION DIFFICULTIES AND PRACTICAL SOLUTIONS IN THE CASE OF A CONSTANT VARIABLE

In the formulae that we have presented, knowing the links between individuals and the services universe is critical. However, these quantities are not known for several reasons:

- a theoretical reason: because the data collection is spread over time, and an individual interviewed at the start of the period cannot anticipate the services that he will use after the interview date (Note that data collection must necessarily be spread over time to ensure good coverage of the target population; synchronous collection, even if technically possible, would not capture the whole target population but only the persons using the services on that date);
- practical reasons: because the memory of the person interviewed becomes questionable after a few days, and because detection by the interviewer or the designer of the survey of the services provided in centres not belonging to the survey frame is very difficult.

In practice, it is therefore impossible to estimate without bias a total of interest over the period of the survey (one month) without making assumptions at the outset (see Section 5.3).

## 5.1 "Average Day" and "Average Week" estimations

This forces us to look at quantities that bring into play links over a short period, for example, a day or week. The population of persons who use the services in the survey field on a given day $j$ is $P_{J}=U_{c, r} P_{c, j, r}$. Let us now introduce the following quantities that relate to day $j$ :

$$
\begin{aligned}
& \Theta_{j}=\sum_{k \in P_{j}} y_{k} \\
& N_{j}=\sum_{k \in P_{j}} 1=\operatorname{card}\left(P_{j}\right) .
\end{aligned}
$$

If $\tau=\operatorname{card}(J)$ is the number of days in the survey reference period, we define the following parameters of interest:

- the total of $y$ in the population of persons who use the services in the survey field on an "average" day, as follows:

$$
\begin{equation*}
\Theta=\frac{1}{\tau} \sum_{j=1}^{\tau} \Theta_{j} \tag{5}
\end{equation*}
$$

A specific case is the number of persons who use the services in the survey field on an "average" day, $\bar{N}=1 / \tau \sum_{j=1}^{\tau} N_{j}$.

In the same way, the average of $y$ in the population of persons who use the services in the survey field on an "average" day is defined as:

$$
\begin{equation*}
\psi=\frac{\Theta}{\bar{N}}=\frac{\sum_{j=1}^{\tau} \Theta_{j}}{\sum_{j=1}^{\tau} N_{j}} \tag{6}
\end{equation*}
$$

Defining totals or averages for a given week or an "average week" follows the same principle.

We can estimate these parameters by simply adapting the formulae in the previous section, noting that the $\mathrm{r}_{k}(J)$ must be replaced by the number of services in the survey field that the person sampled received on the day (or week) of the survey.

Note that $s_{j}$ is the sample of persons interviewed on day $j, r_{k}(j)$ the number of services in the universe received by individual $k$ on day $j$ only, and $s_{k}(j)$ the services sampled on day $j$ that link to individual $k$.

$$
\begin{aligned}
& \Theta_{j} \text { will be estimated by } \hat{\Theta}_{j}=\sum_{k \varepsilon s_{j}} y_{k} \tilde{w}_{k} \\
& \qquad \text { where } \tilde{w}_{k}=\frac{1}{r_{k}(j)} \sum_{i \in s_{k}(j)} w_{i} .
\end{aligned}
$$

Here, the weights of the individuals depend on the day $j$. (But not the weights of the services, $w_{i}$, which are set one time for all (if there are no nonresponses, this would be the inverse of the selection probabilities of services)). The following analogy is useful to convince oneself of the difference between $\Theta$ et $Y_{J}$ : Consider a service window where everyone who comes must fill out a file. $Y_{J}$ corresponds to an approach where the person fills out a file the first time that he arrives at the window and does not fill one out on subsequent visits; the "average day" case corresponds to an approach where everyone who arrives at the window fills out a file, regardless of whether he has come to the window on some other day or not. At the end of a week, for example, the analysis of the characteristics of the persons who filled out the files will be very different in the two cases: in the second case, persons who come to the window often will be over-represented compared to the first case. It is possible to formalize this approach. We refer interested readers to Ardilly and Le Blanc (1999).

### 5.2 Practical Estimation of the Links with the Survey Frame

Even if we restrict ourselves to estimating "average week" and "average day" quantities, it is not generally possible to determine the links with the survey frame on a given day (much less a given week or over the whole of the survey period).

### 5.2.1 "Average Day" Estimation

To share the weights, we must estimate the links relating to the survey day; the situation that presents the most problems is that of persons interviewed at noon in a centre that provides meals; we do not know which centres (meals and/or accommodation) these persons will use that same evening. One option not retained by the INSEE survey designers is to include in the questionnaire questions of the type "Where will you eat (or sleep) this evening?". The answers can be used to determine the links. Of course the issue is whether the answers to these questions reflect the true links and whether the nonresponse rate for the question would be too high. From a more statistical standpoint, (hypothesizing that there is a certain regularity of behaviour) we could use information relating to the same time interval on the day before the survey. The corresponding links are undoubtedly reasonable approximations of the actual links. The practical problem relates to the possible difference in use of the centres depending on the day of week: for example, some centres are not open on weekends and others are open only on specific days.

### 5.2.2 "Average Week" Estimation

To share the weights, we retain all the links relating to the week. Clearly, the first option described in 5.2.1 cannot be used. For a given week estimations, we can use, as an approximation of the services used on day $j$ following the interview date, the services used by the individual on day
$(j-7)$. This is consistent if we assume that there is a certain pattern to the services used depending on the day of the week. This approach would mean that the calendar week would be replaced in estimators by a sliding week, that is, the last seven days beginning on the date of the interview. This is the option that was used for the survey, the questionnaire having been designed to collect the links over the 7 days preceding the interview.

### 5.3 Estimation Over the Whole of the Survey Period

It may seem that estimating totals and averages for the population $P(J)$ is one of the survey's objectives. This estimation calls on the links between individuals sampled and the services in the survey field during the whole of the data collection period, which are not known. This means that we have to model the evolution of the links beyond a week or, what amounts to the same thing, model the use behaviour of the individuals in the centres.

The solution is not simple. For example, the hypothesis that comes to mind is

$$
\begin{equation*}
\forall k, r_{k}(J)=A \cdot r_{k}(S) \tag{7}
\end{equation*}
$$

where $A$ is the number of weeks of the survey and $r_{k}(S)$ is the number of links for individual $k$ with the services of the survey field during a week $S$, leads to estimators for the whole of the period that are identical to the estimators for an average week. In effect, an "average week" estimator weights individual $k$ by

$$
\sum_{i E s_{k}(J)} \frac{w_{i}}{A \cdot r_{k}\left(S_{i}\right)}
$$

where $S_{i}$ is the week during which he received service $i$ and $s_{k}(J)$ is the sampling of services that link to individual $k$, whereas a theoretical "whole period" estimator weights the individual $k$ by

$$
\sum_{i \in s_{k}(J)} \frac{w_{i}}{r_{k}(J)}
$$

Equation (7) is therefore an adequate condition of equality of these estimators. This condition is satisfied in particular when for any $j$ and any $k$

$$
\begin{equation*}
r_{k}(J)=\operatorname{card}(J) \cdot r_{k}(j) \tag{8}
\end{equation*}
$$

that is, when the number of daily links does not depend on $j$.

This hypothesis is definitely too strong. To expand on this point, we will have to use the data provided by the survey itself on the behaviour of the individuals with respect to use of the centres.

The most sought after figure of the survey - in the French context - is undoubtedly an estimate of the size of the "homeless" population, that is, an estimation of the size of $P(J)$. In addition to the issues regarding counting the links that have already been discussed extensively, this estimation runs up against several inadequacies in the survey frame as well as the indirect nature of the sampling.

- The risk of overlooking certain structures when identifying the centres is significant. Even with an exhaustive inventory, the gap between when the inventory is established and the survey itself takes place makes it likely that new unidentified centres will appear in the survey frame. This can introduce a bias to the extent that some individuals who might use these structures would not use any other service in the survey frame. (We might also expect those in charge of certain centres to refuse to cooperate: for the INSEE survey, there was virtually no refusal by the institutions (less than $1 \%$ refusal rate). This was due largely to consideration awareness building at the time the centres were identified and just before the survey.) Further, the lack of bias depends on a correct calculation of the links; use of centres not included in the frame should not be counted in these links.
- Individuals who use the centres only outside the "classic" hours (those in which we have the means to count the services) are outside the survey frame. (Counting them would create significant on-site implementation problems.)
- Another source of bias can come from the careful counting of the total number of services provided in the centres during the survey, these numbers being used to calculate the probability of a service being sampled. For budget reasons, one person only counted the services and did the sampling, a situation that could create problems of rigour in the sampling if there is confusion in the field.
- In terms of the concepts, the only remaining problem was that the survey had to take place over a month and that the target population may have changed during that period.
The estimation of the size of the population is therefore particularly fragile. For this reason, we can expect any errors to be larger for the totals than for the averages.


## 6. ESTIMATION IN THE CASE OF VARIABLES OF INTEREST THAT ARE NOT CONSTANT OVER THE SURVEY PERIOD

Some of the survey's variables of interest depend on the observation date and therefore are not constant over the survey period. This can be the case with answers to questions dealing with the day before the interview, for example "How many meals did you have yesterday?", "How many times did you sleep in the street last week?", etc. The questions on links also fall into this category. It is therefore important to determine the extent to which we can adapt the earlier formalism to estimations involving this type of variable. In other words, where $y$ is such a variable of interest.

If we go back to expression (3), it is easy to see that the constancy of $y_{k}$ during the survey period is the condition
that makes it possible to factor $y_{k}$ and to reveal the links $r_{k}(J)$. From this we can deduce that the above type of calculation is always valid for estimations covering shorter periods than the period for which the $\mathrm{y}_{k}$ are constant.

This means that for variables that are constant for a day, we can appropriately use the "average day" estimators. For variables that are constant over the week, we can use the "average day" or "average week" estimators.

## 7. ADJUSTMENT FOR TOTAL NONRESPONSE

To describe the operation fully, we still need to explain how to move from a set of inclusion probabilities (and thus initial weights of services included in the sample) to a set of weights on respondent services. Some people will agree to the interview, others will not. We will refer to services in the first case as respondent services and those in the second case as nonrespondent services. The usual adjustment methods for total nonresponse can be applied. We suggest a nonresponse adjustment by homogeneous subgroup (for a description of the method, see for example Hambaz and Legendre 1999).

In reality, the main problem relates to the fact that there is no survey frame of individuals and thus no advance information on nonrespondents. In a world that is likely very heterogeneous, this is a considerable handicap. We therefore have to model the service response behaviour. We know from the test surveys of the INED (Institut National des Etudes Démographiques) that nonresponse varies widely depending on the type of centre (Firdion and Marpsat 1997). Other variables in the survey frame can be used to build homogeneous groups (day of the week, period of the day, groups of population centres, ...).

A reweighting of the respondent services produces weights for the respondent services of the type
$w_{i}=1 / \delta_{i} \pi_{i}$, where
$\pi_{i}$ is the probability of inclusion of service $i$ in the sample
$\delta_{i}$ is the probability estimated after the fact that service $i$ will result in a response.
This provides us with a set of weights for the respondent services.

In fact, some of the nonresponses come from the fact that the same individual is sampled several times: obviously, an individual who is sampled twice might respond the first time but not the second. (The frequency of occurrence of this event was not known at the time of writing this paper.) The second selection therefore produces a "false nonresponse". If this is not detected, the total nonresponse adjustment procedure leads to an incorrect reweighting, when the true value can be obtained from a questionnaire that has already been completed. To avoid this problem, the interviewer tries to find out the reason for the refusal and must check off a specific box when the individual states that he has already been interviewed. In this situation, the
interviewer collects some information, including the first name and the date of birth, that can be used to link this questionnaire to the questionnaire that has already been completed. (The ideal situation would be to have an identifier for the respondents. This approach was not used because of confidentiality requirements and consideration of the reaction of the persons interviewed to such a measure.) However, in the field, it can be difficult to obtain a reason for refusal. Even if a reason is given, problems can occur. (It is hard to verify that a person who states that he or she has already been interviewed has in fact been interviewed. Even if the person is showing goodwill, he may have been interviewed a few days earlier for a completely different survey than the INSEE survey.)

## 8. CONCLUSION

In this article, we show how the weight sharing method can be used to weight the survey conducted by the INSEE in order to better understand homeless persons. The method has many advantages. It makes it possible to work on a file of individuals, that is, on the natural statistical units used in the definition of the parameters of interest. Simple to apply, it also makes it easy to move from one reference period to another ("average day", "average week" estimation). Operations following to the survey, such as the nonresponse adjustment and the calculation of variance can be carried out in a traditional framework because they are done on sampled units (services), for which the selection probabilities are known, and not on individuals, for which the selection probabilities are not known. We show that a crucial quality criterion of such a survey is reliable data collection on use of services by the persons interviewed. Without these data, it is not possible to weight the survey. The weight sharing method appears to be a good compromise for a survey in which the purpose is not simply to count a population but to better understand it through the use of a questionnaire. Other alternative methodologies could be used for a survey aimed simply at determining the size of the homeless population. The first such methodology uses capture- recapture techniques to determine the size of animal populations (see for example, Pollock, Turner and Brown 1994). These techniques cannot be easily applied to a population that is often suspicious of any attempt to identify it, which they perceive negatively. Another technique is that of "snowball" sampling, which involves finding individuals of interest through the intermediary of individuals already sampled (Franck and Snijders 1994). It relies on a system of mutual knowledge of persons, who are probably illusive in the community. These methods always run up against the issue of the identifying individuals. In our case, the only places where it is possible to find the persons we are seeking are the centres: it is essential that we work through the centres.

## ACKNOWLEDGEMENTS

The authors thank the journal's Editor and two anonymous referees whose comments helped improve both the content and layout of the article. Any errors that remain are entirely our responsibility.

## APPENDIX 1: THE WEIGHT SHARING METHOD APPLIED TO THE PROBLEM

This appendix briefly presents the principle of the weight sharing method. For a more complete discussion, the reader may consult Lavallée (1995) or Deville (1999) whose notations we have used.

1. We have a population $U$ of $n$ units, and a population $V$ of $m$ units. The units of $U$ are services in the survey field. The units of $V$ are persons who used at least one service during the survey period (otherwise expressed in the present case as $V=P(J)$ with the previous notations).
$\therefore 2$. It is assumed that there are links between the units of the two populations. These links can be written in the form of a matrix

$$
\begin{array}{r}
\left(r_{i k}\right) 1 \leq i \leq n, \\
1 \leq k \leq m
\end{array}
$$

where $r_{i k}=1$ if unit $k$ of $V$ is linked to unit $i$ of $U, r_{i k}=0$ otherwise. In this case, the links connect the services to the persons who used these services: $r_{i k}=1$ if person $k$ used service $i$ of $U, r_{i k}=0$ otherwise.
3. All units of $U$ have at least one link to a unit of $V$. Clearly, that is achieved here by definition of population $V$. Further, in this case, each unit of population $U$ points to one and only one unit of $V$.

In general, we are interested in the total of a variable of interest $y$ in $V$,

$$
Y=\sum_{k \in V} y_{k}
$$

If, for example, we use $y \equiv 1$, the total of interest is the number of persons who used a service in the survey field during the month of the survey.

We can write

$$
r_{k}=\sum_{i \in U} r_{i k}
$$

The identity $Y=\Sigma_{i \in U} \Sigma_{k \varepsilon V}\left(r_{i k} / r_{k}\right) \quad y_{k}$ makes it possible to define for any $i \varepsilon U$ the variable $z_{i}=\Sigma_{k \in V}\left(r_{i k} / r_{k}\right) y_{k}$ which gives:

$$
Z=\sum_{i \in U} z_{i}=\sum_{k \in V} y_{k}=Y .
$$

Let us now assume that we have a sample $s_{U}$ from the population $U$, which is associated with a set of weights $\left(w_{i}\right)_{\text {ies } j}$. This sample implicitly defines a sample in $V, s_{V}$, specifically

$$
s_{V}=\left\{k \varepsilon V ; \exists i \varepsilon s_{U}, r_{i k}=1\right\}
$$

We assume that we collected the $r_{i k}$ for all $k \varepsilon s_{v}$, that is, that all links between individuals and the universe $U$ are known (this point is fundamental).

The total $Z=Y$ is estimated by $\hat{Z}=\sum_{s_{u}} w_{i} z_{i}$.
And consequently, if the weights are unbiased (that is, set so that $\hat{Z}$ is without bias), $\hat{Y}$ estimates $Y$ without bias..

We can rewrite $\hat{Z}=\Sigma_{s v} w_{i} \Sigma_{k \varepsilon v} r_{i k} y_{k} / r_{k}=\hat{Y}$.
The second equation impacts only $s_{V}$ by definition and therefore $\hat{Y}=\Sigma_{s_{v}} y_{k}\left(\Sigma_{s_{U}} w_{i} r_{i k} / r_{k}\right)=\Sigma_{s_{v}} y_{k} \tilde{w}_{k}$, where we have written for all $k \varepsilon s_{V}$ :

$$
\begin{equation*}
\tilde{w}_{k}=\frac{1}{r_{k}} \sum_{s_{U}} w_{i} r_{i k} . \tag{9}
\end{equation*}
$$

We can work directly on the individuals sampled. In our case, $r_{k}$ is the number of links, that is, the number of services used by the person interviewed during the survey reference period. It is the quantity that is written $r_{k}(J)$ in the previous sections, the dependence on $J$ being intended to remind that links affecting the weight can vary by the type of estimator ("average day", "average week") considered. This number is derived from the use data collected in the survey.

## APPENDIX 2: SUMMARY TABLE OF EXPRESSIONS

$J \quad$ All days in the survey reference period
$\tau \quad=\operatorname{card}(J)$, number of days in the reference period
$P(J)$ population of interest, all persons who used at least one service in the survey field during the reference period
$N_{J} \quad=\operatorname{card}(P(J))$, size of the population of interest
$C$ all centres in the population centre, denoted by index $c$
$\Pi_{c, j, t}$ all services provided in centre $c$ on day $j$ during time interval $t$, denoted by index $i$
$\Pi_{j, t} \quad$ all services provided in the population centre on day $j$ during time interval $t$
$P_{c, j, t}$ all persons who visit centre $c$ on day $j$ during time interval $t$, denoted by index $k$
$P_{j, t}$ all persons who visit one of the centres in the population centre on day $j$ during time interval $t$
$P_{j} \quad$ all persons who use services in the survey field on day $j$
$y$ variable of interest
$Y_{J} \quad$ total of variable $y$ in the reference population
$\bar{Y}_{J} \quad$ average of $y$ in the reference population
$\Pi(J)$ all services provided during the reference period in all centres in the survey frame
$r_{k}(J)$ number of services provided to individual $k$ during period $J$ in all centres in the survey field, or "number of links"
$s_{\Pi} \quad$ sample of services
$w_{i} \quad$ weight associated with the services sample
$s_{P} \quad$ sample of individuals, all individuals who received sampled services
$\tilde{w}_{k} \quad$ weight associated with the sample of individuals
$\Theta_{j} \quad$ total of $y$ in $P_{j}$
$N_{j} \quad=\operatorname{card}\left(P_{j}\right)$
$\Theta$ total of $y$ "an average day"
$\bar{N} \quad$ number of persons on "an average day"
$\psi \quad=\frac{\Theta}{\bar{N}}$, average of $y$ "on an average day"
$r_{k}(j)$ number of services received by the individual $k$ on day $j$ only
$s_{j} \quad$ sample of persons interviewed on day $j$
$s_{k}(j)$ all services sampled on day $j$ that point to individual $k$
$s_{k}(J)$ all services sampled during period $J$ that point to individual $k$

## REFERENCES

ARDILLY, P., and LE BLANC, D. (1999). Enquête auprès des personnes sans-domicile : éléments techniques sur l'échantillonnage et le calcul de pondérations individuelles, une application de la méthode du partage des poids. Working Paper, INSEE , F9903.
CHAMBAZ, C., and LEGENDRE, N. (1999). Calcul des pondérations dans le panel européen de ménages. Actes des journées de méthodologie statistique, INSEE Méthodes, 84-86.
DEVILLE, J. C. (1999). Les enquêtes par panel : en quoi different-elles des autres enquêtes? suivi de : comment attraper une population en se servant d'une autre. Actes des journées de méthodologie statistique, INSEE Méthodes, 84-86.
FIRDION, J. M., and MARPSAT, M. (1997). Comptes rendus du groupe *pondérations» de l'enquête auprès des personnes sans-domicile, mimeo.
FRANCK, O., and SNIJDERS, T. (1994). Estimating the size of hidden populations using snowball sampling. Journal of Official Statistics, 10, 53-67.
LAVALLÉE, P. (1995). Cross-sectional weighting of longitudinal surveys of individuals and households using the weight share method. Survey Methodology, 21, 25-32.
POLLOCK, K.H., TURNER, S.C. and BROWN, C.A. (1994). Use of capture-recapture techniques to estimate population size and population totals when a complete frame is unavailable. Survey Methodology, 20, 117-124.
RTI (1993). Prevalence of drug use in the Washington DC metropolitan area, homeless and transient population : 1991. Technical report, 2.

## 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 16, Number 4, 2000
Borrowing Strength When Explicit Data Pooling Is Prohibited
Jerome P. Reiter ..... 295
Attrition and Misclassification of Drop-outs in the Analysis of Unemployment Duration Johan Bring and Kenneth Carling ..... 321
Income Measurement Error in Surveys: A Review
331
331
Jeffrey C. Moore, Linda L. Stinson, and Edward J. Welniak, Jr.
Jeffrey C. Moore, Linda L. Stinson, and Edward J. Welniak, Jr.363
Lennart Nordberg
379
A Functional Form Approach to Calibration
Victor M. Estevao and Carl-Erik Särndal
401
An Assessment of the Current State of Dependent Interviewing in Household Surveys Nancy A. Mathiowetz and Katherine A. McGonagle
419
Remembering Heads and Bushels: Cognitive Processes Involved in Agricultural Establishments' Reports of Inventories Jaki Stanley McCarthy and Martin A. Safer
Accuracy of Using Pneumonia as an Underlying Cause in the Cause-of-Death Register Boo Svartbo, Linda Nilsson, Anders Eriksson, Gösta Bucht, Lars Age Johansson, and Lars Olov Bygren ..... 435
Index to Volume 16, 2000 ..... 445
Volume 17, Number 1, 2001
Can A Statistician Deliver?
Richard Platek and Carl-Erik Särndal ..... 1
Comment
21
21
Barbara A. Bailar ..... 25
Paul P. Biemer
33
33
Alain Desrosières, Jean-Claude Deville, and Olivier Sautory ..... 39
Eva Elvers and Lennart Nordberg
43
43
Ivan P. Fellegi
51
51
Robert M. Groves and Nancy A. Mathiowetz
55
55
David Holt
David Holt
63
63
Margarida Madaleno
71
71
David A. Marker and David R. Morganstein
73
73
Pilar Martín-Guzmán ..... 77
Photis Nanopoulos
87
87
Svein Nordbotten ..... 93
Erkki Pahkinen
Erkki Pahkinen
99
99
Jaki Stanley McCarthy ..... 103
Lynne Stokes ..... 107
Dennis Trewin
Dennis Trewin
113
Rejoinder
Enumeration Accuracy in a Population Census: An Evaluation Using Latent Class Analysis ..... 129Paul P. Biemer, Henry Woltman, David Raglin, and Joan Hill
Estimation of Interviewer Effects on Multivariate Binary Responses in a Community Based Survey Sujuan Gao ..... 149
Swedish Employment in the 1950s - How to Fill the Lacuna
Gudmundur Gunnarsson and Thomas Lindh ..... 163
Practicing What We Preach: The Application of Continuous Improvement in a Preclinical Statistics Department at a Pharmaceutical Company
V. Bill Pikounis, Joseph M. Antonello, Danielle Moore, Edith T. Senderak, and Keith A. Soper ..... 187

## Volume 29, No. 1, March/mars 2001, 1-172

Richard A. LOckHART
Message from the new Editor/Un message du nouveau rédacteur en chef ..... 1
Christian Genest:
Report from the former Editor/Rapport du redacteur en chef sortant ..... 3
Malay Ghosh and Yeong-Hwa Kim
The Behrensy Fisher problem revisited: a Bayes-frequentist synthesis ..... 5
Daniel R. Eno and Keying Ye
Probability matching priors for an extended statistical calibration model ..... 19
Patrick E. Brown and Piet de Jong
Nonparametric smoothing using state space techniques ..... 37
Kert Viele
Evaluating fit in functional data analysis using model embeddings ..... 51
Thomas J. DiCiccio, Michael A. Martin and Steven E. Stern
Simple and accurate one-sided inference from signed roots of likelihood ratios ..... 67
Daniel B. HaLL
On the application of extended quasi-likelihood to the clustered data case ..... 77
Paul Kabaila and John Byrne
Exact short Poisson confidence intervals ..... 99
Vincent F. Melf, Connie Page and Margarida Geraldes
An adaptive randomized design with application to estimation ..... 107
Giseon HEo, Byron Schmuland and Douglas P. Wiens
Restricted minimax robust designs for misspecified regression models ..... 117
Min Tsao and Julie Zhou
On the robustness of empirical likelihood ratio confidence intervals for location ..... 129
Marc Hallin, Amal Mellouk and Khalid Rifi
Projection de Hájek et polynômes de Bernstein ..... 141
Faouzi El Bantli and Marc Halunn
Asymptotic behaviour of M-estimators in $\mathrm{AR}(p)$ models under nonstandard conditions ..... 155
Acknowledgement of referees' services/Remerciements aux membres des jurys ..... 169
Corrigenda and addendum ..... 170
Forthcoming Papers/Articles à paraître ..... 171

## GUIDELINES FOR MANUSCRIPTS

Before having a manuscript typed for submission, please examine a recent issue of Survey Methodology (Vol. 19, No. 1 and onward) as a guide and note particularly the points below. Accepted articles must be submitted in machine-readable form, preferably in WordPerfect. Other word processors are acceptable, but these also require paper copies for formulas and figures.

## 1. Layout

1.1 Manuscripts should be typed on white bond paper of standard size ( $81 / 2 \times 11 \mathrm{inch}$ ), one side only, entirely double spaced with margins of at least $11 / 2$ inches on all sides.
1.2 The manuscripts should be divided into numbered sections with suitable verbal titles.
1.3 The name and address of each author should be given as a footnote on the first page of the manuscript.
1.4 Acknowledgements should appear at the end of the text.
1.5 Any appendix should be placed after the acknowledgements but before the list of references.

## 2. Abstract

The manuscript should begin with an abstract consisting of one paragraph followed by three to six key words. Avoid mathematical expressions in the abstract.

## 3. Style

3.1 Avoid footnotes, abbreviations, and acronyms.
3.2 Mathematical symbols will be italicized unless specified otherwise except for functional symbols such as "exp( $\cdot$ )" and " $\log (\cdot)$ ", etc.
3.3 Short formulae should be left in the text but everything in the text should fit in single spacing. Long and important equations should be separated from the text and numbered consecutively with arabic numerals on the right if they are to be referred to later.
3.4 Write fractions in the text using a solidus.
3.5 Distinguish between ambiguous characters, (e.g., w, $\omega ; \mathbf{o}, \mathrm{O}, 0 ; 1,1$ ).
3.6 Italics are used for emphasis. Indicate italics by underlining on the manuscript.

## 4. Figures and Tables

4.1 All figures and tables should be numbered consecutively with arabic numerals, with titles which are as nearly self explanatory as possible, at the bottom for figures and at the top for tables.
4.2 They should be put on separate pages with an indication of their appropriate placement in the text. (Normally they should appear near where they are first referred to).

## 5. References

5.1 References in the text should be cited with authors' names and the date of publication. If part of a reference is cited, indicate after the reference, e.g., Cochran (1977, p. 164).
5.2 The list of references at the end of the manuscript should be arranged alphabetically and for the same author chronologically. Distinguish publications of the same author in the same year by attaching $\mathrm{a}, \mathrm{b}, \mathrm{c}$ to the year of publication. Journal titles should not be abbreviated. Follow the same format used in recent issues.


[^0]:    ${ }^{1}$ Gad Nathan, Departement of Statistics, Hebrew University, 91905 Jerusalem, Israel.

[^1]:    Avinash C. Singh and Shiying Wu, Statistical Research Division, Research Triangle Institute, Research Triangle Park, N.C. 27709-2194, U.S.A.; Brian Kennedy, Statistics Canada, Ottawa, Ontario, Canada, K1A OT6.

[^2]:    Wayne A. Fuller, Iowa State University, 221 Snedecor Hall, Ames, IA, U.S.A.; J.N.K. Rao, Carleton University, Ottawa, Ontario K1S 5B6.

[^3]:    ${ }^{1}$ Philip Beil, Australian Bureau of Statistics, e-mail: philip.bell@abs.gov.au

[^4]:    1 Jack Gambino, Brian Kennedy and Mangala P. Singh, Statistics Canada, Tunney's Pasture, Ottawa, Ontario, K1A 0T6.

[^5]:    1 Jae-Kwang Kim, Westat, 1650 Research Boulevard, Rockville, Maryland, 20850, U.S.A.

[^6]:    1 Trivellore E. Raghunathan, James M. Lepkowski, John van Hoewyk and Peter Solenberger, University of Michigan, Institute for Social Research, Survey Methodology Program, P.O. Box 1248, Ann Arbor, MI 48106-1248, U.S.A.

[^7]:    I Johane Dufour, François Gagnon, Yves Morin, Martin Renaud and Carl-Erik Särndal, Statistics Canada, Tunney's Pasture, Ottawa, Ontario, K1A 0T6.

[^8]:    1 Pascal Ardilly and David Le Blanc, Institut National de la Statistique et des Etudes Economiques, 18 boulevard Adolphe Pinard, 75675, Paris, Cedex, France. E-Mail : pascal.ardilly@insee.fr, leblanc@ensae.fr .

