SURVEY METHODOLOGY

Catalogue No. 12-001-XPB

A JOURNAL PUBLISHED BY STATISTICS CANADA

DECEMBER 1998

•

VOLUME 24

•

NUMBER 2

Statistics Canada Statistique Canada

Canada

SURVEY METHODOLOGY

A JOURNAL PUBLISHED BY STATISTICS CANADA

DECEMBER 1998 • VOLUME 24 • NUMBER 2

Published by authority of the Minister
responsible for Statistics Canada

January 1999

Catalogue no. 12-001-XPB

Frequency: Semi-annual

ISSN 0714-0045

Ottawa

Statistics Canada Statistique Canada

Canada

SURVEY METHODOLOGY

A Journal Published by Statistics Canada

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

EDITORIAL POLICY

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

Submission of Manuscripts

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

Subscription Rates

The price of Survey Methodology (Catalogue no. 12-001-XPB) is $47 per year in Canada and US $47 per year Outside Canada. Subscription order should be sent to Statistics Canada, Operations and Integration Division, Circulation Management, 120 Parkdale Avenue, Ottawa, Ontario, Canada K1A 0T6 or by dialling (613) 951-7277 or 1 800 700-1033, by fax (613) 951-1584 or 1 800 889-9734 or by Internet: order@statcan.ca. A reduced price is available to members of the American Statistical Association, the International Association of Survey Statisticians, the American Association for Public Opinion Research, and the Statistical Society of Canada.

SURVEY METHODOLOGY

A Journal Published by Statistics Canada

Volume 24, Number 2, December 1998

CONTENTS

Survey Methodology, December 1998
Vol. 24, No. 2, pp. 99-100
Statistics Canada

In This Issue

This issue of *Survey Methodology* begins with a special section entitled "Longitudinal Surveys and Analysis" which contains six of the papers presented at the IASS/IAOS Satellite Meeting on Longitudinal Studies held in Jerusalem in 1997. One or two other papers from that conference, which were not ready on time for this issue, may appear in future issues of the journal. I am very grateful to Gad Nathan and Christopher Skinner who were the Coordinating Editors for this special section. Without their persistence and hard work it would not have been possible.

The first paper in the special section, by Binder, introduces the topic by reviewing the current status and challenges for longitudinal studies as compared to cross-sectional studies. The discussion is divided into four parts, reviewing in turn the special issues and challenges encountered in the design, implementation, evaluation, and analysis of longitudinal surveys.

Bassi, Torelli and Trivellato consider the problem of estimation of gross flows among labour force states when there are classification errors in the data. They first review various strategies for the collection of longitudinal labour force data, and their likely implications for classification errors. They then present a general modeling framework and a modified LISREL model for adjusting gross flows estimates to correct for classification errors. The methods are illustrated by two case studies using data from the U.S. Survey of Income and Program Participation and the French Labour Force Survey.

Clarke and Chambers consider the impact of household level non-response on estimates of labour force gross flows. They propose a class of models for nonignorable household-level nonresponse. They then use simulations to demonstrate that labour force gross flows estimates can be biased in the presence of this nonignorable household level nonresponse, and that estimates using household level nonresponse models can reduce this bias. If the household level nonresponse mechanism is correctly specified then this source of bias is removed completely; however, even incorrectly specified household nonresponse models can reduce the bias.

Salamin considers the problem of estimating a change in proportion for a small area. He shows how a general multivariate logistic regression model can be used to describe the longitudinal data obtained from a rotating panel design. He also considers how the parameters of this model may be restricted to describe various types of dependance among the repeated observations, leading to alternative model based estimates of change. The method is illustrated by estimating changes in probability of being employed for a Canton in Switzerland using data from the Swiss Labour Force Survey. Compared to simple differences of estimated proportions of employed persons, the model based estimates have smaller standard errors.

Dorfman, in his paper, attempts to treat consumer price indices from a statistical point of view. He first reviews price index theory in general, including the stochastic approach and objections to it. He then proposes a modification to the stochastic approach, based on state space modeling, which circumvents the major criticism of it. The approach is illustrated using price and quantity data for canned tuna.

In the last paper in the special section, Tambay, Şchiopu-Kratina, Mayda, Stukel and Nadon describe the treatment of nonresponse in the Canadian National Population Health Survey. Data collected at the first cycle of the survey are considered as potential predictors of nonresponse to the second cycle. A CHAID (Chi-square Automatic Interaction Detection) algorithm is used to determine weighting classes for nonresponse adjustment at the second cycle. The paper also briefly describes the sample design and other steps in the derivation of the estimation weights.

Sinclair and Gastwirth study the problem of misclassification error of labour force status in the Current Population Survey of the U.S. Bureau of the Census. To do so, they extend the method of Hui and Walter, which is appropriate for dichotomous data using reinterview data, to the trichotomous case. Unlike other methods, this method does not assume that reinterview data is error free, but rather assumes an error in both the original interview and the reinterview data. They make an empirical assessment by comparing the estimated error rates generated by their method as opposed to other existing methods such as that of Poterba and Summers, and find that the degree of underestimation of the error tends to be higher when the true unemployment rate is in fact high. Finally, rather than assuming a constant error rate throughout, they attempt an analysis assuming that the error rates are constant only within time groupings having differing levels of unemployment.

Renssen considers the problem of combining information on variables collected from two different large surveys, using auxiliary information from a smaller third survey collecting all of the variables. Using ideas from statistical matching and from calibration, he proposes methods for the production of two-way tables, for the production of microdata files, and for the estimation of correlations. For the production of two-way tables his development leads to consideration of two different sets of calibration constraints, one termed incomplete two-way stratification and the second termed synthetic two-way stratification. In a simulation study using data from a pilot study for the Dutch Household Survey on Living Conditions, the calibration based on synthetic two-way stratification is shown to be much better.

Arnab considers different strategies for sampling on two occasions. The sample at the second occasion is assumed to be a combination of a subsample of the first sample and a new, unmatched sample. Different strategies for subsampling the first sample and estimating a total at the second occasion are compared. He reviews strategies already existing in the literature, and proposes two new ones. Efficiencies of various strategies are compared analytically and empirically.

Finally, Korn and Graubard consider the problem of generating confidence intervals for proportions having a small expected number of positive counts. Noting that the Clopper-Pearson binomial intervals traditionally used in the non-survey setting are inappropriate for use with complex survey data, they propose a modification of these intervals. Via simulation, they then compare the proposed intervals to others commonly used such as: logit-transform intervals, Breeze (1990) intervals based on a Poisson approximation, and normality-based linear intervals. They also illustrate the proposed and three alternative methods with applications using data from both the National Health and Nutrition Examination Survey and the Hispanic Health and Nutrition Examination Survey.

The Editor

Survey Methodology, December 1998
Vol. 24, No. 2, pp. 101-108
Statistics Canada

Longitudinal Surveys: Why Are These Surveys Different From All Other Surveys?

DAVID A. BINDER[1]

ABSTRACT

We review the current status of various aspects of the design and analysis of studies where the same units are investigated at several points in time. These studies include longitudinal surveys, and longitudinal analyses of retrospective studies and of administrative or census data. The major focus is the special problems posed by the longitudinal nature of the study. We discuss four of the major components of longitudinal studies in general; namely, Design, Implementation, Evaluation and Analysis. Each of these components requires special considerations when planning a longitudinal study. Some issues relating to the longitudinal nature of the studies are: concepts and definitions, frames, sampling, data collection, nonresponse treatment, imputation, estimation, data validation, data analysis and dissemination. Assuming familiarity with the basic requirements for conducting a cross-sectional survey, we highlight the issues and problems that become apparent for many longitudinal studies.

KEY WORDS: Frames; Administrative data; Data collection; Nonresponse; Imputation; Estimation; Data analysis.

1. REASONS FOR LONGITUDINAL STUDIES

Each year around the world various statistical agencies conduct thousands of surveys. Usually, these surveys obtain information required for decision or policy making. These surveys are not conducted just for historical purposes, but also to have information on what measures may be taken to assist with making various policy changes. Most surveys are based on cross-sectional data, where a survey is taken of a particular population at a given point in time. Various summaries are taken about the population under consideration at the time of the survey. However, very often the interest is not so much in what actually happened when the survey was taken, but what would be the impact of making various changes. Alternatively, a planned change in policy may be forthcoming and monitoring the effect of this change is desirable. What is most important is the time element. For example, when trying to learn about certain phenomena such as health status or education attainment, one is interested in the various determinants related to these outcomes. Sometimes, the actual temporal relationship is not even clear in terms of what are the causes that precede the effects. These could be measured if, instead of taking a cross-sectional survey, surveys are conducted over time, either as a series of cross-sectional surveys or, alternatively, using the same panel of respondents from one occasion to another. This common sense notion has led to the desire to conduct more longitudinal studies. This also has the benefit that the effects of unobserved variables may be less important when the same respondents are used to compare differences over time.

One of the factors contributing to the increase in the number of longitudinal studies is that administrative data sources can now be used more effectively, thus making certain longitudinal studies feasible. Administrative data are becoming increasingly available. These data are often routinely collected for the same individuals over a period of time. Even if the data collected from the administrative sources is not ideal for the survey-taker, they may provide a good proxy for the information.

The advantage of designing a study as longitudinal is that a common methodology can be used for each of the various waves of the survey. This may lead to more valid conclusions. Often, when trying to understand various patterns of social and economic change, conducting surveys of the same respondents on a number of occasions is best. Less desirable, but possibly satisfactory, is simply to repeat the survey from one occasion to another without necessarily returning to the same respondents. This may be less costly. The main point is that to understand certain phenomena over time, collecting the information on more than one occasion is necessary.

When making decisions on the nature of a new longitudinal study, a number of cost considerations need to be accounted for. Obviously, one needs to consider the benefits against these various costs. Issues that longitudinal studies could address cover many subject-matter areas. We enumerate just a few of them. In the area of health status, one is interested in changes to health status and the determinants that lead to these changes. In other words, what are the health risks, and what, in fact, is the effect of these health risks on health status in the long term? By collecting the data from the same individuals over a period of time, one can assess these factors, not just on small scale studies typical of clinical trials, but on large-scale nationally-based population health surveys. However, the type of information that can be obtained from a nationally-

[1] David A. Binder, Business Survey Methods Division, Statistics Canada, R.H. Coats Building, Tunney's Pasture, Ottawa, Ontario, Canada, K1A 0T6.

based longitudinal survey would be very different from that which is obtainable in a clinical trial.

Another topic where there is interest in observations over time is in the area of labour and income. For example, it is not enough to have information on the net change to labour force status and labour force participation rate over time. It is also of interest to know which individuals move from say, being unemployed to working or to not being in the labour force. In recent times, employment patterns have changed. More women are working and part-time work is more common. Frequency of job changes is also changing. To understand these phenomena, longitudinal surveys can answer many important questions. The characteristics, for example, of entry level jobs taken by those who were previously unemployed may be of interest, as well as effectiveness of different job search strategies by individuals or the effectiveness of various government training schemes.

Length of spells in poverty is of increasing interest. For example, for persons with low income, how long does one remain in that situation? What are the various factors that will determine whether this is a long-term situation? How important are education and other factors with respect to poverty and the length of poverty spells?

In the field of education, an interesting aspect is the school-to-work transition at the time when people finish full-time school and decide to join the labour force. This behaviour may be measured more easily through a longitudinal study than through other types of surveys. Another education-related example is the effectiveness of various types of education such as vocational training and adult training programs.

In justice and victimization, there are many examples where observing the same individuals over time can be beneficial. Persons who have been victimized could be followed up to assess the long-term implications. As well, persons who have been involved with the judicial system may be observed over time to determine the subsequent patterns of behaviour and the determinants for these patterns.

Studies of consumer behaviour are of great interest to marketers and others. This would include purchasing patterns for consumers. Event histories for consumer purchasing would be very useful to many researchers.

Studies on the effects of government transfer payments to individuals over time can be important to policy makers. A longitudinal study can determine how long individuals may be dependent on such government payments, whether or not habits are created because of the existence of some of these payments, what are the characteristics of the individuals and what are the long-term effects of participation in various assistance programs.

On the economic side, the longitudinal characteristics of various businesses are of great interest. One can measure how efficient these businesses are, what the use of technology is in these businesses, what is the long-term effect of this use and how productivity is changing over time. Various interesting questions on business demographics could be asked; for example, what are the characteristics of businesses that result in failure, what are the economic conditions under which businesses are created. As well, mergers and amalgamations are of interest with respect to the conditions under which these occur. Through longitudinal studies these phenomena can be more easily measured.

There have been various structural changes to many businesses over the last few years and it is only through longitudinal studies that one can observe some of these structural changes at the micro level. Many measures can be estimated only when the respondents are measured on more than one occasion.

Another area of interest is in agriculture, where the nature of farming is undergoing transition. Of interest is how farms are changing, both in terms of the products that are being produced and the size of the farms. Changes in the characteristics of who is running the operation are also of interest.

As we have discussed, there are many applications and many facets to longitudinal studies. Also, there are many dimensions to their design and analysis. In the following sections we summarize these issues around Four Questions: design issues, implementation issues, evaluation issues and analysis issues. Many of these issues have been discussed in Kasprzyk, Duncan, Kalton and Singh (1989) and in Armstrong, Darcovich and Lavallée (1993). Some design issues and time series methods are reviewed in Binder and Hidiroglou (1988). We include a few more recent references.

QUESTION 1: DESIGN ISSUES

When designing a longitudinal study, advance planning is vital to the success of the study. For example, one must ensure that only relevant and accurate information is being collected from the respondents so that the potential benefit of the longitudinal survey is maximized. This implies that the longitudinal analyses to be undertaken from the survey should be planned from the outset to ensure that the relevant data are obtained. Duncan and Kalton (1987) give an excellent summary of many of the issues. Webber (1994) describes the testing strategy used in the planning of the Survey on Labour and Income Dynamics. Huggins and Fischer (1994) discuss the plans for the redesign of the Survey of Income and Program Participation based on their experiences. Longitudinal studies can be more expensive than a series of cross-sectional studies. Therefore, the benefits of collecting these data must be even greater since the costs themselves are higher. As well, ensuring that funding for a longitudinal study can be assured is important since the fruits from the longitudinal nature of the study may not be borne until at least the second or third wave of

the study. There is a difference, of course, between planning for a study to be longitudinal from the beginning as opposed to taking a series of cross-sectional data and trying to merge them into a longitudinal database. Obviously, the former is more desirable but often, because of the history of the survey-taking organization, a series of cross-sectional data already exists so that merging these would be a reasonable alternative; see Hughes and Hinkins (1995).

In general, careful attention needs to be paid to the design of the database for any longitudinal survey where the analysis includes longitudinal measures such as the study of episodes and spells. For some statistical agencies and organizations, the survey program is now in transition from cross-sectional surveys to longitudinal surveys. The change from a series of cross-sectional surveys to longitudinal surveys requires careful planning. When conducting longitudinal surveys, the databases need to be maintained and updated in ways that are very different from cross-sectional surveys. There may be many infrastructure and organizational issues within the agency that become apparent as more longitudinal surveys are being conducted, particularly with respect to the maintenance of the databases and the survey operations. The impact of such changes on the statistical organization may be substantial.

An important issue to consider when planning for a longitudinal survey is whether or not the users will also be requiring cross-sectional estimates. Is there a requirement to have information about the respondents who are in the survey over a period of time, and also being able to produce estimates for a single point in time as if it were a cross-sectional survey? If this is the case, there are major implications on the way the survey is designed and implemented; see Lavallée (1995). This concern would also be present if the variables of interest include comparing cross-sectional estimates over time, as opposed to true longitudinal measures such as studying autocorrelations for common units in a business survey.

Concepts and the definitions used in longitudinal surveys are usually obtained through consultations with the data users. Even the definition of the longitudinal unit to be observed over time may need clarification for dynamic populations. This is the case for both household surveys and for business surveys. Understanding the user requirements and discussing what can be measured over time with appropriate quality is important. During the survey planning, these requirements must be carefully weighed against what is operationally feasible in an actual survey context. Given the eventual costs of these studies, conducting thorough tests is often worthwhile, particularly on the survey questionnaires. A point that deserves more attention is the need for more standard longitudinal measures that are common across countries. This would permit governments and researchers to make better international comparisons.

Another major component for designing longitudinal studies is the creation, use and maintenance of sampling frames over time in ways that facilitate the implementation of the study. For example, an establishment panel survey may be based on a business register that can be highly dynamic with respect to births, deaths, mergers and amalgamations. It is important that the definitions of which units are to be included in these panels over time are clear under these conditions.

One reason that longitudinal surveys have become more prevalent in recent years is the fact that there are more administrative data files available now that can be used as frames for conducting the longitudinal studies. The administrative files themselves may also contain useful data information besides just being useful as frames per se. Some data manipulation of the administrative data is usually required to make these data useful for the statistical purpose of the longitudinal study, however. In general, the impact of frame changes to the study must be carefully considered at the design stages.

A common practice is to take a number of different administrative files and to match them to create a sampling frame. As well, some longitudinal studies are based solely on the information contained in various administrative files. The difficulty, of course, is that over time these administrative files will change. This may imply a change to the samples that are being taken from these files, and therefore special measures will need to be taken to keep the analyses relevant.

Often a longitudinal study is based on an existing survey or census conducted at a point in time in the past, and this then becomes the basis for the sampling frame for following up respondents over time. One disadvantage of this is that it becomes difficult to obtain cross-sectional estimates when births to the population are excluded from the frame. Record linkage techniques may be necessary for maintaining the frame and such techniques are usually error-prone.

For rare populations, it is often advantageous to use not just a single frame but to use multiple frame methods. This ensures that there is adequate representation from the populations of interest that might be underrepresented in a single frame, but this may also require the use of record linkage and complex weighting techniques.

An important design issue is the method of sampling from the frame once it has been established. In Kalton and Citro (1993), a number of different types of longitudinal surveys were enumerated. These were repeated surveys, that is, a series of cross-sectional surveys; panel surveys, where certain respondents are selected and followed up over time; repeated panel surveys, where new panel surveys are selected at different points in time; rotating panel surveys, where on each occasion a panel is dropped from the study and a new panel is added; overlapping surveys, where there are common respondents from one occasion to the other, but not necessarily through a fixed panel sample design; split panel surveys that can be a combination of panel surveys and repeated or rotating panel surveys. The

sample design must ensure that there is a sufficient sample from the population of interest as well from any of the control groups. Administrative data have proven to be very useful when designing a sample for many of these surveys as they often provide a suitable frame.

As a referee pointed out, a key issue at the design stage is the strategy for dealing with sample loss though attrition, due to nonresponse, leaving the target population, *etc.* Possibilities include topping up the sample in subsequent waves, but such a strategy can distort the representativity of the cohort. Another strategy would be to start with a larger sample and not replace lost units; see, for example Singh, Petroni and Allen (1994).

When deciding on a particular sample design, consideration must be given to the related weighting and estimation issues. As well, the periodicity or frequency of the survey must be established. Obviously, when the variables of interest change more rapidly, having the survey conducted more frequently would be more desirable. On the one hand, more frequent surveys lead to increased cost and respondent burden; on the other hand, less frequent surveys can lead to larger recall biases. These cost-quality tradeoffs are usually difficult to quantify.

Very often, if both cross-sectional and longitudinal estimates are required, ensuring that there will be valid cross-sectional estimates may be necessary to select supplementary samples. This is because there may be members of the population in the cross-sectional estimates who were not in the sampling frame on previous waves and, therefore, would not be represented in the sample. Czajka (1994) studies this for the case of estimating income.

Designing some evaluation samples is also worthwhile at the planning stage. There are a number of sources of bias in longitudinal surveys. Some of these biases can occur simply because the same respondent has been surveyed on a number of occasions. Therefore, consideration should be given to adding additional samples for evaluation purposes only, in order to be able to measure some of these impacts. These samples would include individuals in the target population that were not in the longitudinal survey. They are most useful for evaluating cross-sectional measures.

QUESTION 2: IMPLEMENTATION ISSUES

The second main issue we discuss is related to the implementation of a longitudinal study. First, one has various choices of modes of data collection. Recently, computer-assisted interviewing has gained popularity. With computer-assisted interviewing, more choices of survey instruments are available. For example, using dependent interviewing where the respondent or the interviewer has access to the responses from previous occasions is easier. This may increase or decrease certain biases. Hill (1994) asseses this in the context of Survey of Income and Program Participation.

Of course, since we are going back to the same respondents on a number of occasions, the question of response burden is even more crucial than in a single cross-sectional survey. We do not want to overload the respondent since this could result in higher refusal rates at later waves of the survey. Michaud, Dolson, Adams and Renaud (1995) suggest respondent burden can be reduced by making more use of administrative data. Reducing attrition due to nonresponse is an important goal in longitudinal surveys and consideration may be given to the use of monetary or other incentives to help keep the integrity of the sample over time; see Lengacher, Sullivan, Couper and Groves (1995). Another means of reducing attrition is to collect information to aid in the tracing efforts and to keep in contact with the respondents over time; McGuigan, Ellickson, Hays and Bell (1995) studies alternatives of tracing, reweighting and sample selection modelling, to cope with attrition problems.

In some longitudinal surveys, some data are collected retrospectively; that is, questions are asked which refer to previous points in time as well as the current point in time. This could lead to what is known as seam effects. As a result, the observed changes over the reference periods may depend on which periods contain data obtained retrospectively.

Administrative records may be useful to enrich the database so that not all data need to be collected directly from the respondent; see Michaud *et al.* (1995). Of course, this could depend on the quality of the administrative data, its availability, and what the interplay is between the information from the administrative records and the survey variables; see Stearns, Kovar, Hayes and Koch (1996) for an example that studies this relationship. When dealing with administrative data or merged sample files, there may be data gaps in these various files and how to handle these data gaps becomes an issue.

In general, changes to the frame structure can result in difficulties when performing the longitudinal analyses. Some key characteristics of the respondents could also be changing over time. For example, in a business register, if the industrial classification information changes because of the fact that businesses change the nature of the products that they are producing over time, being able to keep track of this changing classification on the database to ensure that the longitudinal analyses are as useful as possible is important. This can also complicate the analysis.

Many issues arise when the database is obtained by combining the samples from a series of individual surveys. Integrating this information may present a challenge because different surveys may have used different methodologies. This could result in some inconsistencies in the quality of the information from one database to another.

Important issues for many longitudinal surveys are those related to record linkage. Record linkage is used in many processing steps. In some cases, the longitudinal studies may be based solely on these linked files. Record linkage

is common for creating and maintaining the survey frames, including linking administrative files over time, linking administrative files and survey frames and linking separate survey frames. For example, for surveys of establishments, we may wish to create longitudinal composite records for the establishments that are based on several independent repeated surveys, since many of the establishments are surveyed on each occasion. Record linkage is often used to find which units correspond to the same establishments. Record linkage is also used to identify births to a frame. Of course, the errors due to the record linkage can be important in the analysis; see Scheuren and Winkler (1993).

In some cases, in fact, no real respondents are being followed over time. Instead record linkage is used to create artificial populations through statistical matching. These populations are then analysed as if they were real.

Another implementation issue is that of handling nonresponse. It is known that nonresponse to longitudinal surveys does not occur completely at random. There tends to be differential nonresponse among different subpopulations. Therefore, special attention needs to be placed on how the imputations or reweighting will be performed; see, for example, Tambay, Şchiopu-Kratina, Mayda, Stukel and Nadon (1998). When using administrative data as the basis for the longitudinal study, there may be missing administrative data and special procedures will be necessary to handle this situation.

For missing data, there are generally two methods of treatment: imputation and reweighting. Reweighting is common for situations where there is wave nonresponse. Imputation is more frequently used when there is partial nonresponse within a given wave of the survey. There can be advantages to longitudinal imputation as opposed to cross-sectional. For longitudinal imputation, the longitudinal information from the same individual on the database is used as the basis for doing the imputation, as opposed to using other individuals at the same point in time. For attrition and wave nonresponse, one may wish to model the attrition rates and use these models to compensate for the nonresponse through weight adjustments. A variety of weight adjustments were researched for the Survey of Income and Program Participation and the results were presented in Rizzo, Kalton and Brick (1994), Folsom and Witt (1994), and An, Breidt and Fuller (1994). Singh, Wu and Boyer (1995) study this problem for the difficult case of estimating gross flows.

There are many complexities that may be introduced into the derivation of the weights. There are various approaches and techniques available to calculate both cross-sectional weights and longitudinal weights. Cross-sectional weights are used for measures of the population at a single point in time, whereas the longitudinal weights are necessary when data from individuals over more than one occasion are included. The analyst may wish to have person-level weights that are different from the household-level weights; Kalton and Brick (1995). For example, for some variables such as household income, using household-level weights would be preferable to the individual person-level weights. Weighting becomes more complex with the use of multiple frames. Effective use of administrative data may imply even more complexities in the weighting scheme itself; see, for example Stearns *et al.* (1996).

There are many causes for the samples to become unrepresentative. For example, lack of representativity could be due to problems of coverage due to immigration into the population. Some undercoverage may be due to attrition. Some overcoverage could be due to including some non-sampled co-habitants of a household, thus implying that those individuals could be included in the sample by living with an originally sampled person; see Lavallée (1995) and Kalton and Brick (1995). Other types of systemic overcoverage are also possible. Ensuring that no biases are introduced requires special weighting treatments. For longitudinal surveys in particular, this may become quite complex. Administrative data can be used both to assess whether or not the sample is representative and to provide information for making the appropriate adjustments.

Since much of the estimation for longitudinal study will be associated with measuring change as opposed to measuring the phenomena at a single point in time, there will be questions about how to develop the variances for these estimates of measures of change. Some new procedures may need to be developed for this situation. In general, variance estimates can become quite complicated when the statistics are complex functions of the longitudinal observations. For example, income class boundaries may change over time and studying the transitions of individuals from one class to another is of interest.

Another complexity of estimation may be the desire to include information from ongoing cross-sectional surveys to produce new integrated measures, using all the information that is available from the various available sources.

QUESTION 3: EVALUATION ISSUES

The third set of issues we discuss is related to the evaluation of the information and methods. Even though the evaluations may be conducted separately from the implementation, the results of such evaluations should impact on the survey itself, either by altering the estimation methods or by changing the way the survey is designed and implemented in future waves.

There are many sources of biases that could be studied. Biases may be due to dependent interviewing by giving the respondent and the interviewer information that could refer to a previous occasion of the survey. Seam effects can arise from retrospective studies; see, for example Murray, Michaud, Egan and Lemaître (1991). Other sources of bias could occur when the nonresponse is informative; that is,

when the nonresponse propensity is related to the variable of interest. An example would be when household level nonresponse is correlated with gross flows within the household, where gross flows are the changes in the individual's classification; see Clarke and Chambers (1989). Other biases could be due to measurement or classification errors; see, for example, Bassi, Torelli and Trivellato (1998). Conditioning bias could arise from the fact that since we have been asking the respondents about information, such as labour dynamics, they may have become more sensitized to some of these issues so that their behaviour could change because of the fact that they are included in the survey.

The effect of response errors and interviewer errors on the analysis should be evaluated. Different individual interviewer methods may lead to different error rates. The stability or instability of the turnover of interviewing staff could affect some analyses. Questions such as whether or not the information was collected by proxy can also be relevant.

Other evaluations could be performed to measure the effect of attrition and to evaluate various imputation methodologies and other nonresponse handling strategies; see Tin (1996) for an evaluation of attrition using econometric methods. Schejbal and Lavrakas (1995) study the effect of panel attrition in a dual-frame local telephone survey. Corder, Manton and Woodbury (1994) study ways to improve coverage and reduce attrition in the context of the National Long Term Care Survey. Panel attrition could be the result of non-traceable or refusal cases, the impact of which can be quite different from cross-sectional surveys, and these differences should be studied. Allen and Petroni (1994) discuss the problem of adjusting for movers.

There is a need to develop quality studies that take into account the special features of longitudinal surveys. Many quality control studies are available in the conduct of longitudinal surveys besides the usual ones for cross-sectional surveys, since the repeated nature of the study can lead to a more efficient identification of error-prone cases. Since for longitudinal studies, the stability of the data over time is an issue, methodological changes in the study could have an impact on the longitudinal measures that are of interest and these should be evaluated. Administrative data can provide useful evaluations since some of the data can help validate some of the results.

QUESTION 4: ANALYSIS ISSUES

Analysis concerns are the last set of issues we discuss. It is the potential analysis of the longitudinal study that is its most important facet. The causes or determinants of various outcomes are of major interest to the data users. However, the modelling of these causes can be complex, particularly if the survey itself is of a complex nature. Many of these issues are discussed in Singh and Whitridge (1990) and in Hidiroglou and Michaud (1998).

Examples of the kinds of analyses that are common would be measures of gross flows or other measures of gross change. Gross flows refers to the change of an individual from one category to another. In other words, it is the flow from category A to category B between two points in time, as opposed to net flow that is the change in the margins over time. There are difficult questions about the impact of measurement error on the measurement of gross flows. If fairly large measurement errors are present on each occasion, there will be a significant impact on the bias of the estimates of the gross flows, even if the net flows themselves are not as adversely affected. Sometimes, sample rotation will aggravate this problem, since accounting for sample rotation properly when measuring gross flows can be problematic. Special treatment is needed for those panels that are entering the sample on a given occasion and for those panels that have left the sample on the previous occasion to get good estimates of these flows. The changes to the population when gross flows are being measured need to be sorted out from the gross flows themselves. In other words, the change from one occasion to another is a combination of the changes in size of the population and the individual changes within the population. The situation can become even more complex when the gross flows are themselves analysed with respect to other information such as income dynamics.

As a referee pointed out, an important issue is the need for educating users on how longitudinal data can be analysed effectively. The recent increase in the number of longitudinal surveys raises many opportunities for new types of analysis, but many analysts who have been studying only cross-sectional surveys may not be aware of the most appropriate techniques.

For the many surveys that use frames based on administrative data, accounting for the frame changes in the analysis may be necessary, since inclusion on the frame can be subject to changes in administrative procedures, as well as changing conditions for the individuals. For example a file of unemployment insurance beneficiaries would be subject to changing eligibility criteria, as well as changing personal situations.

The measurement of change can often be decomposed into various components. For example, the movement of units in the sample from one domain to another can be sorted out from the changes of the data for units within the same domain. Holt and Skinner (1989) contains an interesting discussion on various components of change.

For more complex analyses, such as modelling of time series, most classical time series models do not account for the fact that the information is derived from a sample survey. Therefore, the sampling errors resulting from the sample survey are not properly taken into account in the time series modelling.

In the analysis, some measures may depend on other cross-sectional surveys. For example, it may be another cross-sectional survey that determines the income class

boundaries to be used in the analysis of the longitudinal survey. This may add to the complexity of the analysis since the boundaries can change over time.

Whether and how to use the sampling weights have created difficulties for many analysts, since many of the classical models for analysis of data over time do not use the sampling weights. Procedures need to be developed that incorporate the survey weights in the analysis properly. For large-scale surveys, using the weights is often preferable as this provides some protection against model misspecification.

Errors resulting from the processing, such as the record linkage operation, may need to be incorporated in the analysis or at least some studies need to be taken to understand the impact of these kinds of errors; see, for example Dorinski and Huang (1994).

Often administrative data are used as part of the analysis since these data may be available more readily than collected information. However, since there may be conceptual or other difficulties with the administrative data, special analytical methods may need to be developed to use the administrative data effectively.

Finally, we mention the difficulties associated with the data dissemination. Longitudinal summary measures need to be developed for many phenomena. Often these are not suitable for the usual tabular displays that are commonly used in cross-sectional studies. Many analyses require access to the microdata. This could create problems with respect to protecting the confidentiality of the respondents. The usual measures that one takes when releasing microdata files on cross-sectional surveys may not be sufficient when releasing surveys which are longitudinal in nature, because the databases are so much richer so that the risk of being able to identify an individual on such databases becomes much greater. Protecting the respondents' confidentiality is of paramount importance, so a conservative approach that may not fulfill all the users' requirements may be necessary.

SUMMARY

We have briefly discussed many of the questions and issues that are now being investigated by researchers concerned with the design and analysis of longitudinal studies. Based on our discussion, we see that many questions need to be further investigated. As we gain more experience with longitudinal surveys, many of these issues will be better understood and many new issues will arise. The opportunities for important research and investigation are numerous.

ACKNOWLEDGEMENTS

The author is grateful for many useful suggestions from the referees and the Associate Editor.

REFERENCES

ALLEN, T.M., and PETRONI, R.J. (1994). Mover nonresponse adjustment research for the Survey of Income and Program Participation. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 662-667.

AN, A.B., BREIDT, F.J., and FULLER, W.A. (1994). Regression weighting methods for SIPP data. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 434-439.

ARMSTRONG, J., DARCOVICH, N., and LAVALLÉE, P. (Eds.) (1993). *Proceedings: Symposium 92, Design and Analysis of Longitudinal Surveys*, Statistics Canada.

BASSI, F., TORELLI, N., and TRIVELLATO, U. (1998). Data and modeling strategies in estimating labour force gross flows affected by classification errors. *Survey Methodology*, 24, 109-122.

BINDER, D.A., and HIDIROGLOU, M.A. (1988). Sampling in time. *Handbook of Statistics Volume 6: Sampling* (Eds. P.R. Krishnaiah and C.R. Rao). North Holland: Amsterdam, 187-211.

CLARKE, P.S., and CHAMBERS, R.L. (1998). Estimating labour force gross flows from subject to household-level nonignorable nonresponse. *Survey Methodology*, 24, 123-129.

CORDER, L.S., MANTON, K.G., and WOODBURY, M. (1994). Improving coverage, response rates, and nonresponse follow-up via a longitudinal list sample design: The National Long-Term Care Surveys. *Proceedings of the 1994 Annual Research Conference*, U.S. Bureau of the Census, 63-84.

CZAJKA, J.L. (1994). Income stratification in panel surveys: Issues in design and estimation. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 791-796.

DORINSKI, S.M., and HUANG, H. (1994). Use of administrative data in SIPP longitudinal estimation. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 656-661.

DUNCAN, G.J., and KALTON, G. (1987). Issues of design and analysis of surveys across time. *International Statistical Review*, 55, 97-117.

FOLSOM, R.E., and WITT, M.B. (1994). Testing a new attrition nonresponse adjustment method for SIPP. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 428-433.

HIDIROGLOU, M.A., and MICHAUD, S. (Eds.) (1998). *Proceedings: Symposium 98, Longitudinal Analysis for Complex Surveys*, Statistics Canada. To appear.

HILL, D.H. (1994). The relative empirical validity of dependent and independent data collection in a panel survey. *Journal of Official Statistics*, 10, 359-380.

HOLT, D., and SKINNER, C.J. (1989). Components of change in repeated surveys. *International Statistical Review*, 57, 1-18.

HUGGINS, V.J., and FISCHER, D.P. (1994). The redesign of the survey of income and program participation. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 668-673.

HUGHES, S., and HINKINS, S. (1995). Creation of panel data from cross-sectional surveys. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 408-413.

KALTON, G., and BRICK, J.M. (1995). Weighting schemes for household panel surveys. *Survey Methodology*, 21, 33-44.

KALTON, G., and CITRO, C. (1993). Panel surveys: Adding the fourth dimension. *Survey Methodology*, 19, 205-215.

KASPRZYK, D., DUNCAN, G., KALTON, G., and SINGH, M.P. (Eds.) (1989). *Panel Surveys*. Wiley: New York.

LAVALLÉE, P. (1995). Cross-sectional weighting of longitudinal surveys of individuals and households using the weight share method. *Survey Methodology*, 21, 25-32.

LENGACHER, J.E., SULLIVAN, C.M., COUPER, M.P., and GROVES, R.M. (1995). Once reluctant, always a reluctant? Effects of differential incentives on later survey participation in a longitudinal study. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 1029-1034.

McGUIGAN, K.A., ELLICKSON, P.L., HAYS, R.D., and BELL, R.M. (1995). Tracking, weighting, and sample selection modeling to correct for attrition. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 402-407.

MICHAUD, S., DOLSON, D., ADAMS, D., and RENAUD, M. (1995). Combining administrative and survey data to reduce respondent burden in longitudinal surveys. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 11-20.

MURRAY, T.S., MICHAUD, S., EGAN, M., and LEMAÎTRE, G. (1991). Invisible seams? The experience with the Canadian Labour Market Activity Survey. *Proceedings of the 1991 Annual Research Conference*, U.S. Bureau of the Census, 715-730.

RIZZO, L., KALTON, G., and BRICK, M. (1994). Adjusting for panel nonresponse in the Survey of Income and Program Participation. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 422-427.

SCHEJBAL, J.A., and LAVRAKAS, P.J. (1995). Panel attrition in a dual-frame local area telephone survey. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 1035-1039.

SCHEUREN, F., and WINKLER, W.E. (1993). Regression analysis of data files that are computer matched. *Survey Methodology*, 19, 39-58.

SINGH, A.C., and WHITRIDGE, P. (Eds.) (1990). *Analysis of data in time. Proceedings: Symposium 89, Analyses of Data in Time*, Statistics Canada.

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, American Statistical Association*, 396-401.

SINGH, R.P., PETRONI, R.J., and ALLEN, T.M. (1994). Over-sampling in panel surveys. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 674-679.

STEARNS, S.C., KOVAR, M.G., HAYES, K., and KOCH, G.G. (1996). Estimates of national hospital use from administrative data and personal interviews. *Journal of Official Statistics*, 12, 47-61.

TAMBAY, J.L., ŞCHIOPU-KRATINA, I., MAYDA, J., STUKEL, D., and NADON, S. (1998). Treatment of nonresponse in cycle two of the National Population Health Survey. *Survey Methodology*, 24, 147-156.

TIN, J. (1996). Program participation and attrition: The empirical evidence. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 669-674.

WEBBER, M. (1994). The survey of labour and income dynamics: lessons learned in testing. *Proceedings of the 1994 Annual Research Conference*, U.S. Bureau of the Census, 85-99.

Survey Methodology, December 1998
Vol. 24, No. 2, pp. 109-122
Statistics Canada

Data and Modelling Strategies in Estimating Labour Force Gross Flows Affected by Classification Errors

FRANCESCA BASSI, NICOLA TORELLI and UGO TRIVELLATO[1]

ABSTRACT

Gross flows among labour force states are of great importance in understanding labour market dynamics. Observed flows are typically subject to classification errors, which may induce serious bias. In this paper, some of the most common strategies, used to collect longitudinal information about labour force condition are reviewed, jointly with the modelling approaches developed to correct gross flows, when affected by classification errors. A general framework for estimating gross flows is outlined. Examples are given of different model specifications, applied to data collected with different strategies. Specifically, two cases are considered, *i.e.*, gross flows from (i) the U.S. Survey of Income and Program Participation and (ii) the French Labour Force Survey, a yearly survey collecting retrospective monthly information.

KEY WORDS: Correlated classification errors; Latent class models; Longitudinal data; Recall errors; Seam effect.

1. INTRODUCTION

Gross flows among labour force states, are a powerful tool to analyse labour market dynamics. Gross flows regard changes at individual level, and therefore their estimation rests on the availability of longitudinal data.

The effects of erroneous classification of units with respect to their position in the labour market, can cause spurious transitions. Even if one might assume that these errors cancel out when estimating net flows, they cannot be ignored when estimating gross flows.

Various strategies can be adopted, in order to correct gross flows for classification errors. Basically, they depend on:

(a) assumptions about the classification error mechanism, following from
  (a1) the survey design (panel surveys – possibly with a rotating scheme, retrospective surveys, some mixture of retrospective and panel surveys, *etc.*), and/or;
  (a2) the content and structure of the questionnaire (availa-bility of one or more indicators of the variable of interest, format of the questions – episode based or event based, *etc.*);
(b) assumptions about the generating process of the transitions among labour force states.

In this paper, some of the most common strategies used to collect longitudinal information about labour force condition are reviewed, jointly with modelling approaches developed to correct gross flows when affected by classification errors. It is shown that most of the usual specifications proposed in the literature, can be seen as special cases of a general formulation, which allows to elucidate advantages and disadvantages of each specification, and makes it possible to consider a common estimation strategy.

The focus of the paper is on sound applications of this general modelling approach, for estimating gross flows from survey data collected with different strategies. Two cases are considered: (i) the U.S. Survey of Income and Program Participation and (ii) the French Labour Force Survey, a yearly rotating panel survey with retrospective monthly information.

The organization of the paper is as follows. Section 2 briefly discusses various strategies for collecting longitudinal data on labour force participation, and their likely implications for classification errors, as they emerge from the survey methodology literature. In section 3, a fairly general approach for modelling gross flows affected by classification errors, *i.e.*, for jointly estimating true gross flows and conditional response probabilities, is outlined. Examples are also given on how some well known models for correcting observed gross flows, can be specified as special cases of this approach (section 3.1). Attention is then devoted to a convenient framework for formulating the above models, provided by latent class models and, more specifically, by the so-called "modified LISREL model" proposed by Hagenaars (1990), a general tool to describe causal relationships among observed and unobserved categorical variables (section 3.2).

The final, and main part of the paper (section 4), is devoted to a detailed presentation of the two case-studies. The modelling approach is common: *a priori* information on the measurement characteristics of the survey (and possibly on the true process), is combined with specification searches, in order to obtain parsimonious and (hopefully) sensible models. As already noted, the two case-studies are reasonably different, chiefly in terms of the design of the surveys: this diversification turns out to be useful for illustrating different model specifications, and various strategies for reaching/testing the final formulation.

[1] Francesca Bassi, Nicola Torelli and Ugo Trivellato, Dipartimento di Scienze Statistiche, Via San Francesco, 33, 35121, Padova, Italy.

From the two case-studies, the following overall evidence can be drawn:

(a) the modified LISREL model has proved to be a set-up, flexible enough for modelling the error mechanism in longitudinal data collected with different survey designs, as well as the generating process of true labour force transitions;

(b) specifically, in the measurement part of the model, we were able to incorporate the pattern and the effects of correlated classification errors, which are particularly important in surveys with retrospective features;

(c) observed transitions are corrected towards the direction expected, on the basis of theoretical and empirical evidence on measurement errors effects, (not mechanically towards mobility, as strategies based on the assumption of independent classification errors do).

2. THE ROLE OF DATA COLLECTION STRATEGIES

Information for labour gross flows estimation comes from longitudinal data, *i.e.*, observations on the same units pertaining to different time points. Recently, there have been increasing efforts in collecting longitudinal data. This is true also for surveys, whose main goal is to measure the labour force condition of individuals in a given population. On the other side, this focus on collecting, and using longitudinal data, raised new questions about the origin and pattern of measurement (≡ classification) errors, as well as their possible effects on estimates of the quantities of interest. General references about sources of classification errors for longitudinal data, collected by surveys across time, are Duncan and Kalton (1987) and Kalton and Citro (1993). In this section, some main implications of classification errors on modelling strategies, to correct gross flows are briefly discussed.

A typical argument about the effect of measurement error in estimating gross flows, is that it leads to overestimation of changes. This is true when one assumes that measurement errors are not correlated over time. This assumption is not realistic in many cases (see Skinner and Torelli 1993; Singh and Rao 1995; van de Pol and Langeheine 1997), and should be reconsidered taking carefully into account, the data collection strategy actually adopted. Broadly speaking, if longitudinal data are (at least partly) collected by retrospective interrogation, one can argue that memory inaccuracy leads to correlated errors.

Specific assumptions about classification errors can be successfully introduced in appropriate statistical models, only if additional information is available in the form of plausible *a priori* knowledge about the error generating mechanism and/or supplementary data about the labour force state.

Modelling strategies to correct gross flows for classification errors, should then take into account the measurement process actually used, in the sense that the amount of classification errors and the direction of possible bias, are related to the strategy adopted to collect longitudinal data.

As it is well known, longitudinal data can be obtained by different survey strategies. It is convenient to distinguish at least between (i) panel surveys and (ii) retrospective surveys. In addition, the availability of multiple indicators deserves specific attention.

Panel surveys are the most natural ways of collecting longitudinal information. Among these, rotating panel surveys play a prominent role. In fact, this is the scheme adopted in most national Labour Force Surveys (LFSs), whose primary goal is estimation of labour force stocks. For LFSs with a rotating sampling design, longitudinal information on the (usually short) sequence of states, can be easily obtained by matching data on individuals participating in two or more successive surveys. In LFSs, the reference period, concepts and definitions for classifying people, are typically consistent with the International Labour Office (ILO) recommendations (Hussmanns, Mehran and Verma 1990): this makes measures of labour force conditions reasonably accurate and comparable over space and time. Data on labour force participation are collected also through general purpose household surveys. In this case, attention to labour force condition is less prominent than in the preceding type of surveys, and reference periods, concepts and definitions, might be less consistent with ILO recommendations.

Alternatively, longitudinal information can be collected by retrospective surveys. Cross-sectional surveys can include retrospective questions, to get information on the sequence of labour force states experienced by sampled individuals. In this case, the interrogation strategy is crucial to reduce errors due to memory (recall errors, telescoping, *etc.*). Procedures to improve accurate reporting in retrospective surveys, rely upon contributions from cognitive psychology and survey methodology (for a review, see O'Muircheartaigh 1996). Besides, evidence on the amount and the direction of bias due to memory inaccuracy, is found in many empirical studies. It is worth adding, that in retrospective surveys, factors related to length of recall period, salience of events considered, and/or difficulty in retrieving data on past events, usually lead to a simplified format of questions, not consistent with ILO conventions on labour force condition.

Interesting opportunities for estimating gross labour flows in the presence of classification errors, come from the widespread practise of using a mixture of the panel and the retrospective strategies. Panel surveys use retrospective questions, at least on a limited number of topics, to cover the period between two successive waves (this is the case of the Survey of Income and Program Participation, as will be seen in section 4.2). The main characteristics of the measurement process when such a mixed strategy is used,

have to be carefully considered, as they might have a considerable impact in formulating reasonable models for classification errors. More specific traits of the measurement process emerge also from consideration of the peculiarities of the survey design.

From a different perspective, an important opportunity for modelling classification errors is given by the availability of multiple measurements of labour force state, *i.e.*, data on the labour market condition of an individual at a given time, provided by two or more different sources. This information is of great importance in general, and particularly when fairly complicated patterns of correlated classification errors are to be considered. Multiple indicators on labour force state can be collected (i) in the same interview or (ii) in different interviews (*e.g.*, in different waves of a panel survey).

The first case is not very common, but sometimes questions regarding labour force condition are asked in different contexts, and in different ways. For instance, first, a self-classification of the individual with respect to labour force condition is asked; then, in a different section of the questionnaire, a sequence of questions are put forward that allow to classify the respondent according to standard labour force definitions. (For a different example, see the case of the Survey of Income and Program Participation in section 4.2.)

The second case covers several situations. At least two of them are worth considering:

(a) data from reinterview studies, often collected specifically to get information on classification errors probabilities (in such a case, the common practice is to assimilate reinterview data to validation data: for classical procedures to correct gross flows based on reinterview data, see Abowd and Zellner 1985, Poterba and Summers 1986, and Chua and Fuller 1987);

(b) data collected retrospectively in panel surveys, but referring to a time point already covered by the preceding interview, or collected in a supplementary survey carried out occasionally and covering the reference period(s) of the current panel survey. It is obvious that, in this case different measures of the same variable(s) of interest can be polluted by classification errors with largely different characteristics.

Many of the points raised here will be clarified in the case-studies presented in section 4, where the joint presence of panel and retrospective information and of multiple indicators of the same latent variable is exploited in order to get parsimonious models.

3. ESTIMATING GROSS FLOWS AFFECTED BY CLASSIFICATION ERRORS

3.1 A General Framework

Specification of statistical models to adjust labour force gross flows for classification errors, should allow one to take into account, the nature of available data (as reviewed in the previous section), and substantial assumptions on the generating process of (i) transitions among labour force states (*e.g.*, Markov chain structures) and (ii) measurement errors (*e.g.*, uncorrelated *vs.* correlated measurement errors).

In the simplest case, we consider panel data, where at each time period $t = 1, ..., T$, a discrete variable Y_t is observed for a generic unit, in a random sample of size n. In our case-studies, the units will be individuals, and the time periods, months or quarters. Y_t takes one among r possible distinct values or states. Y_t is an imperfect measure of y_t, which denotes the true state of a generic unit at time t. In general, it is not necessary to assume, that y_t varies over the same set of states $1, 2, ..., r$, but for simplicity, and without loss of generality, we will consider here the same set of states as for Y_t.

Strategies for estimating gross flows, rely upon an appropriate specification of the joint probability of the true and the observed process $P(Y_1, ..., Y_T, y_1, ..., y_T)$. Statistical analysis is then based on marginalization with respect to unobserved quantities:

$$P(Y_1, ..., Y_T) = \sum_{y_1=1}^{r} \cdots \sum_{y_T=1}^{r} P(Y_1, ..., Y_T, y_1, \cdots y_T). \quad (3.1)$$

Models are based on parsimonious specifications of the joint probability function $P(Y_1, ..., Y_T, y_1, \cdots y_T)$. Essentially this can be obtained by decomposing it into a product of conditional probabilities, following from an appropriate set of assumptions about the dependence structure among the components $Y_1, ..., Y_T, y_1, ..., y_T$.

For our purposes, a convenient starting point for model specification, comes from assumptions (i) about the structure of the generating process of the true transitions among labour force states and (ii) about the measurement process (exploiting, for instance, substantial knowledge or empirical evidence from the data collection strategy adopted).

In a model aimed at distinguishing between true and observed turnover in the labour market, a typical example that exploits this idea, is provided by Latent Class Markov (LCM) models (van de Pol and Langeheine 1990). For a generic unit, the following probabilities are specified:

$$q_t^{l_t j_t} = P(Y_t = l_t \mid y_t = j_t) \qquad t = 1, ..., T \quad (3.2)$$

$$\pi_t^{j_t j_{t-1}} = P(y_t = j_t \mid y_{t-1} = j_{t-1}) \qquad t = 2, \ldots, T \tag{3.3}$$

$$\pi_1^{j_1} = P(y_1 = j_1) \tag{3.4}$$

Conditional probabilities (3.2) represent the relationship between true and observed states, *i.e.*, the probability of reporting at time t, state l_t, while the true state is j_t. Clearly, this specification implies the local independence assumption, *i.e.*, $Y_1, \ldots, Y_T$ are independent, given $y_1, \ldots, y_T$. Conditional probabilities (3.3) describe the dynamics in the labour market, *i.e.*, the probability that a transition from j_{t-1} to j_t occurs, when moving from time $t - 1$ to t: according to (3.3), the true transition process evolves following a first order Markov chain. Finally, probabilities (3.4) describe the initial condition for the Markov process.

The marginal probability for the observed sequence (3.1) is then given by:

$$P(Y_1 = l_1, \ldots Y_T = l_T) = \sum_{j_1=1}^{r} \ldots \sum_{j_r=1}^{r} \pi_1^{j_1} \prod_{t=2}^{T} q_t^{l_t j_t} \pi_t^{j_t j_{t-1}} \tag{3.5}$$

For four measurement points, model (3.5) is equivalently represented by the path diagram in Figure 1, where arrows indicate direct effects between variables.

$$\begin{array}{ccccccc} y_1 & \rightarrow & y_2 & \rightarrow & y_3 & \rightarrow & y_4 \\ \downarrow & & \downarrow & & \downarrow & & \downarrow \\ Y_1 & & Y_2 & & Y_3 & & Y_4 \end{array}$$

Figure 1. Path Diagram of a LCM Model for Four Measurement Points

It is worth observing, that the assumption of local independence is equivalent to the Independent Classification Errors (ICE) assumption. As noted in the previous section, the ICE assumption has been severely criticised, and seems definitely unreasonable when longitudinal data are collected by retrospective questions.

As another example, for $T = 2$, classical strategies to correct gross flows based on reinterview studies, can be represented within the framework outlined above. In this case, additional information is used, in the sense that the q_t parameters are exogenously estimated from the reinterview study, and are plugged in (3.5) in order to obtain directly $P(y_1, y_2)$.

The same framework can be used, to encompass more general assumptions on both the latent and measurement processes, up to include serially correlated classification errors. As an interesting case, we consider the model by Pfeffermann, Skinner and Humphreys (1998). Ignoring here initial conditions, they reformulate conditional response probabilities as follows:

$$q_t^{l_t j_t} = P(Y_t = l_t \mid y_t = j_t, Y_{t-1} = l_{t-1}) \qquad t = 2, \ldots, T, \tag{3.6}$$

thus overcoming the ICE assumption.

A similar formulation, aimed at introducing, at least partially, dependence between the observed state at time t and the sequence of true states at times t and $t - 1$, has been suggested by van de Pol and Langeheine (1992), who extend the model to allow also for a second order Markov chain, for the true transition process.

The modelling strategy for estimating true flows can be further extended in various directions, namely:

(a) It is straightforward to extend the model, to exploit the availability of multiple indicators of the same unobserved true state. This implies that response probabilities, as those in (3.2), are defined for one or more additional observed variables, treated as imperfect measures of the same latent state y_t. As an example, a LCM model for two indicators per latent variable, and four points in time, is represented in Figure 2. In this model, each couple of indicators referring to a given point in time, is assumed to be independent, conditionally on the corresponding latent variable, in the sense that the correlation between them, is completely explained by their relation with y_t.

(b) Observed heterogeneity at the individual level, in the transition and/or the measurement processes, can be introduced by conditioning on a set of covariates X_t. An example is given in Pfeffermann *et al.* (1998). They use covariate information at the unit level and model their impact on labour market condition by multinomial logit.

(c) Unobserved heterogeneity can also be considered, which leads to mixed latent class models (van de Pol and Langeheine 1990). A simple case is the movers/stayers model, where a different behaviour, at the latent level, is assumed for groups of units, while the group membership of the units cannot be directly observed.

$$\begin{array}{ccccccc} W_1 & & W_2 & & W_3 & & W_4 \\ \uparrow & & \uparrow & & \uparrow & & \uparrow \\ y_1 & \rightarrow & y_2 & \rightarrow & y_3 & \rightarrow & y_4 \\ \downarrow & & \downarrow & & \downarrow & & \downarrow \\ Y_1 & & Y_2 & & Y_3 & & Y_4 \end{array}$$

Figure 2. Path Diagram of a LCM Model for Four Measurement Points and Two Indicators for Each Latent Variable

3.2 Latent Class and Related Models as a Tool for Estimating Gross Flows With Measurement Errors

A special case of the general model formulation outlined in the above section, are latent class models, where the true state in the labour market plays the role of the latent variable, and the observed state acts as its indicator. Some of the specifications outlined in the previous section, include dependence among classification errors. A general and convenient approach for handling it, which includes standard latent class models with correlated classification

errors, is the so called modified LISREL model proposed by Hagenaars (1990).

The modified LISREL approach consists of an extension of Goodman's (1973) path analysis, which is a tool to describe causal relationships among observed categorical variables, through a system of logit equations. Basically, the extension incorporates latent variables. Thus, a modified LISREL model combines a measurement sub-model, which specifies the dependence of the indicators on latent variables, and a structural sub-model, which specifies ordered relations among latent and possible external variables. As the name itself suggests, it can also be viewed as the analogue for discrete variables, of the well known LISREL model for continuous variables (Joreskög and Sörbom 1988).

Modified LISREL models, allow to introduce serially correlated classification errors, by inserting direct effects between the indicators (Hagenaars 1988). The presence of direct effects implies, that the association among observed variables, is not completely explained by the effects of the latent variables on their indicators, but that there exists a source of additional association among the indicators, over and above the part that is explained by their relation with the latent variables.

Once a reasonable model has been specified, identification should be ascertained. The model involves many unobservables, and identification of all parameters is not automatically assured.

Reasonable opportunities to achieve identification, rest on two strategies, possibly used in combination: (i) imposition of plausible equality restrictions among the set of parameters and (ii) availability of multiple indicators of the unobserved true state. The latent class Markov model represented in Figure 1, for example, is not identified without extra restrictions on its parameters. If the latent chain is assumed to be time homogeneous, or response probabilities are restricted to be equal across time, the model can be shown to be identified (Lazarsfeld and Henry 1968). Availability of multiple indicators for the unobserved true state, can also help identification of complex measurement models. Identification criteria for some very special specifications, have been proven (for example, the model in Figure 2 can be shown to be identified), but no general rules have been provided yet to ascertain global identification. It is advisable to check at least local identification, *i.e.*, identifiability of the unknown parameters in a neighbourhood of the maximum likelihood solution. Goodman (1974) stated that a sufficient condition for local identifiability of a latent class model, is that the Information matrix be full of rank. Goodman's condition may be computationally difficult to check. Moreover, with some data sets, it may happen that the Information matrix is not of full rank, simply because some estimates are very close to the boundaries of the parameter space. An alternative, empirical way to check identifiability, is to estimate the model using different sets of starting values. If different sets of starting values result in the same value for the log-likelihood function but in different parameter estimates, then the model is not identifiable.

As for estimation, modified LISREL models may be treated as directed loglinear models with latent variables (Hagenaars 1997). A directed loglinear model results in a sequence of parsimonious multinomial logit models, possibly with latent variables, which are estimated stepwise. At each step, one dependent variable is considered, and a multinomial logit model is estimated on a contingency table, which has been collapsed over the variables, that do not directly influence the dependent variable in the causal order. Estimates obtained at each step are, at the end, combined in order to obtain estimated parameters for the full model. Directed loglinear modelling yields exactly the same parameter estimates, standard errors and test statistics as the Goodman standard procedure, but using simpler marginal tables. If the causal model contains one or more latent variables, an appropriate estimation technique must be used, *e.g.*, an implementation of the EM algorithm (Meng and Rubin 1993).

The empirical validity of the complete causal model may be tested, comparing the estimated expected frequencies with the observed ones in the complete table, by means of the likelihood ratio L^2 and the Pearson X^2 statistics. However, the structure of the observed data on labour market transitions, is such that many cells show very low observed frequencies. For this reason, the usual X^2 and L^2 criteria must be used only as a general indication of fit, since their asymptotic χ^2 distribution is no longer guaranteed, due to the sparse and unbalanced pattern of the contingency table.

Various strategies can be adopted to extend and improve model evaluation, and three of them are worth mentioning in this context:

(i) A restricted model nested within a larger one, can be tested with the conditional test, *i.e.*, considering the difference in the L^2 values of the two models, which is asymptotically distributed as χ^2 under weaker conditions (Goodman 1981, and Haberman 1978).

(ii) In general, using multiple criteria can be a sensible strategy. Indices based on the information criterion, such as AIC or BIC, can be useful to compare alternative non-nested models. Another advantage of AIC and BIC is that, in the selection procedure, they weight the goodness of fit of a model against its parsimony, considering the model degrees of freedom and the sample size. (AIC = L^2 - 2 x degrees of freedom. BIC = L^2 -ln(N+1) x degrees of freedom.) The model that is preferred, in this context, is the one with the lowest value of AIC or BIC.

(iii) Monte Carlo resampling techniques can be implemented to simulate the asymptotic distribution of X^2 and L^2 (Langeheine, Pannekoek and van de Pol 1995).

4. TWO CASE-STUDIES

4.1 The General Set Up

In this section we present two applications of the modified LISREL approach to correct observed gross flows in the labour market. Data come from surveys with partly different designs:

(1) the U.S. Survey of Income and Program Participation (SIPP), a multi panel household survey, which collects retrospective information on the between waves working history;
(2) the French Labour Force Survey (FLFS), a yearly retrospective survey, with one month overlapping reference periods.

For each case-study, a model is specified on the basis of *a priori* information on both the true transition process and the error generating mechanism. *A priori* information is crucial for model specification, in order to obtain parsimonious and plausible models.

All the models are written in the form of a modified LISREL model, and estimated by the EM algorithm. Actually, we used the *l*EM program (Vermunt 1993) and checked all the models for local maxima.

The two final models turn out to be rather complex, since they incorporate correlation among classification errors, and specific assumptions on respondent's behaviour. This fact, together with the sparse and unbalanced pattern of the observed contingency table, typical of labour force transitions, demands for goodness of fit evaluation criteria, other than L^2 and X^2. In the first case-study, alternative models have been judged by means of the BIC index, and on the basis on substantive knowledge on the labour market in the U.S.. In the second case, alternative models have been compared by means of the conditional test.

In the following sections, models are presented in a logical and verbal form, while the mathematical formulation for the final model is given in the relevant Appendix.

4.2 The SIPP Data

SIPP is a multi panel household survey conducted by the U.S. Bureau of the Census, in order to collect information on topics such as employment, income, participation in social programs, *etc*. The reference population is the U.S. noninstitutionalized individuals over 14.

The survey started in 1984, and is a continuing one: as a general pattern, each year a new sample of households, called "panel", has been selected for the survey and followed for two and half years (for a detailed description of SIPP, see U.S. Department of Commerce 1991, and Citro and Kalton (1993)).

Each panel is randomly divided into four "rotation groups" and interviewed at 4-months intervals for eight times. For practical reasons, each rotation group is interviewed in each of four consecutive months, and retrospective questions collect information with reference to the 4-months period elapsing between subsequent interviews. Each set of interviews with the full sample is termed a "wave".

We will refer to the 1986 panel, which started in February 1986 and ended in August 1988. We will consider the intermediate period from January 1986 to January 1987, over which we have information from all four rotation groups. Figure 3 represents the survey design with regard to our sample.

Information on labour force participation, is collected mainly in the "Labour Force and Recipiency" section of the questionnaire (for an additional piece of information, collected in another section of the questionnaire, see below), where each respondent is asked to report on a weekly basis his/her labour market history in the preceding four months (18 weeks), by going through a series of filtered questions. The respondent is first asked whether he/she had a job or a business, at any point in time during the reference period. If the respondent gives a negative answer, he/she is asked whether he/she spent any time looking for work, or was in layoff, and, if so, in exactly which weeks. On the other hand, if the answer to the

Interview Month	Rot. Group	Wave	Reference months										
February	2	1	Oct	Nov	Dec	Jan							
March	3	1		Nov	Dec	Jan	Feb						
April	4	1			Dec	Jan	Feb	Mar					
May	1	1				Jan	Feb	Mar	Apr				
June	2	2					Feb	Mar	Apr	May			
July	3	2						Mar	Apr	May	Jun		
August	4	2							Apr	May	Jun	Jul	
September	1	2								May	Jun	Jul	Aug

Figure 3. Rotation Plan for the 1986 SIPP Panel (First 2 Waves)

starting question is positive (*i.e.*, he/she worked some time), and the respondent declared a job or a business with continuity during the reference period, he/she will move to the following section of the questionnaire. The respondent not declaring a stable situation in the labour market, is asked a long series of questions in order to establish the labour force state occupied, in each single week of the reference period.

The weekly based information is usually recorded, to obtain a monthly classification based on the usual three categories: Employed (E), Unemployed (U) and Not in the labour force (N). For individuals covering different positions during one month, the monthly labour force state is the one identified by the "modal" category with regard to the weeks of that month (Martini 1989).

Observed gross flows between two generic calendar months are then obtained as follows:

(a) For individuals belonging to three rotation groups, on the basis of retrospective data collected in the same interview. These observed flows will be called "within wave" (WW) transitions.
(b) For individuals in the fourth rotation group, by combining information collected in two different interviews, four months apart. These observed flows are termed "between waves" (BW) transitions.

When estimating monthly changes, a peculiar problem with SIPP data, is the so called "seam effect" (Young 1989): more changes are observed when data for two adjacent months are collected in two different waves – the transition covers the seam of the waves – than when they come from the same interview. The seam effect is pervasive in the survey: evidence of it for several variables of interest, is reported in Martini (1988), Marquis and Moore (1989), Kalton and Miller (1991).

Table 1 illustrates this phenomenon for our 1986 SIPP panel sample. Row 4→1 contains average BW transition rates; rows 1→2, 2→3 and 3→4 contain average WW transition rates, pertaining to the position of the two relevant reference months in each wave (for example, row 1→2 contains transition rates between the first two reference months in each wave). From Table 1, there is clear evidence that observed WW transitions describe a more stable labour market than BW ones. Moreover, WW stability increases, moving backwards in the wave (from 3→4 to 1→2).

One reasonable explanation for the seam effect, and for the systematic pattern of observed transitions throughout a wave, is the different role of measurement errors, for data obtained under the BW and WW strategies respectively. Specifically, it is likely that classification errors have a different degree of correlation for WW and BW observed flows: the higher stability documented by WW transitions may be induced by highly correlated classification errors. Indeed, if errors were uncorrelated, specifically for WW transitions, no evidence of seam effect would be expected.

A variety of plausible causes of correlated errors, is suggested by the cognitive psychology and the survey methodology literature on memory effect and recall errors (see, Bernard, Killworth, Kronenfeld and Sailer 1984, and O'Muircheartaigh 1996), among which a "conditioning" effect: respondents tend to give the same answer going backwards within the wave, and in extreme cases, they mechanically repeat the same answer for all four months.

Table 1
Observed Monthly Transition Rates (×100) for the 1986 SIPP Panel, January 1986 to January 1987

	Type	EE	EU	EN	UE	UU	UN	NE	NU	NN
1→2	WW	98.27	1.04	0.69	15.46	79.63	4.91	1.15	1.42	97.43
2→3	WW	97.91	1.13	0.96	17.34	75.96	6.70	1.38	1.71	96.91
3→4	WW	97.85	1.20	0.95	19.23	73.25	7.52	1.28	1.69	97.03
4→1	BW	94.03	2.10	3.87	26.81	42.20	30.99	5.65	3.77	90.58

Abundant empirical literature shows, that this sort of conditioning effect is the main source of classification errors in SIPP data. Other potential sources of error, typical of panel surveys, do not affect SIPP data dramatically. Administrative record check studies find little, if any, evidence of time-in-sample effect (Chakrabartry and Williams 1989; McCormick, Butler and Singh 1992). As a general consideration, we may say that in SIPP data, the seam effect dominates over other sources of error, that potentially bias gross flows estimates.

Summing up, a model-based approach to obtain unbiased gross flows from SIPP data, is justified by two arguments:

(a) the patent presence of correlated classification errors;
(b) *a priori* information on the data generating mechanism, drawn from two sources:
  (b1) specific evidences emerging from SIPP observed gross flows, such as the seam effect, and the increase in stability going backwards within the wave, just documented;
  (b2) general hints provided by the social survey literature on respondent behaviour.

In order to correct SIPP observed labour force gross flows from classification errors, a model has been built, based on the following assumptions/information:

(a) the true transition process follows a first order Markov chain;
(b) WW data transitions are affected by correlated classification errors, according to a pattern that will be specified in the sequel;
(c) for BW, the standard ICE assumption holds;
(d) rotation groups are equivalent samples also for modelling purposes, *i.e.*, respondents behave in the same way in all four rotation groups;
(e) SIPP data provide two indications on the monthly labour force state of each individual: the detailed information collected in the "Labour Force and Recipiency" section

of the questionnaire, just presented, and the additional information collected in the "Earnings and Employment" section, where the respondent is asked if he/she did/did not have a job in the reference period, on a weekly basis.

$$\begin{array}{ccccccc} W_1 & \leftarrow & W_2 & \leftarrow & W_3 & \leftarrow & W_4 \\ \uparrow & & \uparrow & & \uparrow & & \uparrow \\ y_1 & \rightarrow & y_2 & \rightarrow & y_3 & \rightarrow & y_4 \\ \downarrow & & \downarrow & & \downarrow & & \downarrow \\ Y_1 & \leftarrow & Y_2 & \leftarrow & Y_3 & \leftarrow & Y_4 \end{array}$$

Figure 4. Path Diagram of a Modified Lisrel Model for Four Measurement Points and Two Indicators for Each Latent Variable (for the Meaning of Symbols, See Main Text)

Figure 4 contains the path diagram of a simplified version of the model (*i.e.*, a version that does not aim at representing in detail, the pattern of correlated classification errors, nor at taking into account the fact that we are dealing with four rotation groups) for four points in time, *i.e.*, for four consecutive calendar months. Here y_t ($t = 1, 2, 3, 4$) represents latent variables; Y_t and W_t represent indicators; arrows indicate direct effects between pairs of variables. Indicator Y_t refers to the reported labour force state, described by the usual three categories (E, U and N), while W_t refers to the binary variable Job/No Job. Since information is collected in two different sections of the questionnaire, and with different interviewing procedures, Y_t and W_t can be assumed to be independent given y_t. On the other hand, direct effects between the indicators, account for correlated classification errors over time: the response given for time $t + 1$ affects that given for time t. Note also, that an additional variable G with four categories should be added to the diagram, to account for rotation group membership. All indicators depend on G, since units in different groups are interviewed in different calendar months.

The basic equation of the model, decomposes the proportion in the generic cell of the 9-way contingency table, in the product of the conditional probabilities reported in Appendix A, equations (A1) to (A7). A preliminary version of the model has been proposed in Bassi, Croon, Hagenaars and Vermunt (1995).

Equation (A1) defines the probability of belonging to one of the four rotation groups. Equations (A2) and (A3) define the initial condition, and the transition probabilities, of the latent first order Markov chain respectively. Equations (A4) and (A5) define the response probabilities for indicator Y_t; equations (A6) and (A7) the analogous probabilities for the dichotomous indicator W_t. The response probabilities are defined in such a way that the answer given for a certain month, depends jointly on the current true state (y_t) and on the "past" true and "past" reported states (y_{t+1} and Y_{t+1}). The term "past" refers to the way respondents think, while answering retrospective questions: they start recalling from the moment of time nearest to the interview, and go backwards up to the end of the reference period.

A complex set of constraints has been imposed on response probabilities of (A4), (A5), (A6) and (A7), to account for (i) the conditioning effect, and (ii) the fact that the four rotation groups are equivalent samples in terms of the error generating mechanism.

These constraints are formulated in detail in Appendix A. Basically, they incorporate *a priori* knowledge on respondent's behaviour, and allow us to specify a parsimonious model. Specifically, equations (A8) to (A14) correspond to the following statements:

(a) With regard to WW classification errors, following Hubble and Judkins (1989), it is assumed that:
  - (a1) a respondent who reports wrongly his/her labour force state for a certain month, continues to repeat this same answer also for the adjacent month, going backwards within the wave (A8);
  - (a2) if, however, the status at time $t + 1$ is correctly reported, the response probability for the adjacent month depends only on the current true state (A9);
  - (a3) the same error generating mechanism operates for both indicators. For W_t, we state that a correct answer is given when the true state is E and 'Job' is reported and when the true state is U or N and 'No Job' is reported, (A10) and (A11).

(b) Response probabilities are set equal across rotation groups, (A12) to (A15). As an example, equalities in (A12) mean that response probabilities for individuals in rotation group 1 for the month of April, are equal to response probabilities for individuals in group 4 for the month of March, to those for individuals in group 3 for the month of February, and to those for individuals in group 2 for the month of January. (They are set to be equal, since they all refer to the answer given for the last month of the wave.)

The model has been estimated to correct observed monthly gross flows for the quarter January to April 1986 (Table 2). The comparison between observed and estimated flows, highlights that the model reduces the seam effect: WW transitions are corrected towards a more dynamic labour market; BW transitions are corrected in the opposite direction. It is worth noting, that effects of model correction are more evident for flows from unemployment, which are characterised by higher mobility.

The goodness of fit of the model has been judged by multiple criteria such as the BIC index and the conditional test for nested models, together with estimate interpretability and consistency, with substantive knowledge of the dynamics of the U.S. labour market in '80s.

4.3 The French Labour Force Survey Data

The second case-study refers to the flows in the labour market, observed with the French Labour Force Survey (FLFS) conducted yearly by INSEE in France.

Table 2
SIPP Observed and Estimated Monthly Transition Rates (×100), January to April 1986

		EE	EU	EN	UE	UU	UN	NE	NU	NN
J-F	WW	98.11	1.17	0.72	14.53	80.16	5.31	0.90	1.57	97.53
	BW	94.08	2.17	3.75	23.58	44.30	32.12	5.62	3.45	90.93
	Estimated	97.25	1.47	1.28	16.08	77.16	6.76	1.59	1.32	97.09
F-M	WW	98.66	0.92	0.42	16.06	78.67	5.27	0.64	1.65	97.71
	BW	94.88	1.91	3.21	21.90	48.54	29.56	4.99	4.11	90.90
	Estimated	97.83	1.20	0.97	19.40	74.01	6.59	1.21	1.50	97.29
M-A	WW	98.71	0.64	0.65	20.76	71.74	7.50	1.47	1.05	97.48
	BW	95.59	1.52	2.89	30.48	34.92	34.60	6.34	3.78	89.88
	Estimated	98.11	0.95	0.94	26.42	65.75	7.83	2.17	0.71	97.12

The reference population of the FLFS are all members of French households, who are above 15 in the year in which the interview is planned. The survey has a rotating design: each year, one third of the sample is renewed.

Information on labour force participation is collected with retrospective questions, having as a reference period the 13 months preceding the interview. Each respondent is asked to recall his/her position in the labour market on a monthly basis, by filling in a grid in which he/she can classify himself/herself, for each month, over eight categories: self-employed, employed on a fixed term basis, permanently employed, unemployed, on training, student, serving in the Army, other (retired, housewife, *etc.*).

For our analysis, we aggregated the eight categories in the usual three states E, U and N. We consider 'Employed' respondents who classify themselves in the first three categories, 'Unemployed' those who classify themselves in the fourth category and 'Not in the labour force' the remaining ones.

We analyze the information collected in the two consecutive waves of March 1991 and March 1992, on a subsample of individuals: those who answered to three consecutive interviews (January 1990, March 1991 and March 1992) and who were 18 to 29 years old in 1992, for a total of 5,427 individuals. The reference periods of the two waves considered, overlap in March 1991. We have, then, two pieces of information on the labour force state for this month: one collected in March 1991, and the other one collected with a retrospective question 12 months afterwards.

The pattern of observed monthly transitions in our FLFS sample shows some interesting evidence, largely dictated by the characteristics of the subsample – young people.

Transitions exhibit a moderate degree of seasonal variation, related to the school calendar. From June to July, for example, we observe a proportion of people who enter the labour market as employed, greater than the average; on the contrary, from August to September, a proportion greater than the average leaves employment (presumably to education).

The marginal distribution of the three states from March 1990 to March 1992, shows that the individuals in our sample progressively enter the labour market: in March 1990, 44% are observed to be Employed or Unemployed, whereas by March 1992, this proportion has risen to 54%.

The double information for March 1991, provides some crude evidence on response error in the data: 8% of respondents declare a different state in the two interviews. For the period from February to April 1991, two types of flows may be observed: a within wave (WW) one, *i.e.*, information about the labour force state is collected in the same interview, and a between waves (BW) one, *i.e.*, information is collected in two different interviews (Table 3).

Table 3
FLFS Observed Monthly Transition Rates (×100) from February to April 1991

		EE	EU	EN	UE	UU	UN	NE	NU	NN
F-M	WW	98.19	1.67	0.14	9.11	90.65	0.24	0.28	0.11	99.61
	BW	93.17	3.58	3.25	25.18	65.23	9.59	3.75	1.96	94.29
M-A	WW	98.60	1.04	0.36	8.89	90.37	0.74	0.24	0.29	99.47
	BW	93.24	3.33	3.43	25.90	63.79	10.31	3.79	2.07	94.14

As expected, WW transitions describe a more stable labour market than the BW ones. This can be considered as an indication of correlated classification errors in the data. Patterns and causes of errors correlation in retrospective surveys, have been extensively discussed in the two previous sections, and the above considerations can largely be extended to the FLFS data.

In general, we expect that, in a retrospective survey with such a long recall period, lack of memory results in the major cause of classification errors. We also expect that the probability of answering incorrectly, increases as the distance between the reference month and the interview month gets longer. This may be considered as the major source of correlation among classification errors, together with telescoping and conditioning effects, which possibly affect FLFS data as well (see Magnac and Visser 1995).

The overall effect of correlated classification errors, reasonably results in an underestimation of mobility in the French labour market.

Moving from these considerations, we specified a model to correct observed quarterly gross flows, from measurement error (Table 4). The last column of Table 4 contains the percentage of individuals who are observed to change state, between the two months considered (OM = observed mobility). On the average over the five WW transitions, 6.122% of mobility between two consecutive months is observed.

As in the previous case-studies, let us denote with $y_t (t = 1, 2, 3, 4, 5, 6)$ true labour force states, and with upper case letters their indicators: $Y_t (t = 2, 3, 4, 5, 6)$ represents labour force states observed in March 1992 (referring to March, June, September, December 1991 and March 1992); $W_t (t = 1,2)$ represents labour force states observed in March 1991 (referring to December and March 1991). As usual, y_t, Y_t and W_t distribute over the three categories of E, U and N.

The model is specified by decomposing the proportion in the generic cell of the 7-way contingency table as in Appendix B, equations (B1) to (B6).

Since we observe two indicators only for one month, a model which assumes direct effects between the indicators, would be under identified. Thus, we can not explicitly model dependencies between observed states. The only way to account for correlated classification errors in FLFS data, is to let observed states depend on latent transitions. By the way, this seems to be a sensible assumption in retrospective surveys. Indeed, flows between two different states may easily undergo wrong placements in time, because in some situations, events might truly be difficult to place exactly. As an example, employees who loose their job or retire (flows EU and EN), will generally use the holidays they are entitled to, and may not clearly know when they exactly left employment. The moment people entered the labour force, may also be hard to recall, especially when they left school (flows NU and NE) (van de Pol and Langeheine 1997).

The modified LISREL model, formulated in mathematical terms in Appendix B, is based on the following substantive assumptions.

At the latent level, transitions follow a first order non stationary Markov chain (equations (B1) and (B2)). Indeed, the evidence on seasonality in observed transitions, suggest avoiding the imposition of stationarity of any order, on the latent Markov chain.

Response probabilities for data collected in both waves, depend on the latent transition occurring between t and $t + 1$ (equations (B3) and (B4) refer to data collected in March 1992, equations (B5) and (B6) to data collected in March 1991).

In order to describe the error generating mechanism in detail, and specify a more parsimonious model, the following constraints have been imposed on response probabilities:

(a) response probabilities referring to the same month of subsequent years (December and March) are set equal;
(b) response probabilities at time t, given that the true state has not changed between time t and time $t + 1$, are set constant over time;
(c) response probabilities are set equal for June and September 1991;
(d) in general, respondents who move between month t and $t + 1$ (transitions EU, EN, UE, UN and NU), at time t, report either the true state occupied at time t, or the true state occupied at time $t + 1$, *i.e.*, they, do not report a state they have not been moved from/to;
(e) if however, the latent transition occurs between states N and E, we admit all three answers at time t, *i.e.*, we consider that people who find a job may confuse their previous position (at time t), and be uncertain between U and N.

Constraint (c) is imposed mainly for reasons of model parsimony. It captures the notion that response probabilities for months that are placed more or less in the central part of the reference period, do not vary too much.

Constraints (b) and (d) reflect the fact that response probabilities depend on latent transitions. We expect that these probabilities do not vary too much over time when there is no latent change (constraint (b)), whereas we expect that the probability of misplacing change, especially in ambiguous situations, increases with the length of the recall

Table 4
FLFS Observed Quarterly Transition Rates (×100), December 1990 to March 1992
(OM = Observed Mobility)

		EE	EU	EN	UE	UU	UN	NE	NU	NN	OM
D90-M91	WW	94.77	4.25	0.98	24.53	72.40	3.07	0.98	0.66	98.36	5.08
	BW	91.50	4.86	3.64	31.60	56.84	11.56	4.40	2.10	93.50	10.16
M91-J91	WW	96.03	3.02	0.95	23.21	74.32	2.47	1.28	0.68	98.04	4.54
	BW	91.48	4.63	3.89	35.01	54.20	10.79	4.84	2.14	93.02	12.04
J91-S91	WW	94.29	3.94	1.77	20.93	78.29	0.78	4.71	2.95	92.34	7.85
S91-D91	WW	93.73	4.48	1.79	23.63	74.89	1.48	3.22	1.65	95.13	7.23
D91-M92	WW	93.90	4.80	1.30	21.67	76.74	1.59	1.70	0.59	97.71	5.91

period. Constraints under (d) aim at catching the telescoping effect.

Figure 5 gives the path diagram of the estimated model.

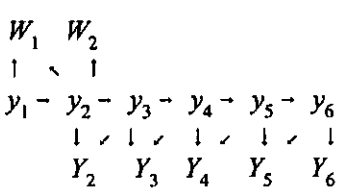

Figure 5. Path Diagram of a Modified Lisrel Model for Six Measurement Points and Two Indicators for One Latent Variable

Table B.1 in Appendix B reports the pattern of restrictions on response probabilities, (a) to (e); it shows which parameters are set equal, and which are fixed to 0, in order to introduce into the basic model, as defined by equations (B1) to (B6), the above constraints.

The final model has been selected after comparing a sequence of models, as can be seen from Table 5.

Table 5
Model Selection (EM = Estimated Mobility)

MODEL	L^2	df	ΔL^2	p-value cond. test	EM
A	2509.5759	2124			5.424
A1	3450.1716	2154	940.5957	0	4.918
A2	3849.9470	2178	399.7754	0	5.798
B	816.1620	2076			5.888
B1	855.2282	2094	39.0662	0.01	5.818
B2	864.9657	2106	9.7375	0.40	5.906
B3	879.5996	2121	14.6339	0.10	6.252

We started the analysis by estimating a model based on the ICE assumption (model A in the table), which, as expected, shows a bad fit.

The following models (A1 and A2) are based on the work by Magnac and Visser (1995). These authors consider monthly transitions over a period longer than ours (from January 1989 to March 1992), but on the same sample of individuals. They assume that the labour force state in the interview month is correctly reported, while the probability of making mistakes increases with the distance between the reference month and the time of interview, according to a deterministic function of time. Response probabilities are assumed to be constant over the survey waves, and true transitions are assumed to follow a first order stationary Markov chain. Our model A1 is a less restricted version of Visser and Magnac's model – no stationarity assumption is made, applied to quarterly transitions from December 1990 to March 1992. Our model A2 adds to model A1, the hypothesis of first order stationarity at the latent level. Both models perform quite badly, and (from column EM), we see that, on average, they correct the observed labour market towards stability: a result which contradicts the evidence on the effects of classification errors in retrospective surveys.

Model B introduces correlation among classification errors, by letting each indicator to depend on the true transition that occurred between times t and $t+1$; moreover, it encompasses constraint (a). The fit increases dramatically (see L^2). All subsequent models are nested in model B, and additional restrictions may be evaluated by a conditional test. Model B1 introduces constraints under (b); model B2 the additional constraints under (c); and model B3 is our final model.

Table 6 presents estimated transition rates with our best fitting model. The French labour market is corrected towards a greater mobility. The average estimated mobility amounts to 6.252%. Moreover, estimated response probabilities show a pattern consistent with the notion, that the probability of making mistakes gets bigger, the longer the recall period.

ACKNOWLEDGEMENTS

Research for this paper was supported by grants from CNR n. 94.02242.CT10 and MURST n. 02.09.02.110 and n. 02.09.02.124. We are grateful to Michael Visser and Thierry Magnac for providing us with anonimyzed individual data from the French Labour Force Survey. A preliminary version of this paper was presented at the IASS/IAOS Satellite Meeting on Longitudinal Studies, Jerusalem, August 27-31, 1997, where we benefited from comments and discussion. Detailed criticisms and suggestions from two referees are especially acknowledged.

Table 6
FLFS Estimated Quarterly Transition Rates (×100), December 1990 to March 1992
(EM = Estimated Mobility)

	EE	EU	EN	UE	UU	UN	NE	NU	NN	EM
D90-M91	94.85	4.48	0.67	12.70	66.28	21.02	1.09	1.55	97.36	6.27
M91-J91	95.65	1.37	2.98	28.43	62.35	9.22	3.61	1.48	94.91	7.49
J91-S91	93.71	4.25	2.04	14.88	82.50	2.62	4.11	3.49	92.40	7.70
S91-D91	98.32	1.67	0.01	15.42	83.75	0.83	3.80	0.47	95.73	4.24
D91-M92	93.23	5.02	1.75	9.99	88.65	1.36	2.07	1.28	96.65	5.56

APPENDIX A

Final Model Specification for the SIPP Data, in Terms of Conditional Probabilities

(1) Basic model decomposition

$$z_g = P(G = g) \tag{A1}$$

$$\pi_1^{j_1} = P(y_1 = j_1) \tag{A2}$$

$$\pi_t^{j_t j_{t-1}} = P(y_t = j_t \mid y_{t-1} = j_{t-1}) \qquad t = 2, 3, 4 \tag{A3}$$

$$q_{yt}^{l_t j_t l_{t+1} j_{t+1} g} = P(Y_t = l_t \mid y_t = j_t, Y_{t+1} = l_{t+1}, y_{t+1} = j_{t+1}, G = g) \qquad t = 1, 2, 3 \tag{A4}$$

$$q_{y4}^{l_4 j_4 g} = P(Y_4 = l_4 \mid y_4 = j_4, G = g) \tag{A5}$$

$$q_{wt}^{m_t j_t m_{t+1} j_{t+1} g} = P(W_t = m_t \mid y_t = j_t, W_{t+1} = m_{t+1}, y_{t+1} = j_{t+1}, G = g) \qquad t = 1, 2, 3 \tag{A6}$$

$$q_{w4}^{m_4 j_4 g} = P(W_4 = m_4 \mid y_4 = j_4, G = g) \tag{A7}$$

g varies over 1, 2, 3 and 4; l_t and $j_t, t = 1, 2, 3, 4,$ vary over the categories E, U and N, $m_t, t = 1, 2, 3, 4,$ vary over the categories 'Job' and 'No Job'.

(2) Constraints on conditional probabilities

$$q_{yt}^{l_{t+1} j_t l_{t+1} j_{t+1} g} = P(Y_t = l_{t+1} \mid y_t = j_t, Y_{t+1} = l_{t+1}, y_{t+1} = j_{t+1}, G = g) = 1 \qquad \text{for } l_{t+1} \neq j_{t+1}\ t = 1, 2, 3 \tag{A8}$$

$$q_{yt}^{l_t j_t l_{t+1} j_{t+1} g} = P(Y_t = l_t \mid y_t = j_t, G = g) \qquad \text{for } l_{t+1} = j_{t+1} \text{ and } t = 1, 2, 3 \tag{A9}$$

$$q_{wt}^{m_{t+1} j_t m_{t+1} j_{t+1} g} = P(W_t = m_{t+1} \mid y_t = j_t, W_{t+1} = m_{t+1}, y_{t+1} = j_{t+1}, G = g) = 1 \qquad \text{for } m_{t+1} \neq j_{t+1} \text{ and } t = 1, 2, 3 \tag{A10}$$

$$q_{wt}^{m_t j_t m_{t+1} j_{t+1} g} = P(W_t = m_t \mid y_t = j_t, G = g) \qquad \text{for } m_{t+1} = j_{t+1} \text{ and } t = 1, 2, 3 \tag{A11}$$

$$q_{t=4}^{g=1} = q_{t=3}^{g=4} = q_{t=2}^{g=3} = q_{t=1}^{g=2} \tag{A12}$$

$$q_{t=4}^{g=2} = q_{t=3}^{g=1} = q_{t=2}^{g=4} = q_{t=1}^{g=3} \tag{A13}$$

$$q_{t=4}^{g=3} = q_{t=3}^{g=2} = q_{t=2}^{g=1} = q_{t=1}^{g=4} \tag{A14}$$

$$q_{t=4}^{g=4} = q_{t=3}^{g=3} = q_{t=2}^{g=2} = q_{t=1}^{g=1} \tag{A15}$$

APPENDIX B

Final Model Specification for the FLFS Data, in Terms of Basic Model Decomposition and Pattern of Restrictions on Parameters

(1) Basic model decomposition

$$\pi_t^{j_1} = P(y_1 = j_1) \tag{B1}$$

$$\pi_t^{j_t j_{t-1}} = P(y_t = j_t \mid y_{t-1} = j_{t-1}) \qquad t = 2, 3, 4, 5 \tag{B2}$$

$$q_{yt}^{l_t j_t j_{t+1}} = P(Y_t = l_t \mid y_t = j_t, y_{t+1} = j_{t+1}) \qquad t = 2, 3, 4, 5 \tag{B3}$$

$$q_{y6}^{l_6 j_6} = P(Y_6 = l_6 \mid y_6 = j_6) \tag{B4}$$

$$q_{wt}^{m_1 j_1 j_2} = P(W_1 = m_1 \mid y_1 = j_1, y_2 = j_2) \tag{B5}$$

$$q_{w2}^{m_2 j_2} = P(W_2 = m_2 \mid y_2 = j_2) \tag{B6}$$

j_t, l_t and m_t vary over E, U and N.

(2) Pattern of restrictions on response probabilities

Table B.1
Month of Observation

Probability of observing a state given a latent transition	March 91	June 91 & Sept. 91	Dec. 90 & Dec. 91
E|ee	1	1	1
U|ee	2	2	2
N|ee	3	3	3
E|eu	F	F	F
U|eu	F	F	F
N|eu	*	*	*
E|en	F	F	F
U|en	*	*	*
N|en	F	F	F
E|ue	F	F	F
U|ue	F	F	F
N|ue	*	*	*
E|uu	4	4	4
U|uu	5	5	5
N|uu	6	6	6
E|un	*	*	*
U|un	F	F	F
N|un	F	F	F
E|ne	F	F	F
U|ne	F	F	F
N|ne	F	F	F
E|nu	*	*	*
U|nu	F	F	F
N|nu	F	F	F
E|nn	7	7	7
U|nn	8	8	8
N|nn	9	9	9

Equal numbers indicate response probabilities fixed to be equal.
* indicates a probability fixed to 0.
F indicates a free parameter.

REFERENCES

ABOWD, J.M., and ZELLNER A. (1985). Estimating gross labour force flows. *Journal of Business and Economics Statistics*, 3, 254-283.

BASSI, F., CROON, M., HAGENAARS, J., and VERMUNT, J. (1995). Estimating Latent Turnover Tables When Data are Affected by Correlated and Uncorrelated Classification Errors. WORC PAPER 95.12.25/7, Tilburg University.

BERNARD, H.R., KILLWORTH, P., KRONENFELD, D., and SAILER, L. (1984). The problem of informant accuracy: The validity of retrospective data. *Annual Review of Anthropology*, 13, 495-517.

CHAKRABARTRY, R.P., and WILLIAMS, T.R. (1989). Time-in-sample biases in the Survey of Income and Program Participation. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 309-314.

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

CITRO, C.F., and KALTON, G. (1993). *The Future of the SIPP*. Washington, D.C.: National Academy Press.

DUNCAN, G., and KALTON, G. (1987). Issues of design and analysis of surveys across time. *International Statistical Review*, 55, 97-117.

GOODMAN, L.A. (1973). The analysis of a multidimensional contingency table when some variables are posterior to the others. *Biometrika*, 60, 179-192.

GOODMAN, L.A. (1974). Exploratory latent structure analysis using both identifiable and unidentifiable models. *Biometrika*, 79, 1178-1259.

GOODMAN, L.A. (1981). Three elementary views of log-linear models for the analysis of cross-classifications having ordered categories. In *Sociological Methodology*, (Ed. S.Leinhardt), 193-293. San Francisco: Jossey Bass.

HABERMAN, S.J. (1978). *Analysis of Qualitative Data. Vol.1. Introductory Topics*. New York: Academic Press.

HAGENAARS, J.A. (1988). Latent structure models with direct effects between the indicators, local dependence models. *Sociological Methods and Research*, 16, 379-405.

HAGENAARS, J.A. (1990). *Categorical Longitudinal Data: Log-linear, Panel, Trend and Cohort Analysis*. Newbury Park: Sage.

HAGENAARS, J.A. (1997). Categorical Causal Modeling: Directed Loglinear Models With Latent Variables. WORC PAPER 97.04.002/7, Tilburg University.

HUBBLE, D.L., and JUDKINS, D.R. (1989). Measuring the Bias in Gross Flows in the Presence of Autocorrelated Response Errors. SIPP Working Paper No. 8712, U.S. Bureau of the Census.

HUSSMAN, R., MEHRAN, F., and VERMA, M. (1990). *Surveys of Economically Active Population, Employment and Underemployment: an ILO Manual on Concepts and Definitions*. Geneva: ILO.

JORESKÖG, K.G., and SÖRBOM, D. (1988). *Lisrel 7: A Guide to the Program and Applications*. Chicago: SPSS INC.

KALTON, G., and CITRO, C.F. (1993). Panel surveys: adding the fourth dimension. *Survey Methodology*, 19, 205-215.

KALTON, G., and MILLER, M.W. (1991). The seam effect with social security income in the SIPP. *Journal of Official Statistics*, 7, 235-245.

LANGEHEINE, R., PANNEKOEK, J., and van de POL, F. (1995). Bootstrapping goodness-of-fit measures in categorical data analysis. *Sociological Methods and Research*, 24, 492-516.

LAZARSFELD, P.F., and HENRY, N.W. (1968). *Latent Structure Analysis*. New York: Houghton Mufflin.

MAGNAC, T., and VISSER, M. (1995). Transition Models With Measurement Errors. Working Paper, Institut National de la Recherche Agronomique (INRA), Paris.

MARQUIS, K.H., and MOORE, J.C. (1989). Some response errors in SIPP with thoughts about their effects and remedies. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 381-386.

MARTINI, A. (1988). Retrospective versus panel data in estimating labour force gross flows: comparing SIPP and CPS. *Proceedings of the Social Science Section, American Statistical Association*, 109-114.

MARTINI, A. (1989). Seam effect, recall bias, and the estimation of labour force transition rates from SIPP. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 387-392.

MENG, X.L., and RUBIN, D.B. (1993). Maximum likelihood estimation via ECM algorithm: a general framework. *Biometrika*, 80, 267-278.

McCORMICK, M., BUTLER, D., and SINGH, R. (1992). Investigating time-in-sample effects for the Survey of Income and Program Participation. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 554-559.

O'MUIRCHEARTAIGH, C. (1996). Measurement errors in panel surveys: implications for survey design and for survey instruments. *Proceedings of the Scientific Reunion of the Italian Statistical Society*, 1, 207-218. Rimini: Maggioli.

PFEFFERMAN, D., SKINNER, C.J., and HUMPHREYS, K. (1998). The estimation of gross flows in the presence of measurement error using auxiliary variables. *Journal of the Royal Statistical Society*, Series A, 161, 13-32.

POTERBA, J.M., and SUMMERS, L.H. (1986). Reporting errors and labor market dynamics. *Econometrica*, 54, 1319-1338.

SINGH, A.C., and RAO, J.N.K. (1995). On the adjustment of gross flows estimates for classification errors with application to data from the Canadian Labor Force Survey. *Journal of the American Statistical Association*, 90, 1-11.

SKINNER, C.J., and TORELLI, N. (1993). Measurement error and the estimation of gross flows from longitudinal economic data. *Statistica*, 3, 391-405.

U.S. DEPARTMENT OF COMMERCE (1991). *SIPP User's Guide*. Washington D.C.

van de POL, F., and LANGEHEINE, R. (1990). Mixed Markov latent class models. In *Sociological Methodology*, (Ed. C.Clogg), 213-247. Oxford: Blackwell.

van de POL, F., and LANGEHEINE, R. (1992). Analysing Measurement Error in Quasi-experimental Data: An Application of Latent Class Models to Labour Market Data. Working Paper of the European Scientific Network on household Panel Studies, 57, Colchester, University of Essex.

van de POL, F., and LANGEHEINE, R. (1997). Separating change and measurement error in panel surveys with an application to labour market data. In *Survey Measurement and Process Quality*, (Ed. L.Lyberg *et al.*), 671-688. New York: Wiley.

VERMUNT, J.K. (1993). Log-linear and event history analysis with missing data using the EM algorithm. WORC PAPER 93.09.015/7, Tilburg University.

YOUNG, N. (1989). Wave seam effects in the SIPP. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 393-398.

Survey Methodology, December 1998
Vol. 24, No. 2, pp. 123-129
Statistics Canada

Estimating Labour Force Gross Flows From Surveys Subject to Household-level Nonignorable Nonresponse

PAUL S. CLARKE and RAY L. CHAMBERS[1]

ABSTRACT

Measurement of gross flows in labour force status is an important objective of the continuing labour force surveys carried out by many national statistics agencies. However, it is well known that estimation of these flows can be complicated by nonresponse, measurement errors, sample rotation and complex design effects. Motivated by nonresponse patterns in household-based surveys, this paper focuses on estimation of labour force gross flows, while simultaneously adjusting for nonignorable nonresponse. Previous model-based approaches to gross flows estimation have assumed nonresponse to be an individual-level process. We propose a class of models that allow for nonignorable household-level nonresponse. A simulation study is used to show, that individual-level labour force gross flows estimates from household-based survey data, may be biased and that estimates using household-level models can offer a reduction in this bias.

KEY WORDS: Gross flows; Household-based surveys; Nonignorable nonresponse.

1. INTRODUCTION

Labour force gross flows are typically defined as transitions over time between the three major labour force states, employed, unemployed and economically inactive. Gross flows estimates are an important tool in the study of labour force dynamics (for example, see Vanski 1985). Large-scale on-going surveys such as the British Labour Force Survey and the U.S. Current Population Survey, provide data for gross flows estimation. However, nonresponse, measurement error, sample rotation and complex design effects, affect gross flows estimation from these surveys. A discussion of these and other factors affecting gross flows estimation, is given in Hogue (1985). Here we focus on the problem of nonresponse.

We assume that a nonresponse mechanism leads to the observed data being incomplete. If the probability of not responding depends on the missing data, then the nonresponse mechanism is nonignorable (Rubin 1976). The model-based approach to analysing incomplete survey data, is detailed in Little (1982). Model-based approaches to the estimation of labour force gross flows, involve modelling both the labour force flows and the nonresponse mechanism, and simultaneously fitting both models to the incomplete data. Examples of such models are given in Stasny and Fienberg (1985), Stasny (1986) and, for nonignorable nonresponse, in Little (1985). We call these individual-level models, because individuals are modelled as responding or not responding, independently of other sampled individuals.

Both the Labour Force Survey and the Current Population Survey, are examples of household-based surveys, that is, surveys based on a random sample of households, rather than individuals. Household-based surveys can lead to correlated nonresponse behaviour within households. For example, in the Current Population Survey, a single household member (usually the head-of-household) acts as a proxy for the other household members; thus, if the chosen household member is a nonrespondent, so are other household members. It follows that, due to correlated within-household nonresponse behaviour, individual-level nonresponse models are unsuitable for the estimation of labour force gross flows, using household-based survey data.

In this paper, we propose a class of models for individual-level labour force flows, and household-level nonresponse, that account for correlated within-household nonresponse behaviour. A number of plausible nonresponse models that are estimable from the observed data, both ignorable and nonignorable, are also presented. We then simulate household-based survey data, using these household-level models, to demonstrate the potential utility of our approach: first, individual-level labour force gross flows estimates are shown to be biased, when fitted to household-based survey data; and second, the bias of individual-level and household-level gross flows estimates are compared, to show the advantages of fitting household-level models to household-based survey data. To conclude, we summarise the findings of our simulation studies and discuss ideas for further research in this area.

2. A MODEL FOR HOUSEHOLD-LEVEL NONRESPONSE

2.1 The Data

A gross flow is the probability or frequency of individuals in the population, making a state transition between two points in time, t_1 and $t_2 (t_1 < t_2)$. Labour force gross flows refer to transitions between the three main

[1] Paul S. Clarke and Ray L. Chambers, Department of Social Statistics, University of Southampton, Highfield, Southampton, SO17 1BJ, United Kingdom.

labour force states: 1 = 'employed', 2 = 'unemployed' and 3 = 'not in labour force', where the last category refers to economically inactive individuals, such as retired individuals and students. Let S denote a simple random sample of households, indexed by h. Within household h, there are n_h eligible individuals, of which $n_h(ab)$ have labour force flow (a, b) between t_1 and t_2, where $\sum_{a,b} n_h(ab) = n_h$, and $a, b = 1, 2, 3$. We refer to $\{n_h(ab)\}$ as the complete data, that is, the frequencies that would be observed in the absence of nonresponse.

Table 1 shows the complete labour force flows data for household h as a 3×3 contingency table. If h responds at both times, the observed data are the cells of this 2-way table. However, if the household does not respond at t_1 or t_2, the observed data correspond to the margins of the table: $n_h(1+), n_h(2+), n_h(3+)$ are the observed data if h responds at t_1, but does not respond at t_2; and $n_h(+1)$, $n_h(+2), n_h(+3)$ are the observed data if h responds at t_2 but does not respond at t_1. (An index replaced by '+' denotes summation over all levels of that index.) Furthermore, if h does not respond at both t_1 and t_2, the observed data is the household size, n_h, which we take to be known and fixed between t_1 and t_2.

Table 1
Complete Labour Force Flows Data for Household h

Status		t_2 1	2	3	
t_1	1	$n_h(11)$	$n_h(12)$	$n_h(13)$	$n_h(1+)$
	2	$n_h(21)$	$n_h(22)$	$n_h(23)$	$n_h(2+)$
	3	$n_h(31)$	$n_h(32)$	$n_h(33)$	$n_h(3+)$
		$n_h(+1)$	$n_h(+2)$	$n_h(+3)$	n_h

2.2 Model Specification

It is inappropriate to treat the nonresponse behaviour of individuals within a household as independent, in household-based surveys. In the Labour Force Survey, for example, one eligible household member determines whether the household can be interviewed. Therefore, if no eligible individual can be contacted, each household individual is a nonrespondent. To construct a model for household-level nonresponse, we take the ideas behind individual-level nonresponse and extend them to the household, by considering a household to be an entity with its own nonresponse flow between t_1 and t_2. To allow for nonignorable nonresponse, the probability of a household nonresponse flow is modelled as a function of its individual labour force flows, as shall now be described.

Let $\boldsymbol{N}_h = (N_h(11), N_h(21), ..., N_h(33))$ be the random vector of labour force flows frequencies for household h, where $N_h(ab)$ is the random variable, whose outcome corresponds to the number of individuals with labour force flow (a, b), $a, b = 1, 2, 3$. Further, denote the random vector for the nonresponse flow of household h by $\boldsymbol{R}_h = (R_{h1}, R_{h2})$, where

$$R_{hj} = \begin{cases} 1, & \text{if household responds at } t_j \\ 0, & \text{otherwise} \end{cases}$$

is the nonresponse status random variable for h at t_j, $j = 1, 2$. The realisations of these random quantities are denoted by $\boldsymbol{n}_h$ and $\boldsymbol{r}_h$. We now assume that $\boldsymbol{n}_h$ and $\boldsymbol{r}_h$ are known, and write the joint probability of $\boldsymbol{N}_h$ and $\boldsymbol{R}_h$ as

$$\Pr(\boldsymbol{N}_h = \boldsymbol{n}_h, \boldsymbol{R}_h = \boldsymbol{r}_h) = \Pr(\boldsymbol{N}_h = \boldsymbol{n}_h)\Pr(\boldsymbol{R}_h = \boldsymbol{r}_h \mid \boldsymbol{N}_h = \boldsymbol{n}_h),$$

where $\Pr(\boldsymbol{N}_h = \boldsymbol{n}_h)$ is the labour force flows model, and $\Pr(\boldsymbol{R}_h = \boldsymbol{r}_h \mid \boldsymbol{N}_h = \boldsymbol{n}_h)$ is called the nonresponse flows model.

The labour force flows model is taken to be multinomial, with probability function

$$\Pr(\boldsymbol{N}_h = \boldsymbol{n}_h; \boldsymbol{\omega}) = n_h! \prod_{a,b} \frac{\omega(ab)^{n_h(ab)}}{n_h(ab)!}, \tag{1}$$

where $\omega(ab) > 0$ is the probability of an individual having labour force flow (a, b) and $\sum_{a,b} \omega(ab) = 1$. The vector of labour force flows parameters is denoted by $\boldsymbol{\omega} = (\omega(11), \omega(21), ..., \omega(33))$, of which 8 are free. The assumption of multinomial sampling in (1), implies that individuals' labour force flows behaviour, is independent within households, and that households are homogeneous with respect to their labour force flows behaviour. These assumptions are unrealistic, but (1) can easily be extended to a more realistic model for the labour force flows, as we discuss in Section 4.

The probability of household h having nonresponse flow (u, v), is taken to be

$$\pi(uv \mid \boldsymbol{n}_h) = \Pr(\boldsymbol{R}_h = (u, v) \mid \boldsymbol{N}_h = \boldsymbol{n}_h; \boldsymbol{\psi}) = \frac{1}{n_h} \sum_{a,b} n_h(ab)\psi(uv \mid ab), \tag{2}$$

for $u, v = 0, 1$, namely, a weighted average of the nonresponse model parameters. By setting $n_h = 1$, it can be seen that $\psi(uv \mid ab) > 0$ is the probability of a household of size one (*i.e.*, an individual) having nonresponse flow (u, v), given it has labour force flow (a, b). Thus, $\sum_{u,v} \psi(uv \mid ab) = 1$ and $\boldsymbol{\psi} = (\psi(11 \mid 11), \psi(01 \mid 11), ..., \psi(00 \mid 33))$ is the vector of nonresponse parameters, of which 27 are free.

Before defining the likelihood function for the complete data, partition S into 4 mutually exclusive and exhaustive subsets

$$S = S_{11} \cup S_{01} \cup S_{10} \cup S_{00},$$

where $S_{uv} = \{h : r_h = (u, v)\}$ is the subset of households with nonresponse flow (u, v). Thus, since S is a simple random sample of households, the likelihood function for the complete data is

$$L(\omega, \psi; \{n_h, r_h\}) = \prod_{u,v} \prod_{h \in S_{uv}} L_h(\omega, \psi; n_h, (u, v)), \quad (3)$$

where $L_h(\omega, \psi; n_h, (u, v))$ is the contribution of household $h \in S_{uv}$ to the likelihood, the product of (1) and (2).

2.3 Model Fitting

2.3.1 Maximum Likelihood Estimation

Since the complete data are unavailable, (3) must be modified to give the likelihood based on the observed data. Denote the observed data by $\{n_h^*\}$. As discussed in Section 2.1, the observed data for households that respond at t_1 and t_2, is the full cross-classification in Table 1, namely, $n_h^* = n_h$. Similarly, if $h \in S_{10}$ then $n_h^* = (n_h(1+), n_h(2+), n_h(3+))$; if $h \in S_{01}$ then $n_h^* = (n_h(+1), n_h(+2), n_h(+3))$; and if $h \in S_{00}$, then $n_h^* = n_h$.

The contribution of household $h \in S_{uv}$ to the observed data likelihood, is obtained by summing $L_h(\omega, \psi; n_h, (u, v))$ over all possible values that the full 3×3 cross-classification of labour force flows can take, given the observed margin. Representing this set of tables by $n_h : n_h^*$, the observed data likelihood for S is

$$L(\omega, \psi; \{n_h^*, r_h\}) = \prod_{u,v} \prod_{h \in S_{uv}} \sum_{n_h : n_h^*} L_h(\omega, \psi; n_h, (u, v)). \quad (4)$$

Model fitting requires calculating (4) at each stage of an iterative optimization process. This is computationally intensive, because the complete data likelihood function must be summed explicitly over the missing data. For example, the observed data for $h \in S_{10}$ is $n_h^* = (n_h(1+), n_h(2+), n_h(3+))$ and the likelihood contribution of this household to the observed data likelihood is

$$\sum_{n_h : n_h^*} L_h(\omega, \psi; n_h, (1, 0)).$$

To explicitly calculate this contribution, each 3×3 complete data table n_h for fixed n_h^* is generated and $L_h(\omega, \psi; n_h, (1, 0))$ evaluated for each. For household size $n_h = 5$, there are at least 21 and at most 108 possible tables, depending on the values in the fixed margin; for $n_h = 15$, a very large household size, the respective numbers are 136 and 9,261. A similar procedure is used for $h \in S_{01}$, except here $n_h^* = (n_h(+1), n_h(+2), n_h(+3))$ is the fixed margin. If $h \in S_{00}$, then no data about labour force status are observed, only the household size n_h. So each 3×3 table with total n_h must be generated, and the likelihood function calculated for each: for $n_h = 5$ there are 1,287 tables and for $n_h = 15$ there are 490,314. It is not infeasible, in terms of computer run-time, to calculate such sums directly. The number of explicit calculations can be reduced, by recognising that each household is defined only by its observed labour force flows frequencies and nonresponse flow. Thus, summation over the missing data need only be performed once for a household with a particular nonresponse flow and labour force flows frequencies; the contribution of this household to the likelihood is then raised to the power of the number of similarly defined households in S.

2.3.2 Parameter Estimability

If we fix $n_h = 1$ for all h, the complete data have no household structure, and form a 4-way table cross-classified by labour force status and nonresponse status at t_1 and t_2. The observed data log-likelihood (4) is now equivalent to that of the individual-level models in Stasny and Fienberg (1985), Little (1985) and Stasny (1986). For these models, estimability requires that the number of model parameters does not exceed 15 (one for each observed table cell, less one for the multinomial sampling constraint). Hence, (ω, ψ) are inestimable because there are $8 + 27 = 35$ free parameters. Since interest is focused on the labour force gross flows probabilities, ω, it is neccessary to constrain ψ to ensure estimability.

When $n_h > 1$, determining parameter estimability is more difficult, because (4) has a complicated closed-form expression. Fitzmaurice, Laird and Zahner (1996) use a numerical method to determine estimability, that involves showing that the information matrix is non-singular in the neighbourhood of the maximum likelihood estimate. However, not only is this impractical for problems of a high dimension, but evaluating the information matrix for the household-level model, is particularly difficult in this case. Instead, we adopt a pragmatic approach for determining parameter estimability: first, we restrict attention to models that satisfy the necessary condition for estimability when $n_h = 1$; and second, different starting values are used to for each fit. If the different starting values reveal a non-unique maximum likelihood estimate, or any parameter estimate is unchanged from its starting value then the model parameters are taken to be inestimable.

2.4 Nonresponse Models

To enable parameter estimates to be obtained from the observed data, θ and ψ must be constrained in accordance with assumptions about the nonresponse mechanism. The nonresponse parameters are interpreted as individual nonresponse probabilities, but within the household framework established thus far, it is inappropriate to talk about individuals not responding. However, in reality, it is individuals within households that determine a household's nonresponse flow, not the household itself. Therefore, constraints are placed on the nonresponse parameters at the individual level, that apply at the household level through the functional dependence of $\pi(uv \mid n_h)$ on ψ in (2). For example, if the nonresponse parameters are constrained such that $\psi(uv \mid ab) = \psi(uv)$ for all a, b, then the household nonresponse mechanism is ignorable, because household

nonresponse flows are independent of the labour force flows.

We now present four models for the nonresponse mechanism, two of which are ignorable, and two nonignorable.

- Ignorable models.
  - Model I_A: Constant nonresponse probability,

$$\psi(uv \mid ab) = \lambda^{1-u}(1-\lambda)^u \times \lambda^{1-v}(1-\lambda)^v,$$

    which has 1 parameter, λ, the probability of an individual not responding;
  - Model I_B: Independent of labour force status, but different nonresponse probabilities, at t_1 and t_2,

$$\psi(uv \mid ab) = \lambda^{1-u}(1-\lambda)^u \times \theta^{1-v}(1-\theta)^v,$$

    which has 2 parameters, λ, θ, the probabilities of nonresponse at t_1 and t_2, respectively.
- Nonignorable models.
  - Model N_A: The nonresponse distributions at t_1 and t_2 are independent but depend on labour force status at t_1 and t_2, respectively,

$$\psi(uv \mid ab) = \lambda(a)^{1-u}(1-\lambda(a))^u \times \theta(b)^{1-v}(1-\theta(b))^v$$

    which has 6 parameters, $\boldsymbol{\lambda} = (\lambda(1), \lambda(2), \lambda(3))$ and $\boldsymbol{\theta} = (\theta(1), \theta(2), \theta(3))$, where $\lambda(a)$ is the probability of not responding at t_1, given labour force status a at t_1, and $\theta(b)$ that at t_2, given labour force status b at t_2;
  - Model N_B: The nonresponse distributions at t_1 and t_2 depend on labour force status at t_1 and t_2 respectively, *i.e.*, a first-order Markov process. Unlike N_A, the nonresponse distributions at t_1 and t_2 are dependent: if the nonresponse status at t_1 is 1, then the nonresponse distribution at t_2 is the same as at t_1; but if the nonresponse status at t_1 is 0, the nonresponse distributions are distinct,

$$\psi(uv \mid ab) = \lambda(a)^{1-u}(1-\lambda(a))^u \times \begin{cases} \lambda(b)^{1-v}(1-\lambda(b))^v, & \text{if } u = 1, \\ \theta(b)^{1-v}(1-\theta(b))^v, & \text{if } u = 0, \end{cases}$$

for $a, b = 1, 2, 3$ and $u, v = 0, 1$. Under model I_A, there are a total of $8 + 1 = 9$ free parameters, satisfying the necessary condition for estimability of an individual-level model. Models I_B, N_A and N_B have 10, 14 and 14 free parameters, respectively, and so also satisfy the necessary condition for estimability.

3. SIMULATION STUDY

3.1 Simulation Procedure

We used a simulation study to investigate the consequences of failing to account for the household structure of household-based survey data, and to compare labour force gross flows estimates for individual-level and household-level models. For this purpose, household-based survey data was generated using Monte Carlo sampling. Each sample data set consisted of 10,000 individuals arranged into households of size $n_h = k$ for all h. Within each household, labour force flows were generated from (1), and the nonresponse flow was generated from (2), under one of models N_A or N_B. The data were made incomplete by collapsing each complete labour force flows data table, to be consistent with the household nonresponse flow. In total, 1,000 independent data sets were generated in this way.

The population parameters used to generate the labour force flows are shown in the following table:

		b		
	$\omega(ab)$	1	2	3
	1	0.43	0.245	0.035
a	2	0.02	0.160	0.01
	3	0.015	0.035	0.05

This is clearly a population in recession, since the probability of moving from being employed to unemployed is very large ($\omega(12) = 0.245$). Under models N_A and N_B, the population parameters are

	i		
	1	2	3
$\lambda(i)$	0.2	0.8	0.5
$\theta(i)$	0.5	0.2	0.8

It should be noted that these parameter values do not represent realistic nonresponse flows behaviour, they were chosen for the purpose of illustrating this methodology. However, this does not affect the general conclusions of the paper, which are also relevant for realistic values of the true nonresponse probabilities.

3.2 Simulation Results

Estimates for individual-level models are obtained by fitting (4) with $n_h = 1$ to each incomplete data set. Figure 1 summarises the sampling distributions of the individual-level maximum likelihood estimate of $\omega(12)$, $\hat{\omega}(12)$, for nonresponse models I_A, I_B, N_A and N_B (estimates for ignorable models I_A and I_B are included together, because both yield the same estimates of the labour force flows). The vertical lines represent the intervals between the 2.5-percentile and the 97.5-percentile of each estimate's sampling distribution, and the bold point represents its median. There are three distributions obtained for each individual-level estimate: the left-most distribution is that when the household size is $k = 1$, *i.e.*, the simulated data

have no household structure; and reading from left to right, the next two distributions are those obtained when the household size is $k = 2$ and $k = 5$, respectively. The solid horizontal line denotes the true flow probability, $\omega(12) = 0.0245$. The behaviour of the sampling distribution of $\hat{\omega}(12)$ in this study, reflects that of the other labour force gross flows estimates.

Figure 1a summarises the sampling distributions when N_A is the true model. If the fitted individual-level model is I_A, I_B or N_B, the labour force gross flows estimates have large biases, whatever the household size. As would be expected, the median estimate for correct model N_A, is unbiased if $k = 1$ and a small bias is apparent for $k = 2$ and $k = 5$ (although this bias is smaller for $k = 5$ than $k = 2$). Bias reduction with increasing k is also apparent for individual-level estimates I_A, I_B and N_B. This behaviour is unexpected, since it seems natural to expect the bias of the individual-level estimates, to increase with the household size. The results are slightly different in Figure 1b when N_B is true. Here the estimate for individual-level model N_B, becomes more biased as k increases, but the bias decreases for mis-specified individual-level models I_A, I_B and N_A. Furthermore, the misspecified estimates for I_A and I_B have a small bias, when compared to those for misspecified model N_A. These results are discussed in Section 3.3.

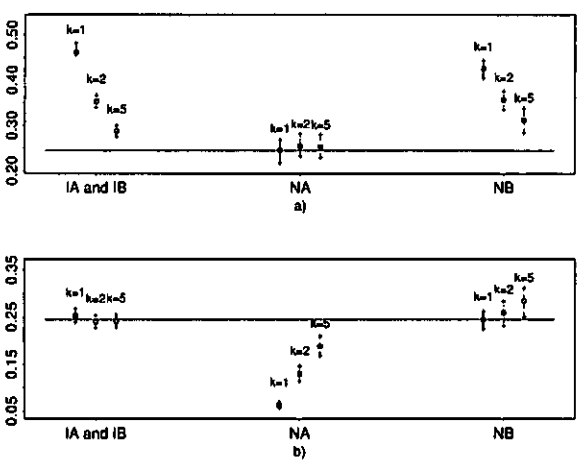

Figure 1. Sampling Distribution of $\hat{\omega}(12)$ for Individual-Level Models I_A, I_B, N_A and N_B When the True Nonresponse Model is a) N_A and b) N_B and the Household Size is $k = 1, 2, 5$.

A comparison of the median estimates of $\omega(12)$ for the fitted individual-level and household-level models when N_B is true, is presented in Figure 2. There are four sampling distributions associated with each model: the first two represent those from fitting an individual-level nonresponse model, and a household-level nonresponse model, when the household size is $k = 2$; and similarly, the next two distributions are those when the household size is 5.

For a particular pair of individual-level and household-level sampling distributions, it can be seen that the household-level estimate is less biased than its equivalent individual-level estimate, and the spread of each household-level sampling distribution, is narrower. The exception to this, is when fitting model I_A, where the household-level and individual-level distributions are identical. This equality occurs because the observed data likelihood for the individual-level and household-level models, are equivalent when the nonresponse model is ignorable. Another feature is that, if the nonresponse model is correctly specified, the household-level estimates are unbiased.

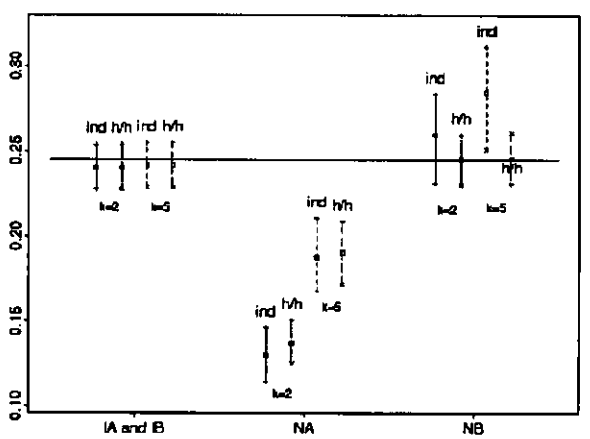

Figure 2. Sampling Distributions of $\hat{\omega}$ (12) for Individual-Level and Household-Level Models I_A, I_B, N_A and N_B When the True Nonresponse Model is N_B and the Household Size is $k = 2, 5$.

3.3 Summary

The estimates of the labour force gross flows under individual-level models, are never less biased than those of household-level models, when fitted to household-based survey data in our study. It should be noted, that if the true model is ignorable, it is unnecessary to utilise a household-level nonresponse model, because the individual-level and household-level models are equivalent. For example, if I_A is true, (2) reduces to $\lambda^{u+v}(1-\lambda)^{1-u-v}$, and (4) factorizes into two components, dependent on ω only and λ only; the factor dependent on ω can be shown to be equivalent to that for the individual-level model, and thus the labour force flows estimates are the same.

It appears, as the household size increases, that the bias of the labour force flows estimates decreases, if the true model is nonignorable. In fact, this result arises because we use (1) to generate the labour force flows, and not because the model estimates are unbiased for large n_h. To see why, consider the household formation process, used to generate

each Monte Carlo sample: as n_h increases, each household frequency tends to the same value, *i.e.*, $n_h(ab)$ converges to $n_h\omega(ab)$; hence,

$$\pi(uv \mid \boldsymbol{n}_h) \to \frac{1}{n_h}\sum_{a,b} n_h\omega(a,b)\psi\ (uv \mid ab)$$

$$= \sum_{a,b} \omega(ab)\psi(uv \mid ab),$$

which is independent of $\boldsymbol{n}_h$, that is, the simulated household nonresponse mechanism is ignorable. Therefore, the labour force flows estimates are unbiased, because fitting the nonignorable models to the simulated data, yields parameter estimates that are consistent with ignorable nonresponse. To generate nonignorable household-level nonresponse, it is necessary to prevent $n_h(ab) \to n_h\omega(ab)$, by extending (1), to allow for differential labour force flows between households. Such extensions to the labour force flows model are discussed in Section 4.

Figure 1b) shows two anomalous results that contradict the above explanation, when N_B is the true model. First, the bias of individual-level model N_B's estimate, increases as n_h increases. However, further simulations with household size $n_h = 10$, revealed that the individual-level estimate bias is zero. Thus, asymptotic ignorable nonresponse is also evident when N_B is true, but n_h must be large before its effect becomes apparent for individual-level model N_B. Second, the bias of the ignorable individual-level model estimates is small, almost zero, when N_B is true. This small bias reduces even further as n_h increases, in line with asymptotic ignorability, but we have yet to arrive at a satisfactory explanation as to why the ignorable models perform so well in this situation. Further study is necessary to investigate this finding.

4. DISCUSSION

In Sections 3 and 4, it is demonstrated by means of a simulation based study, that modelling household-level nonignorable nonresponse, when estimating labour force gross flows from household-based surveys, leads to reduced bias in the flows estimates, compared to those from individual-level models. If the nonresponse model is ignorable, it is unnecessary to use household-level models, because the individual-level and household-level models are equivalent. Furthermore, it is shown that controlling for household-level nonresponse does not necessarily remove all bias from the estimates of the labour force flows. Correct specification of the nonresponse model is still seen to be imperative, although taking the household structure of the data into account, may lead to a refinement of the flows estimates if the nonresponse model is misspecified. In particular, we show that household-level estimates are less biased than their equivalent individual-level estimates.

Our nonresponse model is an extension of the idea that nonresponse can depend upon the characteristics of a unit, in this case, the labour force flows of household members. Nonresponse in household-based surveys can occur for more than one reason, *e.g.*, refusal, non-contact, moving house or sample rotation. The current model can easily be extended to model more complex nonresponse patterns, by specifying the nonresponse indicator as a polytomous variable, and parameterizing the nonresponse model in accordance with the complex nonresponse patterns. It should also be noted, that we do not assume that the household-level model is an accurate representation of household nonresponse behaviour; rather, we assume that the household-level model, offers an approximation of within-household nonresponse dynamics.

An important problem, highlighted by the results from the simulation study, is our assumption that individual labour force flows behaviour is homogeneous within households. Clearly, this is an unrealistic assumption. The model is easily extended, by specifying the labour force flows and nonresponse flows probabilities, as regression models to accommodate individual-level, household-level, or higher level covariate information. For example, the labour force flows probabilities could be specified as a multinomial-logistic regression:

$$\log\left(\frac{\omega_{hi}(ab)}{\omega_{hi}(11)}\right) = \beta_0^{(ab)} + \boldsymbol{\beta}_1^{(ab)}\boldsymbol{x}_{hi}^T,$$

where $\omega_{hi}(ab)$ denotes the probability of individual i in household h, making labour force flow (a, b), $\boldsymbol{x}_{hi}$ is a (row) vector of covariates, and $(\beta_0^{(ab)}, \beta_1^{(ab)})$ are the regression coefficients for multinomial-logit (a, b). However, fitting these models requires conditional independence assumptions to be made, about the relationship between the distributions of the covariates, the labour force flows and the nonresponse flows, because the covariate information may be missing for nonresponding households. An alternative solution, is to allow for heterogeneous between household labour force flows, using random effects, by making assumptions about the distribution of between household differences. Fitting these models is also complicated and would require, for example, a Markov chain Monte Carlo procedure to perform the necessary integration. If S is not a simple random sample, auxiliary design variables can be incorporated into the fitting process, using the regression framework just described.

ACKNOWLEDGEMENTS

The work of Paul Clarke on this paper was funded by an Economic and Social Research Council studentship (award no. R00429614273); the work of Ray Chambers was funded by the contract between the Office for National

Statistics and the University of Southampton for the provision of research services in statistical methodology. Both authors would like to thank the referees, whose comments and practical advice helped make the final version of the manuscript considerably more readable.

REFERENCES

FITZMAURICE, G.M., LAIRD, N.M., and ZAHNER, G.E.P. (1996). Multivariate logistic models for incomplete binary responses. *Journal of the American Statistical Association*, 91, 99-107.

HOGUE, C.R. (1985). History of the problems encountered in estimating gross flows. *Proceedings of the Conference on Gross Flows in Labor Force Statistics*, U.S. Bureau of the Census, 1-8.

LITTLE, R.J.A. (1982). Models for nonresponse in sample surveys. *Journal of the American Statistical Association*, 77, 237-250.

LITTLE, R.J.A. (1985). Nonresponse adjustments in longitudinal surveys: models for categorical data. *Bulletin of the International Statistical Association*, 15, 1-15.

RUBIN, D.B. (1976). Inference and missing data. *Biometrika*, 63, 581-592.

STASNY, E.A. (1986). Estimating gross flows using panel data with nonresponse: an example from the Canadian Labour Force Survey. *Journal of the American Statistical Association*, 81, 42-47.

STASNY, E.A., and FIENBERG, S.E. (1985). Some stochastic models for estimating gross flows in the presence of nonrandom nonresponse. *Proceedings of the Conference on Gross Flows in Labor Force Statistics*, U.S. Bureau of the Census, 25-39.

VANSKI, J.E. (1985). Uses of gross change data in assessing demographic labor market dynamics. *Proceedings of the Conference on Gross Flows in Labor Force Statistics*, U.S. Bureau of the Census, 9-12.

Survey Methodology, December 1998
Vol. 24, No. 2, pp. 131-138
Statistics Canada

Longitudinal Analysis of Swiss Labour Force Survey Data by Multivariate Logistic Regression

PAUL-ANDRÉ SALAMIN[1]

ABSTRACT

In longitudinal surveys, simple estimates of change, such as differences of percentages may not always be efficient enough to detect changes of practical relevance, especially in sub-populations. The use of models, which can represent the dependence structure of the longitudinal survey, can help to solve this problem. One of the main characteristics observed by the Swiss Labour Force Survey (SLFS) is the employment status. As the survey is designed as a rotating panel, the data from the SLFS are multivariate categorical data, where a large proportion of the response profiles are missing by design. The multivariate logistic model, introduced by Glonek and McCullagh (1995) as a generalisation of logistic regression, is attractive in this context, since it allows for dependent repeated observations and incomplete response profiles. We show that, using multivariate logistic regression, we can represent the complex dependence structure of the SLFS by a small number of parameters, and obtain more efficient estimates of change.

KEY WORDS: Longitudinal binary data; Multivariate logistic model; Labour force survey.

1. INTRODUCTION

One of the main objectives of the Swiss Labour Force Survey (SLFS), is to produce estimates of change for the percentages of the population in different employment statuses. Typically, simple estimates of change, such as the difference of the percentages of employed individuals between two years, are calculated for the whole population, and for a large number of sub-populations. In general, this is unsatisfactory, as the estimates for the sub-populations may not always be efficient enough to detect changes of practical relevance. The work presented here was motivated by the question, whether the use of models, which can represent the dependence structure of the survey, could help to solve this problem.

As the SLFS is designed as a rotating panel, we are dealing with longitudinal categorical data, for which a fairly large proportion of the response profiles, are incomplete by design. The focus of interest is on modelling marginal probabilities, namely, the probabilities to be in a given employment status, as a function of time and other covariates that define sub-populations. If the repeated observations of the employment status were independent, a natural approach would be to use logistic regression. The multivariate logistic model, introduced by Glonek and McCullagh (1995) as a generalisation of logistic regression, is attractive in this context, since it allows for dependent repeated observations and incomplete response profiles.

The aim of this paper is to show that, the ability of multivariate logistic regression to model the complex dependence structure of the SLFS data, leads to more efficient estimators of change. Although we illustrate the method using the SLFS data only, it is clearly of wider applicability.

There are a number of important issues that are not dealt with in this paper. As the SLFS data come from a complex survey, it can be argued that any analysis should take the sampling weights into account (Pfeffermann 1993). Here we use the unweighted data only. However, it can be shown, using the pseudo-likelihood approach of Binder (1983), that multivariate logistic regression can be extended to that situation (Salamin 1998). Non-response is always of great concern in sample surveys. Here, we consider only the incomplete response profiles that arise through the rotation of the panel, in which case, the hypothesis of missing completely at random, is reasonable. Note however, that multivariate logistic regression, is flexible enough to incorporate extra parameters for the incomplete profiles, arising from panel, attrition. Thus, the individuals which dropped out of the panel, could also have been included into the analysis. Finally, it is well known that classification errors may introduce large biases in the observed response profile probabilities, see *e.g.*, Pfeffermann, Skinner and Keith (1998). It would certainly be desirable to investigate how these biases affect the parameter estimates of multivariate logistic regression, which have interpretations in terms of marginal moments.

Log-linear models and marginal models are closely related to multivariate logistic regression, and are further discussed in Section 3. Here we discuss briefly transition models, random effects models, and survival analysis, in the context of the SLFS. Under a transition model, see *e.g.*, Diggle, Liang and Zeger (1994, Ch. 10) or Zeger and Liang (1992), the repeated observations of the employment status are correlated, because past employment statuses influence the present employment status. The focus of interest, are the transition probabilities between the different employment statuses, *e.g.*, the probability of being

[1] Paul-André Salamin, Statistical Methods Unit, Swiss Federal Statistical Office, Espace de l'Europe 10, CH-2010 Neuchâtel, Switzerland.

employed, conditional on being unemployed in the past. In the regression setting, the past responses are treated as additional explanatory variables. An important issue, is the determination of the number of past responses to include as predictors. If the model for the transition probabilities is correctly specified, we can treat the repeated transitions for an individual as independent events, and use standard statistical methods, such as logistic regression. Under a random effect model, see *e.g.*, Diggle *et al.* (1994, Ch. 9), the probability of being in a given employment status, is a function of explanatory variables, where the regression coefficients vary from one individual to the next. This variability of the regression coefficients, reflects the natural heterogeneity of the individuals, due to unmeasured factors. Given the regression coefficients, the repeated observations of the employment status, are assumed to be independent. The correlation among the repeated observations, arises solely because we are unable to observe the true regression coefficients. This approach is most useful, when inference about individuals rather than population averages, is the focus of interest. In survival analysis, also called event history analysis in the econometric literature (Lancaster 1990), the focus is on modelling the transitions between employment statuses over time, as a function of explanatory variables. Here, the exact time at which a transition takes place, is important. In the SLFS, the employment status is observed once a year. The changes in employment status, that took place during the year preceding the interview, can be reconstructed. However, since this reconstruction is based on the self-assessment of the subjects, there may be some imprecision as regards prior status, and time of change of status. An analysis of the SLFS data based on this approach can be found in Gerfin (1996).

The article is organized as follows. We begin in Section 2 by describing the data, a subset of about 5000 individuals from the SLFS, which are used in the examples of Sections 4 and 5. In Section 3, we discuss multivariate logistic regression, and contrast it with the log-linear and marginal models. In Section 4, we illustrate the ability of multivariate logistic regression, to represent the complex dependence structure of the SLFS data, by a small number of parameters. In Section 5, we compare multivariate logistic regression with a simple estimator of change. It is shown that, using multivariate logistic regression, results in a gain in efficiency. Finally, we present in Section 6 our conclusions, and give directions for further work.

2. SWISS LABOUR FORCE SURVEY DATA

A detailed description of the sampling design and weighting procedure of the SLFS, can be found in Hulliger, Ries, Comment and Bender (1997). Here, we just recall some of the relevant aspects of this survey. The SLFS collects information on the employment of resident persons of age 15 or more in Switzerland. Starting in the second quarter of 1991, a sample of about 16,000 persons are interviewed each year. The survey is designed as a rotating panel, with a time-in-sample of 5 years. During the start-up phase, *i.e.*, from 1992 to 1996, approximately one fifth of the original sample was rotated out each year, and replaced by a renewal sample. The units in the renewal samples then stayed in the panel for a full period of 5 years.

In the examples of Sections 4 and 5, we use the observations of the employment status, for the years 1992 to 1995, obtained from the individuals in the sample, of the canton of Vaud. The structure of the data, as well as the longitudinal and cross-sectional sample sizes, are shown in Table 1. Due to the sampling design, some of the response profiles are incomplete. For example, for the individuals that were selected in 1991 and rotated out of the sample in 1994, the period of observation, denoted (1)234, goes from 1991 to 1994. We use the notation (1)234, to emphasise the fact, that we do not use the observations taken in 1991.

Table 1
Structure of the Data, Longitudinal and Cross-sectional Sample Sizes Canton of Vaud, 1992-1995

First year in sample	Observation times for various parts of the sample				Period of observation	
91	92				(1)2	622
	92	93			(1)23	412
	92	93	94		(1)234	527
	92	93	94	95	(1)2345	481
92	92	93	94	95	2345	612
93		93	94	95	345	722
94			94	95	45	728
95				95	5	877
	2,654	2,754	3,070	3,420		4,981

Employment status is a nominal variable with three categories, defined as "employed", "unemployed" and "out of the labour force". In the examples of Section 4 and 5, we work with a binary variable, taking the value 1 if an individual is employed, and 2 if an individual is unemployed or out of the labour force. This is done solely to simplify the presentation of the multivariate logistic models. As the method can handle an arbitrary number of categories, it would be preferable, not to collapse the statuses in a real analysis. Caution must be exercised, if it is nevertheless necessary to combine some of the statuses, as heterogeneity of the statuses may introduce bias.

3. MULTIVARIATE LOGISTIC MODELS

The multivariate logistic model, introduced by Glonek and McCullagh (1995), can handle multivariate responses of either nominal or ordinal types, and either discrete or continuous explanatory variables. Here, we consider only multivariate binary responses and discrete predictors. The

multivariate logistic model, is an example of a generalized linear model, see McCullagh and Nelder (1989). Its link function, also called the multivariate logistic transformation, expresses the joint distribution of the response profiles, in terms of marginal moments of increasing order, the first two being marginal logits, and marginal log odds ratios. The link function has the property, termed reproducibility, that a multivariate logistic model, applies to any subset of the response vector. This property ensures that, the interpretations of the parameters are the same, regardless of the number of response variables, and whether or not higher order parameters are included. This makes multivariate logistic regression, especially attractive for the analysis of longitudinal data, where the repeated observations of an outcome arise on an equal footing, and where the number of repeated observations may vary from one individual to the next. Reproducibility is also the key to the ability of the model, to accommodate observations with incomplete responses. Note however, that we need to assume, that the data are missing completely at random, if the same parameters are to be used to model the complete and incomplete response profiles. The parameter estimates are found by maximum likelihood. A key step, is the inversion of the multivariate logistic transformation. For more than three responses, this may not always be possible, as there are then constraints among the parameters (Glonek and McCullagh 1995, Liang, Zeger and Qaqish 1992). Also, the presence of empty cells, may limit the order of the parameters that can be fitted.

The log-linear model is widely used to model multivariate binary data. In the saturated log-linear model, see *e.g.*, Liang *et al.* (1992), the canonical parameter associated with a subset of the variables, has an interpretation in terms of conditional probabilities given the rest of the variables, *e.g.*, the first and second order parameters are logits and log odds ratios, conditional on all the other responses. It follows that, the log-linear model is not reproducible, which makes it less preferable than multivariate logistic regression, for the analysis of longitudinal data. It is nevertheless possible, to build log-linear models that, as in the multivariate logistic model, have marginal logits as parameters. This leads to the marginal models (Diggle *et al.* 1994, Ch. 8). In these models, the dependence of the marginal probabilities on explanatory variables, is modelled separately from within-unit correlation. Under this approach, the parameters are not estimated by maximum likelihood. Rather, only the structure of the correlation, between the repeated observations of an outcome is specified, and the parameters are estimated by solving generalized estimating equations (GEE), a multivariate analogue of quasi-likelihood (McCullagh and Nelder 1989). A number of specifications of the correlation structure have been proposed, for example Liang *et al.* (1992) use the marginal log odds ratios, as in Glonek and McCullagh (1995). We have made some comparisons between multivariate logistic regression and PROC GENMOD of SAS (release 6.12). This procedure has the ability to fit correlated response models by the GEE method. We found very similar estimates of the marginal logits. The GEE method appeared to be slightly less efficient than multivariate logistic regression. A limitation of the GEE method is that, it cannot yield estimates of the response profile probabilities, but only of the marginal probabilities. By contrast, the multivariate logistic model does not have this limitation, since its parameters are estimated by maximum likelihood.

Following Glonek and McCullagh (1995), we discuss in Section 3.1 the multivariate logistic transformation, and we give, in Section 3.2, the algorithm for maximum likelihood.

3.1 Multivariate Logistic Transformation

Let $Y_1, Y_2, ..., Y_d$ be d repeated observations, taken at times $t_1 < t_2 < ... < t_d$, of the same binary variable, and let

$$\pi_{i_1 i_2 ... i_d} = P(Y_1 = i_1, Y_2 = i_2, ..., Y_d = i_d),$$

where $i_1, i_2, ..., i_d$ are all either 1 or 2, be the joint probabilities of the random variables $Y_1, Y_2, ..., Y_d$. In the multivariate logistic model, the joint probabilities of $Y_1, Y_2, ..., Y_d$ are parameterized in terms of marginal logits, marginal log odds ratios, and contrasts of marginal log odds ratios. This parameterization can be written as $\eta = C^T \log(L\pi)$, where π is the vector of dimension $q = 2^d$

$$\pi = (\pi_{11...11}, \pi_{11...12}, ..., \pi_{22...21}, \pi_{22...22})^T,$$

and where, the matrices L and C are tensor products of suitably chosen marginal indicator and contrast matrices respectively. The matrices L and C, which depend on the length d of the observation period, are defined recursively, beginning with $L_0 = C_0 = 1$, as

$$L_d = \begin{bmatrix} L_{d-1} \otimes 1_2^T \\ L_{d-1} \otimes \tilde{L} \end{bmatrix}$$

and

$$C_d = \begin{bmatrix} C_{d-1} & 0 \\ 0 & C_{d-1} \otimes \tilde{C} \end{bmatrix},$$

where $1_2^T = (1, 1)$, $\tilde{L}$ is the two by two identity matrix, and $\tilde{C} = (1, -1)^T$ (Glonek and McCullagh 1994).

To illustrate matters, we consider periods of observation of length $d = 1, 2, 3, 4$. For $d = 1$, $\pi = (\pi_1, \pi_2)^T$ and $\eta = (\eta_0, \eta_1)^T = (\log \pi_+, \text{logit } Y_1)^T$, where the plus subscript indicates summation, and logit Y_1 is defined as

$$\text{logit } Y_1 = \log \frac{P(Y_1 = 1)}{P(Y_1 = 2)} = \log \frac{\pi_1}{1 - \pi_1} = \log \frac{\pi_1}{\pi_2}.$$

In that case the multivariate logistic transformation is equivalent to the usual logistic transformation. Note that, although the parameter $\eta_0 = \log \pi_+ = 0$ is strictly superfluous, it is convenient to retain it, as a means of ensuring that the mapping $\pi \to \eta$ is of full rank, and also expressing the requirement that $\pi_+ = 1$.

For $d = 2, \pi = (\pi_{11}, \pi_{12}, \pi_{21}, \pi_{22})^T$ and

$$\eta = (\eta_0, \eta_1, \eta_2, \eta_{12})^T =$$

$$(\log \pi_{++}, \text{logit } Y_1, \text{logit } Y_2, \log OR(Y_1, Y_2))^T$$

where

$$OR(Y_1, Y_2) =$$

$$\frac{P(Y_1 = 1, Y_2 = 1)\, P(Y_1 = 2, Y_2 = 2)}{P(Y_1 = 1, Y_2 = 2)\, P(Y_1 = 2, Y_2 = 1)} = \frac{\pi_{11}\pi_{22}}{\pi_{12}\pi_{21}}$$

is the odds ratio, a quantity measuring the association between the variables Y_1 and Y_2. The parameters η_1 and η_2 are the marginal logits at times t_1 and t_2, for example

$$\eta_1 = \text{logit } Y_1 = \log \frac{\pi_{1+}}{(1 - \pi_{1+})}.$$

For $d = 3, \pi = (\pi_{111}, \pi_{112}, ..., \pi_{221}, \pi_{222})^T$ and

$$\eta = (\eta_0, \eta_1, \eta_2, \eta_{12}, \eta_3, \eta_{13}, \eta_{23}, \eta_{123})^T.$$

The parameters η_1, η_2 and η_3 are the marginal logits at times t_1, t_2 and t_3. The parameters η_{12}, η_{13} and η_{23} are the log odds ratios of the corresponding two-dimensional marginal tables, for example

$$\eta_{23} = \log OR(Y_2, Y_3) = \log \frac{\pi_{+11}\,\pi_{+22}}{\pi_{+12}\,\pi_{+21}}.$$

The parameter η_{123} is a contrast of log odds ratios given by

$$\begin{aligned}\eta_{123} &= \log OR(Y_1, Y_2 \mid Y_3 = 1) - \log OR(Y_1, Y_2 \mid Y_3 = 2) \\ &= \log \frac{\pi_{111}\pi_{221}}{\pi_{121}\pi_{211}} - \log \frac{\pi_{112}\pi_{222}}{\pi_{122}\pi_{212}}.\end{aligned}$$

For $d = 4, \pi = (\pi_{1111}, \pi_{1112}, ..., \pi_{2221}, \pi_{2222})^T$ and

$$\eta = (\eta_0, \eta_1, \eta_2, \eta_{12}, \eta_3, \eta_{13}, \eta_{23}, \eta_{123},$$

$$\eta_4, \eta_{14}, \eta_{24}, \eta_{124}, \eta_{34}, \eta_{134}, \eta_{234}, \eta_{1234})^T.$$

The parameters η_i, η_{ij} and η_{ijk}, where $1 \le i < j < k \le 4$, are defined as above, using the appropriate marginal tables. The parameter η_{1234} is a contrast of log odds ratios given by

$$\begin{aligned}\eta_{1234} &= \log OR(Y_1, Y_2 \mid Y_3 = 1, Y_4 = 1) \\ &\quad - \log OR(Y_1, Y_2 \mid Y_3 = 1, Y_4 = 2) \\ &\quad - \log OR(Y_1, Y_2 \mid Y_3 = 2, Y_4 = 1) \\ &\quad + \log OR(Y_1, Y_2 \mid Y_3 = 2, Y_4 = 2).\end{aligned}$$

A key step in maximum likelihood estimation is the computation of the inverse of the multivariate logistic transformation. To ensure that $\pi > 0$, we work with $\pi = \exp v$, *i.e.*, we seek to solve for v in the equation $\eta = C^T \log(L \exp v)$. In general, no explicit solution is available, so an iterative method must be used. In particular, the Newton-Raphson iterations can be applied as described below. For clarity, we define the two functions $\varphi(\pi) = C^T \log(L\pi)$ and $\psi(v) = \varphi(\exp v)$.

(i) Begin with an initial approximation v_0.
(ii) Then take $v_{k+1} = v_k - [D\psi(v_k)]^{-1}(\varphi(\exp v_k) - \eta)$, where $D\psi(v)$ is the Jacobian matrix of the function $\psi(v)$, and iterate until convergence.

The Jacobian matrices of the function $\varphi(\pi)$ and $\psi(v)$ are given respectively by $D\varphi(\pi) = C^T (\text{diag}\, L\pi)^{-1} L$ and $D\psi(v) = D\varphi(\exp v) \cdot \text{diag}(\exp v)$.

3.2 Maximum Likelihood Estimation

For a binary response variable observed at d time points, there are $q = 2^d$ possible response profiles $i = (i_1, ..., i_d)$, where $i_1, i_2, ..., i_d$ are all either 1 or 2. For each profile $i = (i_1, ..., i_d)$, we define the indicator variable $Y_{i_1 ... i_d}$, which is equal to 1 if the profile i has been observed, and 0 otherwise. We then have

$$P(Y_{i_1 ... i_d} = 1) = P(Y_1 = i_1, ..., Y_d = i_d) = \pi_{i_1 ... i_d}.$$

Defining the q-dimensional vectors

$$Y = (Y_{11...11}, Y_{11...12}, ..., Y_{22...21}, Y_{22...22})^T$$

and

$$\pi = (\pi_{11...11}, \pi_{11...12}, ..., \pi_{22...21}, \pi_{22...22})^T,$$

we may then write $Y \sim M(1, \pi)$, *i.e.*, Y is a multinomial vector with $q = 2^d$ categories, whose probabilities are given by the vector π.

The multivariate logistic regression models, are then defined to be those of the form $\eta = X\beta$ where X is a $q \times p$ matrix of explanatory variables, β is a p-dimensional vector of unknown parameters, and $\eta = C^T \log(L\pi) = \varphi(\pi)$.

If we let y be one observation of the random vector Y, then we may write the kernel of the log likelihood function as $l(\beta; y) = y^T \log \pi(\beta)$ where, using the inverse of the

multivariate logistic transformation, we can express the joint probabilities π as a function of the unknown parameter β, as $\pi(\beta) = \varphi^{-1}(X\beta)$. The score vector is given by

$$s(\beta) = s(\beta, y, X) = D\pi(\beta)^T (\mathrm{diag}\,\pi(\beta))^{-1} y,$$

where $D\pi(\beta)$, the Jacobian matrix of the function $\pi(\beta)$, relating the parameter β to the vector of probabilities π, is given by $D\pi(\beta) = [D\varphi(\varphi^{-1}(X\beta))]^{-1} X$, and where $D\varphi(\pi) = C^T(\mathrm{diag}\,L\pi)^{-1} L$, is the Jacobian matrix of the link function. The information matrix is defined as $\mathfrak{I}(\beta) = Es(\beta)s(\beta)^T$. Now it follows from the assumption on the distribution of Y, that $E(YY^T) = \mathrm{diag}\,\pi$, from which we may deduce that

$$\mathfrak{I}(\beta) = \mathfrak{I}(\beta, X) = D\pi(\beta)^T (\mathrm{diag}\,\pi(\beta))^{-1} D\pi(\beta).$$

If we have n independent observations $y_k \sim M(1, \pi_k)$, $k = 1, \ldots, n$, where $\eta_k = C^T \log(L\pi_k) = X_k\beta$, then the score vector and the information matrix are given by $s(\beta) = \sum_{k=1}^n s(\beta, y_k, X_k)$ and $\mathfrak{I}(\beta) = \sum_{k=1}^n \mathfrak{I}(\beta, X_k)$.

The maximum likelihood estimator of β is the solution of $s(\beta) = 0$, that can be found by using the Fisher scoring algorithm which, starting from some initial value β_0, iterates the sequence $\beta_{m+1} + \beta_m + \mathfrak{I}_m^{-1}(\beta_m)s(\beta_m)$ until convergence.

Incomplete response profiles can readily be incorporated into the analysis. In particular, if some subset of the response variables $Y_1, Y_2, \ldots, Y_c$ is recorded for a particular unit, then the probability distribution on that c-dimensional marginal table is multinomial, and, as a consequence of the reproducibility of the multivariate logistic transformation, a multivariate logistic regression model applies to the table of probabilities. Furthermore, the design matrix relating the marginal probabilities to β, is constructed by selecting the appropriate rows of the full design matrix, that would be used if complete data were available for that unit.

4. MODELS FOR LONGITUDINAL DEPENDENCE

In this section we illustrate, using the SLFS data of Section 2, how multivariate logistic regression can be applied to describe the dependence between the repeated observations of the employment status. We do not intend to carry out an exhaustive search for a best model, but rather to demonstrate the ability of the method, to represent a complex dependence structure by a small number of parameters.

We consider 6 models of decreasing complexity, see Table 2. For all 6 models, we have one parameter for each of the marginal logits corresponding to a given observation time. Symbolically, this is denoted by $\eta_i = \beta_i$. Since the observation times are the 2nd quarter of the years 1992 to 1995, we take $i = 2, 3, 4, 5$. Thus β_3, say, corresponds to the logit of the probability of being employed in 1993. Similarly, the indices for the higher order parameters run from 2 to 5. For model 1 we take a saturated model for the longitudinal dependence, *i.e.*, we have one parameter for each of the interactions of order 2, 3 or 4 within each period of observation. For the models 2 to 5, we assume that the interactions of order 3 and 4, are all equal to zero. The longitudinal dependence is then described in terms of log odds ratios only. For model 2, we take a saturated model for the log odds ratios. In model 3 we drop the covariate period of observation, *i.e.*, we suppose that the log odds ratios are the same for all the periods of observation. In model 4, we use stationary log odds ratios, *i.e.*, log odds ratios which depend only on the difference between times of observation. Note that the parameter γ_1 in model 4, corresponds to the constraint $\beta_{23} = \beta_{34} = \beta_{45}$ on the parameters of model 3, and similarly for γ_2 and γ_3. In model 5, a linear model for the stationary log odds ratios is assumed. In model 6, finally, we assume that the observations taken at different times, are independent. Note that, in that case, multivariate logistic regression is equivalent to ordinary logistic regression.

Table 2
Six Models for Longitudinal Dependence

	Parameters		
Model	Marginal logits	Log odds ratios	3rd and 4th order parameters
1	$\eta_i = \beta_i$	$\eta_{ij} = \beta_{ij,\text{period}}$	$\eta_{ijk} = \beta_{ijk,\text{period}}$, $\eta_{ijkl} = \beta_{ijkl,\text{period}}$
2	$\eta_i = \beta_i$	$\eta_{ij} = \beta_{ij,\text{period}}$	$\eta_{ijk} = 0$, $\eta_{ijkl} = 0$
3	$\eta_i = \beta_i$	$\eta_{ij} = \beta_{ij}$	$\eta_{ijk} = 0$, $\eta_{ijkl} = 0$
4	$\eta_i = \beta_i$	$\eta_{ij} = \gamma_{\lvert i-j \rvert}$	$\eta_{ijk} = 0$, $\eta_{ijkl} = 0$
5	$\eta_i = \beta_i$	$\eta_{ij} = \delta + \gamma . \lvert i-j \rvert$	$\eta_{ijk} = 0$, $\eta_{ijkl} = 0$
6	$\eta_i = \beta_i$	$\eta_{ij} = 0$	$\eta_{ijk} = 0$, $\eta_{ijkl} = 0$

The parameter estimates for the models 2 to 6, are given in Table 3. The number of parameters and the values of the log likelihood function at the maximum likelihood estimates, can be found in Table 4 where, for comparison, we also included the log likelihood for the fully saturated model.

Overall, we notice that the assumed form of the longitudinal dependence, appears to have little effect on the estimates of the marginal logits. This is a desirable property, as the marginal logits would typically be the parameters of interest. The standard errors of the marginal logits, are almost the same for the models that take into account the longitudinal dependence, but are inflated by about 15% for ordinary logistic regression (model 6). It can also be shown that the estimates of the marginal logits are positively correlated under the models that assume a longitudinal dependence, and uncorrelated for ordinary logistic regression. For the example considered here, the

correlation was found to lie between 0.4 and 0.8. Thus, modelling the longitudinal dependence, leads also to more efficient estimates of the difference of marginal logits.

It can be seen from the fit of model 1, that the interaction parameters of order 3 and 4, are not significantly different from 0. This suggests that the longitudinal dependence can be described by the log odds ratios only. This hypothesis is corroborated by the incremental deviance of model 2 with respect to model 1, which is found to be 7.9, on 12 degrees of freedom. Further, all the parameters of model 2 are significantly different from 0, and an examination of the standardised residuals for the fitted probabilities of the response profiles, does not reveal any anomaly. For applications in official statistics, model 2 would be the preferred model, since it is based on as few assumptions as possible, while still allowing a substantial reduction in the number of parameters, thus rendering less acute the danger of sparse tables when longer periods of observation and models with more covariates are considered.

The models 3, 4 and 5 show that, it would nevertheless be possible to greatly simplify the description of the longitudinal dependence, without losing too much information. In going from model 2 to model 5, we observe that the deviance from the fully saturated model, does not increase much, see Table 4. Further, an examination of the residuals shows that, the models 3, 4 and 5 fit the data almost as well as model 2. On the other hand, while model 2 requires 20 parameters to describe the longitudinal dependence, model 5 needs only 2 parameters. This must be contrasted with model 6, which assumes independence between observations taken at different times: the log likelihood is much smaller than for the fully saturated model, see Table 4, and the fit to the data is poor.

Table 3
Parameter Estimates and Standard Errors

Parameter	Period	Model 2	Model 3	Model 4	Model 5	Model 6
logit 92		0.6348 (0.0350)	0.6360 (0.0352)	0.6348 (0.0352)	0.6347 (0.0352)	0.6471 (0.0409)
logit 93		0.5555 (0.0335)	0.5570 (0.0338)	0.5597 (0.0335)	0.5601 (0.0335)	0.5509 (0.0396)
logit 94		0.5440 (0.0324)	0.5407 (0.0325)	0.5402 (0.0326)	0.5397 (0.0325)	0.5377 (0.0374)
logit 95		0.4699 (0.0317)	0.4711 (0.0320)	0.4710 (0.0320)	0.4712 (0.0320)	0.4705 (0.0351)
β_{23}	(1)23	4.2563 (0.3311)	4.2579 (0.1465)			
	(1)234	4.2003 (0.2894)				
	(1)2345	4.0859 (0.2954)				
	2345	4.4830 (0.2841)				
β_{34}	(1)234	4.0894 (0.2794)	4.1111 (0.1310)			
	(1)2345	3.9611 (0.2840)				
	2345	4.0989 (0.2600)				
	345	4.2490 (0.2468)				
β_{45}	(1)2345	5.3992 (0.3854)	4.5561 (0.1389)			
	2345	3.9779 (0.2544)				
	345	4.7288 (0.2735)				
	45	4.5069 (0.2600)				
β_{24}	(1)234	3.7168 (0.2641)	3.8371 (0.1442)			
	(1)2345	4.2560 (0.3059)				
	2345	3.5330 (0.2370)				
β_{35}	(1)2345	4.4000 (0.3098)	3.7913 (0.1334)			
	2345	3.6493 (0.2396)				
	345	3.6116 (0.2192)				
β_{25}	(1)2345	4.3984 (0.3173)	3.5774 (0.1530)			
	2345	3.2209 (0.2256)				
γ_1				4.3260 (0.0928)		
γ_2				3.8519 (0.1050)		
γ_3				3.5340 (0.1495)		
δ					4.7341 (0.1266)	
γ					-0.4191 (0.0653)	

Table 4
Number of Parameters and Value of the Log Likelihood Function at the Maximum Likelihood Estimates

Model	Number of parameters of order					Log likelihood
	1	2	3	4	Total	
Full Model	20	20	10	2	52	-5342.7
1	4	20	10	2	36	-5345.4
2	4	20	0	0	24	-5349.4
3	4	6	0	0	10	-5365.2
4	4	3	0	0	7	-5368.9
5	4	2	0	0	6	-5369.5
6	4	0	0	0	4	-7815.3

5. COMPARISON WITH SIMPLE ESTIMATE OF CHANGE

In this section we concentrate on the estimation of the difference of the probabilities of being employed between any two given years. We show that, estimates based on multivariate logistic regression, are more efficient than simple estimates defined as the difference of the proportions of employed individuals.

The model considered here, is the model 2 of Section 4, with sex as an additional explanatory variable. We have, for each sex, one parameter for each of the marginal logits corresponding to a given year. The longitudinal dependence is accounted for by a saturated model for the log odds ratios. The third and fourth order parameters are set to 0. This model has therefore 8 parameters for the marginal logits, and 40 parameters for the log odds ratios: 2 sexes × 20 odds ratios within periods of observation, see Table 3. By inverting the multivariate logistic transformation, estimates of the probability of being employed, and of their differences between any two given years, can also be computed.

A simple estimator of change is given by the difference of the proportions of employed individuals between any two given years. Its variance, which takes into account the overlap of the two samples, is given by

$$\frac{1}{n+r}\pi_{1+}(1-\pi_{1+})+\frac{1}{n+c}\pi_{+1}(1-\pi_{+1})$$
$$-2\frac{n}{(n+r)(n+c)}(\pi_{11}-\pi_{1+}\pi_{+1}),$$

where n is the number of cases for which observations are available for both years, r and c are the number of cases for which observations are available for only one year, π_{11} is the probability of being employed during both years, and π_{1+} and π_{+1} are the marginal probabilities of being employed.

Tables 5 shows, for the SLFS data of Section 2, the estimates of the difference of the probability of being employed under both methods. Note that both methods yield similar estimates of change. The standard errors of the simple estimates, are on the average, 30% larger than for multivariate logistic regression. The corresponding mean relative efficiency of multivariate logistic regression, with respect to the simple estimates, is 1.7. By comparison, the mean relative efficiency of multivariate logistic regression with respect to ordinary logistic regression, is 3.2.

Table 5
Change in the Probability of Being Employed
Canton of Vaud, 1992-1995

	Comparison	Multivariate logistic regression	Simple estimate
Woman	92 *vs.* 93	0.0138 (0.0090)	0.0136 (0.0115)
	92 *vs.* 94	0.0184 (0.0102)	0.0168 (0.0134)
	92 *vs.* 95	0.0375 (0.0109)	0.0356 (0.0149)
	93 *vs.* 94	0.0047 (0.0087)	0.0031 (0.0107)
	93 *vs.* 95	0.0238 (0.0095)	0.0219 (0.0128)
	94 *vs.* 95	0.0191 (0.0076)	0.0188 (0.0100)
Men	92 *vs.* 93	0.0220 (0.0095)	0.0283 (0.0116)
	92 *vs.* 94	0.0245 (0.0102)	0.0334 (0.0133)
	92 *vs.* 95	0.0387 (0.0106)	0.0452 (0.0144)
	93 *vs.* 94	0.0024 (0.0092)	0.0052 (0.0111)
	93 *vs.* 95	0.0167 (0.0098)	0.0169 (0.0130)
	94 *vs.* 95	0.0143 (0.0080)	0.0117 (0.0102)

6. CONCLUSIONS

The analyses of the SLFS data presented here, have shown the usefulness of multivariate logistic regression. Modelling the longitudinal dependence is necessary, in order to obtain a satisfactory fit of the observed response profile probabilities. Ignoring the longitudinal dependence, we still obtain acceptable point estimates of the marginal logits, but the information on the detailed structure of the data is lost. Modelling the longitudinal dependence leads also to more efficient estimates of the marginal parameters and of change, when compared with ordinary logistic regression, and a simple estimator of change. Finally, the ability of multivariate logistic regression to represent a complex dependence structure, by a small number of parameters, has also been illustrated.

Using the results of Glonek and McCullagh (1995), it is possible to extend the examples presented here, to multivariate responses of either nominal or ordinal types, with either discrete or continuous explanatory variables. The method can also be extended, to take the sampling weights into account (Salamin 1998). For the SLFS, it was found that the sampling weights have little effect on the parameter estimates of the multivariate logistic model. The standard error of the parameter estimates, was inflated by about 15%. This moderate increase of the variability of the parameter estimates due to the sampling weights, is plausible.

Indeed, as in the SLFS, only one person per household is selected, a large cluster effect was not expected.

Apart from the sort of analyses presented here, multivariate logistic regression may also be used for modelling non-response probabilities in longitudinal studies. Such models may be useful when the sampling weights need to be adjusted for non-response. The ability of multivariate logistic regression to give a parsimonious model of the data, may also be of interest in small-area estimation. In particular, estimators for a given geographical region could be based on models for an appropriately chosen larger region.

Although we did not encounter serious problems in the examples presented here, further work may need to be done on the problem of sparse tables. A critical step, when there are a large number of empty cells, is the inversion of the multivariate logistic transformation. The approach of Lang (1996), where the inversion of the link function is avoided, by specifying the models through constraints, may be of interest in this context. Another area of investigation is the influence of the classification errors on the parameter estimates of the multivariate logistic model.

ACKNOWLEDGEMENTS

This paper benefited from discussions with colleagues at the Swiss Federal Statistical Office, among which Beat Hulliger and Philippe Eichenberger deserve special mention. The help of Ariane Bender, from the Section responsible for the Swiss Labour Force Survey, in preparing and understanding the data, was extremely valuable. The author also thanks the editor and two referees for their excellent suggestions that have led to a significant improvement of this paper.

REFERENCES

BINDER, D.A. (1983). On the variances of asymptotically normal estimators from complex surveys. *International Statistical Review*, 51, 279-292.

DIGGLE, P.J., LIANG, K.-Y., and ZEGER, S.L. (1994). *Analysis of Longitudinal Data*. Oxford: Oxford University Press.

GERFIN, M. (1996). *Entwicklung von ökonometrischen Modellen zur Analyse der Dynamik auf dem schweizerischen Arbeitsmarkt*. SLFS-News, Swiss Federal Statistical Office, Berne.

GLONEK, G.F.V., and McCULLAGH, P. (1994). Multivariate Logistic Models. Technical Report 94-31, School of Information Science and Technology, Flinders University of South Australia, Adelaide.

GLONEK, G.F.V., and McCULLAGH, P. (1995). Multivariate logistic models. *Journal of the Royal Statistical Society*, B, 57, 533-546.

HULLIGER, B., RIES, A., COMMENT, T., and BENDER, A. (1997). Weighting the Swiss Labour Force Survey. (Eds. C. Malaguerra, S. Morgenthaler and E. Ronchetti). In *Conference on Statistical Science Honoring the Bicentennial of Stefano Franscini's Birth, Monte Verità, Switerland*, Basel: Birkhäuser Verlag.

LANCASTER, T. (1990). *The Econometric Analysis of Transition Data*. Cambridge: Cambridge University Press.

LANG, J.B. (1996). Maximum likelihood methods for a generalized class of log-linear models. *Annals of Statistics*, 24, 726-752.

LIANG, K.-Y., ZEGER, S.L., and QAQISH, B. (1992). Multivariate Regression Analysis for Categorical Data. *Journal of the Royal Statistical Society*, B, 54, 3-40.

McCULLAGH, P., and NELDER, J.A. (1989). *Generalized Linear Models*, (2nd edn.). London: Chapman and Hall.

PFEFFERMANN, D. (1993). The role of sampling weights when modeling survey data. *International Statistical Review*, 61, 317-337.

PFEFFERMANN, D., SKINNER, C., and KEITH, H. (1998). The estimation of gross flows in the presence of measurement error using auxiliary variables. *Journal of the Royal Statistical Society*, A, 161, Part 1, 13-32.

SALAMIN, P.-A. (1998). Multivariate logistic regression for data from complex surveys. To appear *Proceedings: Symposium '98, Longitudinal Anlysis for Complex Surveys*, Statistics Canada, May 1998.

ZEGER, S.L., and LIANG, K.-Y. (1992). An overview of methods for the analysis of longitudinal data. *Statistics in Medicine*, 11, 1825-1839.

Survey Methodology, December 1998
Vol. 24, No. 2, pp. 139-145
Statistics Canada

Price Index Surveys as Quasi-Longitudinal Studies

ALAN H. DORFMAN[1]

ABSTRACT

To calculate price indexes, data on "the same item" (actually a collection of items narrowly defined) must be collected across time periods. The question arises whether such "quasi-longitudinal" data can be modeled in such a way as to shed light on what a price index is. Leading thinkers on price indexes have questioned the feasibility of using statistical modeling at all for characterizing price indexes. This paper suggests a simple state space model of price data, yielding a consumer price index that is given in terms of the parameters of the model.

KEY WORDS: Random walk plus noise model; State space model; Laspeyres index; Paasche index; Geometric price index.

1. INTRODUCTION

Survey sampling for calculation of a consumer price index is characterized by following a given item across time to determine its prices at a succession of times. Only it is not, typically, exactly the same item that is followed – it is not this particular can of Brand Y Tomato Soup at Outlet Z the price of which is repeatedly ascertained, for this particular can is likely to have been sold and consumed, by the time of the next visit of the survey sampler – but rather a succession of items, each fitting the same description ("Brand Y 8 oz. Can of Tomato Soup with Herring, sold at Outlet Z"), the price of which is collected at different times. In other words, it is essentially a group of items fitting a narrow description which is followed across time. For this reason consumer price index surveys may be termed "quasi-longitudinal" as opposed to longitudinal surveys, which would follow individual items across time. Nonetheless, it is reasonable to hope that, having repeated measurements across time might lead to estimation procedures which could capitalize on the time series aspect of such surveys.

In the light of that hope, this paper considers a question which has by and large been ignored by statisticians and economists, or, when not ignored, been answered in the negative: Can a consumer price index (CPI) be treated from a statistical point of view? That is, can the parameter, which characterizes the "change in the cost of living" from one period to another, and which price index surveys attempt to estimate, be defined in terms of a stochastic model?

Aldrich (1992) gives an historic interpretation of early attempts by Jevons and especially Edgeworth, to incorporate distributional assumptions into the CPI. Recent papers on stochastic modeling of the CPI, are those by Balk (1980), Clements and Izan (1981,1987), Bryan and Cechetti (1993), Kott (1984) and Selvanathan and Rao (1994). Diewert (1995) reviews and criticizes these attempts, taking an argument of Keynes (1930) as decisive grounds for rejecting the stochastic approach.

In this paper, a specific approach to modeling the price index using state space models is suggested, and a specific state space model tentatively suggested. This model is applied to scanner data to demonstrate the feasibility of an index based on it. The approach we contemplate, circumvents the Keynesian criticism in fundamental ways, and offers the prospect of the many advantages that sound statistical modeling can bring, including, possibly, simplifications of the survey sampling process.

In what follows, we first briefly review the definition of a price index, and the two (non-stochastic approaches) which have dominated consideration of choice of index (Section 2). We review the Bryan and Cecchetti (1993) example of a statistical model for the price index, and Diewert's formulation of Keynes' objection (Section 3). We then introduce an approach to modeling a consumer price index, that circumvents the Keynes-Diewert difficulties, and that leads naturally to the use of state space models (Section 4). We present results of applying a relatively simple random walk plus noise model to scanner data from the A.C. Nielsen Academic Data Base (section 5). We assess the new index in Section 6, mentioning further research that might be useful.

2. BACKGROUND

What is meant by a Consumer Price Index (CPI) is a single number indicating how the purchasing power of the consumer has changed from one period t' to another t. Its raw ingredients consist of prices for the variety of available items at (at least) the two time periods

$$p_\tau = (p_{\tau 1}, \ldots, p_{\tau N}), \tau = t', t$$

as well as quantities of the items sold

$$q_\tau = (q_{\tau 1}, \ldots, q_{\tau N}), \tau = t', t.$$

[1] Alan H. Dorfman, U.S. Bureau of Labor Statistics, Room 4915, 2 Massachusetts Ave. N.E., Washington, D.C., 20212-0001, U.S.A.; e-mail: dorfman_a@bls.gov.

(Often however in practice quantity data from the periods in question are unavailable, and one makes do with some form of surrogate.) The CPI is derived from a "formula" that uses these raw ingredients:

$$I_{t't} = f(p_{t'}, p_t, q_{t'}, q_t),$$

where $f(\cdot)$ is a function of one of many possible forms. Most such forms have a long history, and have been extensively discussed in the index literature.

As examples, we mention here the Laspeyres index

$$L_{t't} = \frac{\sum_{i=1}^{N} q_{t'i} p_{ti}}{\sum_{i=1}^{N} q_{t'i} p_{t'i}} = \sum_{i=1}^{N} f_{t'i} r_{t't},$$

with $f_{t'i} = q_{t'i} p_{t'i} / \sum_{i=1}^{N} q_{t'i} p_{t'i}$ the "relative expenditures", and $r_{t'i} = p_{ti}/p_{t'i}$ the "price relatives". The Laspeyres index uses the quantities from the earlier time period, as a fixed basis of comparison of the earlier and later prices. The Laspeyres index (or a close variant) has tended to be the index most targeted by governments, because of its simplicity and intelligibility to the layperson.

The natural counterpart to the Laspeyres is the Paasche index

$$P_{t't} = \frac{\sum_{i=1}^{N} q_{ti} p_{ti}}{\sum_{i=1}^{N} q_{ti} p_{t'i}}$$

which standardizes the prices by the later period quantities. Most indices following other formulas will tend to fall between the Paasche and Laspeyres.

For later reference in this paper, we mention an index based on the geometric mean, with fixed non-negative weights f_i, adding to 1:

$$G_{t't} = \prod_{i=1}^{N} \left(\frac{p_{ti}}{p_{t'i}} \right)^{f_i}.$$

This is sometimes referred to as the "Geomean".

Fisher (1922) discusses these and many other index formulae. He introduces what has come to be called the "Test Approach", for deciding among the variety of candidates for the formula $f(\cdot)$: this approach lays out properties ("tests"), which a reasonable index would seem to require, and then examines to what extent each index formula satisfies them.

One of the tests is the Time Reversal Test: $I_{t't} I_{tt'} = 1$. Two indices which continue to exercise their sway in the world, but fail this test are, the Carli-Sauerbach index $C_{t't} = \sum_{i=1}^{N} f_i p_{ti}/p_{t'i}$ and a geomean $\tilde{G}_{t't} = \prod_{i=1}^{N} (p_{ti}/p_{t'i})^{f_{t'i}}$ which employs first period expenditures instead of fixed weights. One readily shows that $C_{t't} C_{tt'} \geq 1$, using the Cauchy-Schwartz inequality, suggesting that this index will run too high.

If an increase in prices on item i tends to give an increase in expenditure share, then $\tilde{G}_{t't} \tilde{G}_{tt'} \leq 1$, so that under such conditions, the first-period-geomean tends to run too low. If an increase in prices on item i tends to give a decrease in expenditure share, then $\tilde{G}_{t't}$ runs too high. In general, we can expect this to be a rather erratic index.

This suggests the following maxim: price indices of the form of a geometric mean, should not have weights tied to prices at one of the periods being compared; those of the form of an arithmetic mean should not have weights independent of those prices.

By contrast with $\tilde{G}_{t't}$, the geomean $G_{t't} = \prod_{i=1}^{N} (p_{ti}/p_{t'i})^{f_i}$ which has fixed weights, is the unique index which satisfies the five axioms on price indices in Balk (1995), and the "circularity test", which says that, for $t' < t^* < t$, $I_{t't} = I_{t't^*} I_{t^*t}$. Time reversal is an immediate consequence.

Indices which pass most of the tests, tend to be ones incorporating quantity information from both periods; for example, the Fisher index

$$F_{t't} = (L_{t't} P_{t't})^{1/2}$$

and the Törnqvist index

$$T_{t't} = \prod_{i=1}^{N} \left(\frac{p_{ti}}{p_{t'i}} \right)^{f_{t'ti}},$$

with $f_{t'ti} = (f_{t'i} + f_{ti})/2$. The Fisher and Törnqvist are frequently practically indistinguishable. Further discussion of the test approach, may be found in Balk (1995), Diewert (1987), and Eichhorn and Voeller (1976).

The second approach to assessing index formulas is the "economic" approach. This defines a generic index of the form

$$I_{t't} = \frac{C(p_t, U)}{C(p_{t'}, U)},$$

where $U = U(q_1, ..., q_N)$ is a well-defined "utility function", and $C(p_t, U)$ is the minimal cost at prices p_t, of achieving the standard of living, or "utility" U. For a particular utility function U, one inquires whether a particular formula can be regarded as a good approximation to the corresponding cost of living index. Like the test approach, this tends to yield indexes incorporating quantity information from both periods. See Diewert (1987) for further elaboration.

3. THE STOCHASTIC APPROACH

Aldrich (1992) gives the early history of attempts to model price relatives or logarithms of price relatives, using a common parameter that represents the overall rate of growth in prices. A basic theme of his paper is, that the stochastic approach to price indices, while being an early example of the application of statistics to economic concerns, died a natural death. Diewert (1995) also discusses these, as well as more recent examples of the statistical modeling of price relatives. The difficulty which, following Keynes (1930), Diewert finds with such use of models is exemplified by a model of Clements and Izan (1987).

The period from t' to t is divided into equi-temporal pieces, giving relatively short intervals generically represented as being from $t-1$ to t. The logarithm of the price relatives for such a "micro-period", is given by

$$\log\left(\frac{p_{ti}}{p_{t-1,i}}\right) = \pi_t + \beta_i + \varepsilon_{ti}, \qquad (1)$$

with $\varepsilon_{ti} \sim (0, \sigma_t^2/f_i)$. In their model, the f_i's are the average expenditure share of the i-th item over the period t' to t. For the sake of identifiability, it is assumed that $\sum_{i=1}^{N} f_i\beta_i = 0$. These assumptions lead to a maximum likelihood estimator

$$\hat{\pi}_t = \sum_{i=1}^{N} f_i \log\left(\frac{p_{ti}}{p_{t'i}}\right),$$

giving an MLE of the price short period price trend as

$$\exp\left(\hat{\pi}_t\right) = \prod_{i=1}^{N}\left(\frac{p_{ti}}{p_{t'i}}\right)^{f_i};$$

that is, based on their stochastic model, one derives a geometric index, with weights f_i, akin to that for the Törnqvist.

Estimates of the β_i and of σ^2 can also be derived, as well as estimates of precision, for example, of the variance of $\hat{\pi}$. Thus, a new statistical foundation seems to be put under an old estimator.

Diewert (1995) raises several objections, none of which can be taken lightly. The chief of these is

> "... the fundamental objection of Keynes (Keynes 1930, p. 78): 'The hypothetical change in the price level [$\exp(\pi_t)$] which should have occurred if there had been no changes in relative prices, is no longer relevant if relative prices have in fact changed – for the change in relative prices has in itself affected the price level'."

If, say, the price of bread relative to the price of automobiles changes, then by that very fact, the overall price level changes.

Keynes' objection is not entirely clear. Why can't there be two aspects of price change, one overall, and the other particular? However, it is not hard to agree that the individual price trends are primary; an overall price trend can only be some weighted sum of these. Diewert offers the following translation into terms of a model, of Keynes' objection. Since we must have the overall price trend of the form

$$\pi_t^* = \sum_{i=1}^{N} f_i \beta_{ti}^*,$$

the model (1) needs to be replaced by

$$\log\left(\frac{p_{ti}}{p_{t-1,i}}\right) = \pi_t + \beta_{ti} + \varepsilon_{ti}, \qquad (2)$$

with $\beta_{ti} = \pi_t - \beta_{ti}^*$ and $\sum_{i=1}^{N} f_i \beta_{ti} = 0$. The crucial difference between this and (1) is that now the item parameters β_{ti} are indexed by time. But "then the resulting model has too many parameters to be identified." This would suffice to nullify the approach.

Diewert (1995) does not discuss the much more complicated time-series model of Bryan and Cecchetti (1993). Of preceding papers, it is probably the closest to our present paper, involving a complicated state space model and use of the Kalman Filter. Like the other papers Diewert reviews, it is subject to Keynes' objection.

4. PRICE INDEXES RECONSIDERED

4.1 Common Presuppositions

The stochastic modeling of price behavior given in the last section, whether embodied in equation (1) or (2), or some similar model, has three notable characteristics; the modeling is:

1. *Comprehensive* in the sense that it aims straight for an overall "inflation rate" encompassing all items.
2. *Atomistic*: every item is modelled individually, having its "private" parameter, its own rate of inflation [$\exp(\pi_t + \beta_i)$], apart from all other items.
3. *Time isolated*: price relatives modeling for period $t-1$ to t is disjoint from that for period $t-2$ to $t-1$ *etc.*

It is the combination of these suppositions that yields Diewert's "over-parameterized" argument. The primary thrust of Keynes' criticism is against 1: an overall inflation rate or rise/fall in the cost of living has to be a weighted mixture of several price trends. This may be granted without going so far as to embrace item 2. Item 2 is tacitly accepted in almost all (non-stochastic) constructions of price indices. However, it is not at all clear that every single item has its unique price trend. Different items (for example, Brand X ice cream at several supermarkets) are likely to have a tendency to rise and fall together (at least in the long run). There are degrees of homogeneity between

items. In any case, none of these assumptions is a necessary component of a stochastic view of price indices.

4.2 An Elementary State Space Model

We divide the time period t' to t into sub-periods t', $t'+1, \ldots, t-1, t$, and the collection of heterogeneous items into homogeneous sub-groups g, where the defining characteristic of homogeneity is a tendency to similarity of price change behavior. We make two assumptions:

1. $I_{t't}$ is a mixture of "homogeneous" indices $I_{gt't}$;
2. $I_{gt't}$ can be attained through chaining: $I_{gt't} = \prod_\tau I_{g\tau-1,\tau}$, where $\tau = t'+1, \ldots, t$.

We focus on a single group index $I_{gt't}$, dropping the subscript g for simplicity of notation. Thus, for the remainder of this paper, we focus on the "sub-index" $I_{t't} \equiv I_{gt't}$.

We proceed to develop an elementary state space model (Harvey 1990, Chapter 3) for the logarithms of the within-group price relatives. Suppose the group contains n items. For $i = 1, \ldots, n$, let $r_{ti} \equiv p_{ti}/p_{t-1,i}$ be the micro-period price relatives, and $y_{ti} \equiv \log(p_{ti}/p_{t-1,i}) = \log(p_{ti}) - \log(p_{t-1,i})$, their logs. The reason for using logs is that considerable empirical work, beginning with Edgeworth (see Diewert (1995)), suggests that the logs of price relatives will be much closer to having a normal distribution than the price relatives themselves, which can be considerably skewed. Normal distribution of errors is a standard assumption in state space models. Let $y_t \equiv (y_{t1}, \ldots, y_{tn})$ and $\mathbf{1}$ be a vector of ones of length n.

Consider the multivariate random walk plus noise (RWPN) model

$$y_t = \mathbf{1}\mu_t + \varepsilon_t,\ \varepsilon_t \sim \mathrm{MVN}\left(0, \textstyle\sum_{\varepsilon\varepsilon}\right)$$

$$\mu_t = \mu_{t-1} + \eta_t,\ \eta_t \sim N\left(0, \sigma_{\eta\eta}\right) \tag{3}$$

with $\varepsilon_\tau, \eta_\tau, \tau \in (t', t'+1, \ldots, t-1, t)$ mutually independent. The model implies that the amount that overall group prices are rising (or falling) in one micro-period, tends to hover around the amount they tended to rise (or fall) in the previous micro-period. This is a matter of common observation: if the price rise in one month tends to be high (low), then in the next month it tends to be correspondingly high (low). Since we are considering a homogeneous set of items, it makes sense that their log price relatives have a common mean. We leave for later work, the question of how to join sub-indices into an overall index.

The model (3) implies the simpler univariate RWPN model

$$\bar{y}_t = \mu_t + \bar{\varepsilon}_t,\ \bar{\varepsilon}_t \sim N\left(0, \sigma_{\bar{\varepsilon}\bar{\varepsilon}}\right)$$

$$\mu_t = \mu_{t-1} + \eta_t,\ \eta_t \sim N\left(0, \sigma_{\eta\eta}\right) \tag{4}$$

with $\bar{y}_t = n^{-1}\mathbf{1}'y_t$, $\bar{\varepsilon}_t = n^{-1}\mathbf{1}'\varepsilon_t$, and $\bar{\sigma}_{\varepsilon\varepsilon} = n^{-1}\mathbf{1}'\sum_{\varepsilon\varepsilon}\mathbf{1}$. Some information is thrown away in using (4); on the other hand, the normality assumption is even more likely to hold. For convenience, calculations in the study described in Section 5, were based on the univariate model.

The Kalman Filter (Harvey 1990, Section 3.2) can be used to give estimates $\hat{\mu}_\tau$, and $\hat{\sigma}_{\bar{\varepsilon}\bar{\varepsilon}}, \hat{\sigma}_{\eta\eta}$ of the state parameters μ_τ and the variances $\sigma_{\bar{\varepsilon}\bar{\varepsilon}}, \sigma_{\eta\eta}$ respectively.

Then we define $I_{t't} \equiv E(G_{t't} \mid S_t)$, where $G_{t't} = \prod_i (p_{ti}/p_{t'i})^{f_i}$ is a geomean dependent on fixed shares f_i, and S_t represents the totality of state parameters μ_τ through time t, and also the "hyperparameters" $\sigma_{\bar{\varepsilon}\bar{\varepsilon}}, \sigma_{\eta\eta}$. In other words, we condition on what we take to be the underlying process through time t. Then

$$I_{t't} = \exp\left(\mu_t + \mu_{t-1} + \cdots \mu_{t'+1} + \frac{1}{2}\nu\right), \tag{5}$$

where $\nu = (t-t')\sum_i\sum_{i'}\sigma_{ii'}f_i f_{i'}$, with $\sigma_{ii'}$ the covariance of ε_{ti} and $\varepsilon_{ti'}$ typically of lower order than the state parameters $\mu_{\tau i}$. The natural estimator of $I_{t't}$ is $\hat{I}_{t't} \equiv \exp(\hat{\mu}_t + \hat{\mu}_{t-1} + \ldots \hat{\mu}_{t'+1})$; then

$$E\left(\hat{I}_{t't} \mid S_t\right) = \exp\left(\mu_t + \mu_{t-1} + \cdots \mu_{t'+1} + \frac{1}{2}\tilde{\nu}\right), \tag{6}$$

where $\tilde{\nu}$, given in the Appendix, does not in general equal ν, but is frequently close, and in any case is of the same order of magnitude. The difference $\Delta(\nu) = \tilde{\nu} - \nu$ can be estimated, by say $\hat{\Delta}(\nu)$, yielding a bias-corrected estimator $\check{I}_{t't} \equiv \hat{I}_{t't}\exp(-1/2\hat{\Delta}(\nu))$. Expressions for ν and $\tilde{\nu}$, and a suggestion for a maximal $\hat{\Delta}(\nu)$, are given in the Appendix. It may be noted that $\hat{\Delta}(\nu)$, and hence $\check{I}_{t't}$, depends on the weights f_i, but that $\hat{I}_{t't}$ does not.

5. EMPIRICAL STUDY

To determine the feasibility of the calculation of price indices using the RWPN model and gain some idea of the behavior of the RWPN index, a small empirical study was made, using price and quantity data for Canned Tuna in the A.C. Nielsen Academic Data Base. Canned tuna has somewhat volatile price and quantity behavior, due to frequent sales, at sometimes very marked discount.

The study covered the Northeast USA and the 104 weeks of the years 1992-1993. The original data set was rather large. To make the investigation manageable, weekly data was combined into 4-week periods, giving a total of 26 periods over two years. Thus for purposes of this study, the data were cumulative quantities and quantity-weighted average prices over four week periods.

The homogeneous groups were defined by brand and type, as follows: 3 brands here labeled A, B, C of "premium" tuna in water, the same three brands of "light" tuna in oil, and again the same three brands of "light" in water, making 9 groups in all.

The study focused on 83 outlets which had positive quantities over most of the 4 week periods, for each of the 9 distinct groups.

The RWPN based index $\hat{I}_{t't}$ and the adjusted RWPN based index $\check{I}_{t't}$ were calculated for four time intervals. In each case, the final period $t = 26$, and the early period was taken successively as $t' = 3, 6, 10, 14$. For the purpose of comparison, we also calculated the corresponding Laspeyres and Paasche Indices. These two standard indices provide also a basis of indirect comparison to the Fisher and Törnqvist, which will be about half way between them.

Figures 1 and 2, for premium and light tuna respectively, give the values of the four indices for the four time intervals, the points representing the state space indices, the lines used to indicate the Laspeyres and Paasche. The adjusted RWPN $\check{I}_{t't}$ is invariably larger than the unadjusted RWPN $\hat{I}_{t't}$. Note that, since it is the first period that we are varying, where the path of indices is monotone up, this would suggest a downward trend in the cost of the particular tuna group (and vice versa).

We observe that the new indices are not out of line with the traditional indices, frequently lying between the Laspeyres and Paasche, but they tend to be considerably more stable as t' changes, suggesting possibly that the traditional indices are reacting to "noise" in the data, and that, in fact, basically very little change is going on in this two year period. It may also be observed in Figure 2, that Light in Oil and Light in Water have similar within brand behavior, suggesting that we might have taken a broader "homogeneous" grouping.

6. FURTHER WORK

The investigation described in this paper suggests several topics for further research.

Measures of precision and estimates of the RWPN indices, in terms of variances or confidence intervals based on the state space model, need to be worked out. Even those who are dubious about the viability of a stochastic methodology in price indices, find the possibility of having a measure of precision appealing (Diewert 1995). It would also be of interest to get measures of precision of more standard indices, based on the state space model.

Empirical work is desirable that investigates more closely what groups of items might best be considered "homogeneous". Also, models possibly more elaborate than the simple RWPN model require investigation. In this respect, the use of scanner data will be a great help, supplying as it does, quantity data as well as prices, in great detail.

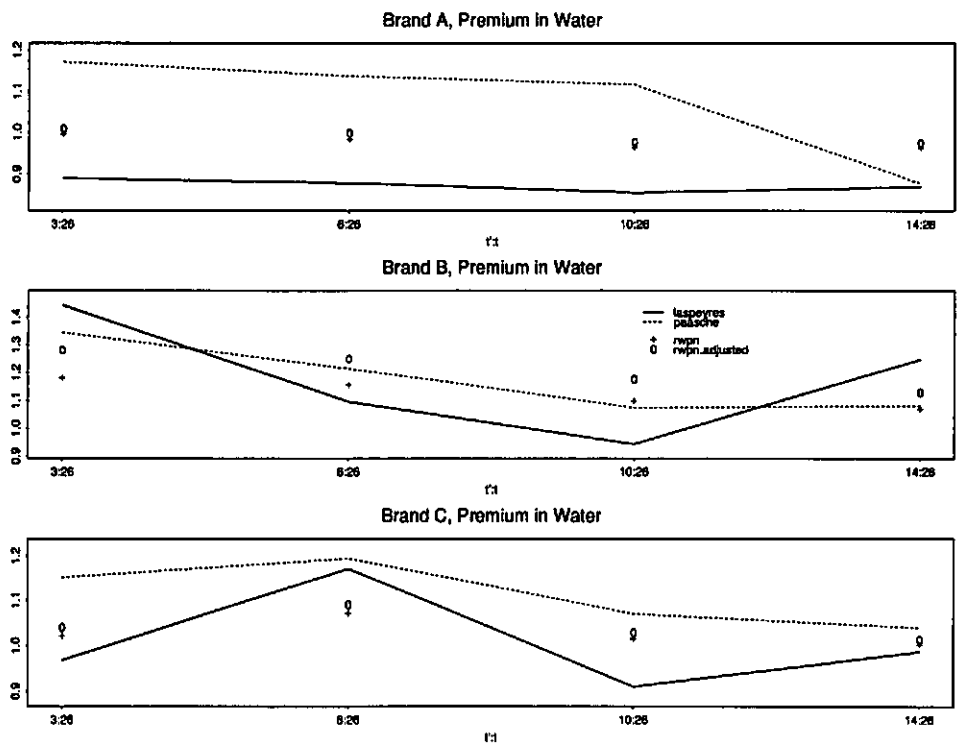

Figure 1. Four Price Indexes for Four Time Intervals, Premium Tuna

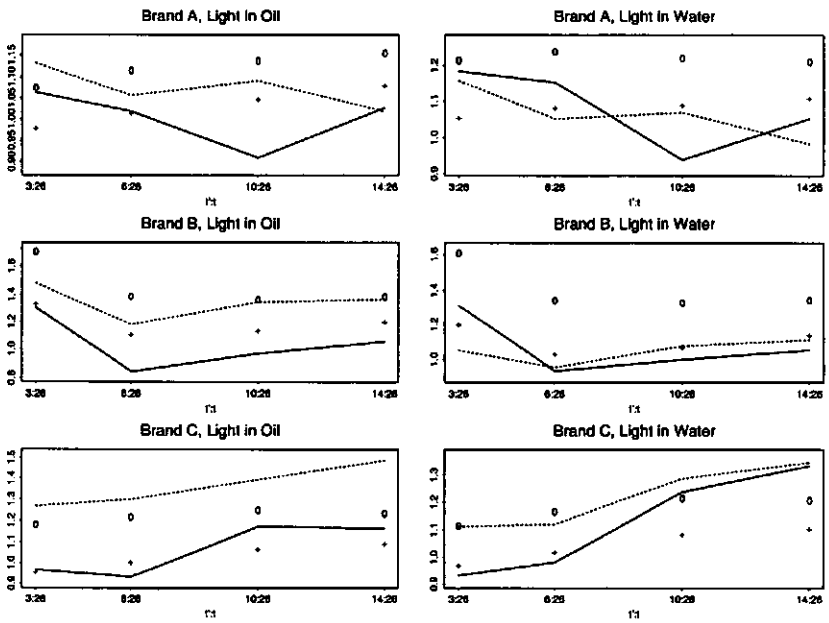

Figure 2. Four Price Indexes for Four Time Intervals, Light Tuna

The state space methodology has methods of handling missing data (Harvey 1990, Section 3.4.7). A point of major concern is how well these models will handle missing data in estimating price indices. In particular, since in practice most data for calculating price indices is based on a small sample of items available, we need to know the robustness of state space indices to the absence of data.

Algorithms for smoothing and forecasting of state space models, are well known. Their use in revising and forecasting indices, might be of great interest.

Finally, in this paper we have focussed only on getting an index for a single homogeneous group. It would be of interest to develop a state space model that combines groups and enables us to get an overall measure of purchasing power.

ACKNOWLEDGEMENTS

The author thanks B. Moulton, S. Scott, M. Reinsdorf, R. Tiller, B. Balk, and J. Aldrich for discussions of the ideas in this paper.

APPENDIX

Details of expressions (5) and (6).

We have that

$$G_{t't} = \prod_i \left(\frac{p_{ti}}{p_{t-1,i}} \frac{p_{t-1,i}}{p_{t-2,i}} \cdots \frac{p_{t'+1,i}}{p_{t'i}} \right)^{f_i}$$

$$= \prod_i \left(r_{ti} r_{t-1,i} \cdots r_{t'+1,i} \right)^{f_i},$$

and letting

$$H_{t't} = \log(G_{t't}) = \sum_i f_i \log(p_{ti}/p_{t'i}),$$

we have that

$$H_{t't} = \sum_i f_i \log\left(r_{ti} r_{t-1,i} \cdots r_{t'+1,i}\right)$$

$$= \sum_i f_i \left(y_{ti} + y_{t-1,i} + \cdots y_{t'+1,i}\right)$$

and also that

$$I_{t't} = E(G_{t't}) = \exp(E(H_{t't}) + 1/2 \operatorname{var}(H_{t't})),$$

where the moments are calculated conditional on the state S_t, as in Section 4.3. Let $\nu = \mathrm{var}(H_{t't})$. Then

$$E(H_{t't}) \equiv E(H_{t't} \mid S_t) = \sum_i f_i(\mu_t + \mu_{t-1} + \cdots \mu_{t'+1}) = \mu_t + \mu_{t-1} + \cdots \mu_{t'+1}$$

and

$$\nu = \mathrm{var}(H_{t't}) \equiv \mathrm{var}(H_{t't} \mid S_t) = \mathrm{var}\left(\sum_{\tau=t'+1}^{t} \sum_i f_i \varepsilon_{\tau i} \mid S_t \right) = (t - t') \sum_i \sum_{i'} \sigma_{ii'} f_i f_{i'},$$

where $\sigma_{ii'}$ is the covariance of ε_{ti} and $\varepsilon_{ti'}$. We note that $\nu = (t - t') \sum_i f_i^2 \sigma_{\varepsilon\varepsilon}$, in the special case that the errors ε_{ti} are independent and identically distributed at each time period.

We now consider estimator $\hat{I}_{t't} \equiv \exp(\hat{\mu}_t + \hat{\mu}_{t-1} + \cdots \hat{\mu}_{t'+1})$. We find that $E(\hat{I}_{t't}) = \exp(\mu_t + \mu_{t-1} + \cdots \mu_{t'+1} + 1/2\, \tilde{\nu})$, where

$$\tilde{\nu} \equiv \mathrm{var}\left(\sum_{t'+1}^{t} \hat{\mu}_\tau \mid S_t \right) = \left\{ \sum_{t'+1}^{t} \gamma_\tau^2 \right\} \mathrm{var}(\bar{y}_t \mid S_t) +$$

$$\gamma_{t'}^{*2}\, \mathrm{var}(\hat{\mu}_{t'} \mid S_t) = \left\{ \sum_{t'+1}^{t} \gamma_\tau^2 + \gamma_{t'}^{*2} p_{t'-1} \right\} \sigma_{\bar{\varepsilon}\bar{\varepsilon}}$$

with

$$\gamma_\tau = k_\tau \left(1 + \sum_{v=\tau+1}^{t} \prod_{u=\tau+1}^{v} (1 - k_u)\right)$$

and

$$\gamma_{t'}^{*} = \sum_{v=t'+1}^{t} \prod_{u=t'+1}^{v} (1 - k_u),$$

where

$$k_\tau = p_{\tau|\tau-1} \big/ (p_{\tau|\tau-1} + 1),$$

and $p_{\tau|\tau-1}$, p_τ are the mean square errors of $\hat{\mu}_\tau$ given data up to $\tau - 1$, τ respectively, and are estimated using the Kalman Filter.

This result follows from the equations used in estimating μ_τ:

$$\hat{\mu}_t = k_t \bar{y}_t + (1 - k_t)\hat{\mu}_{t-1}$$

$$\hat{\mu}_{t-1} = k_{t-1} \bar{y}_{t-1} + (1 - k_{t-1})\hat{\mu}_{t-2}$$

$$\vdots$$

$$\hat{\mu}_{t'+1} = k_{t'+1} \bar{y}_{t'+1} + (1 - k_{t'+1})\hat{\mu}_{t'}$$

(*cf.* Harvey 1990, equation 3.2.8), by expressing each $\hat{\mu}_\tau$ in terms of $\bar{y}_\tau, \bar{y}_{\tau-1}, \ldots, \bar{y}_{t'+1}, \hat{\mu}_{t'}$.

In comparing ν and $\tilde{\nu}$, we find, empirically that

$$\sum_{t'+1}^{t} \gamma_\tau^2 + \gamma_{t'}^{*2} p_{t'-1} \approx t - t'.$$

We here consider the simple case where $\mathrm{var}(\varepsilon_{ti}) = \sigma_{\varepsilon\varepsilon}$ and $\mathrm{cov}(\varepsilon_{ti}, \varepsilon_{ti'}) = \rho\sigma_{\varepsilon\varepsilon}$, with $\rho \geq 0$, for $i' \neq i$, that is where not only variances, but all covariances are equal and non-negative. It then can be shown that

$$\sigma_{\bar{\varepsilon}\bar{\varepsilon}} = n^{-2} \sum_i \sum_{i'} \sigma_{ii'} \leq \sum_i \sum_{i'} \sigma_{ii'} f_i f_{i'} \leq n \sum_i f_i^2 \sigma_{\bar{\varepsilon}\bar{\varepsilon}},$$

where n is the number of items in the group. The lower bound is achieved in the case $f_i = 1/n$, and the upper in the case $\rho = 0$. In the first case, no bias adjustment is necessary; in the second, we would take $\hat{\Delta}(\nu) = \hat{\tilde{\nu}} - \hat{\nu}$, where $\hat{\nu} = (t - t') n \sum_i f_i^2 \hat{\sigma}_{\bar{\varepsilon}\bar{\varepsilon}}$ and $\hat{\tilde{\nu}} = \{ \sum_{t'+1}^{t} \gamma_\tau^2 + \gamma_{t'}^{*2} p_{t'-1} \} \hat{\sigma}_{\bar{\varepsilon}\bar{\varepsilon}}$. These correspond respectively to $\hat{I}_{t't}$ and $\tilde{I}_{t't}$.

REFERENCES

ALDRICH, J. (1992). Probability and depreciation: a history of the stochastic approach to index numbers. *History of Political Economy*, 24, 657-87.

BALK, B.M. (1980). A method for constructing price indexes for seasonal commodities. *Journal of the Royal Statistical Society, A*, 143, 68-75.

BALK, B.M. (1995). Axiomatic price theory: a survey. *International Statistical Review*, 63, 69-93.

BRYAN, M.F., and CECCHETTI, S.G. (1993). The consumer price index as a measure of inflation. *Economic Review, Federal Reserve Bank of Cleveland*, 29, 15-24.

CLEMENTS, K.W., and IZAN, H.Y. (1981). A note on estimating Divisia index numbers. *International Economical Review*, 22, 745-747.

CLEMENTS, K.W., and IZAN, H.Y. (1987). The measurement of inflation: a stochastic approach. *Journal of Business and Economic Statistics*, 5, 339-350.

DIEWERT, W.E. (1987). Index numbers. In *The New Palgrave: A Dictionary of Economics*, (Eds. J. Eatwell, M. Milgate, and P. Newman). London: MacMillan.

DIEWERT, W. E. (1995). On the Stochastic Approach to Index Numbers. Discussion Paper No. DP 95-31, Department of Economics, University of British Columbia.

EICHHORN, W., and VOELLER, J. (1976). *Theory of the Price Index*. Berlin: Springer-Verlag.

FISHER, I. (1922). *The Making of Index Numbers*. Boston: Houghton Mifflin.

HARVEY, A.C. (1990). *Forecasting, Structural Time Series Models and the Kalman Filter*. Cambridge: Cambridge University Press.

KEYNES, J.M. (1930). *A Treatise on Money*. New York: Harcourt, Brace and Company.

KOTT, P.S. (1984). A superpopulation approach to the design of price index estimators with small sampling biases. *Journal of Business and Economic Statistics*, 2, 83-90.

SELVANATHAN, E.A., and RAO, D.S.P. (1994). *Index Numbers: A Stochastic Approach*. Ann Arbor: The University of Michigan Press.

Survey Methodology, December 1998
Vol. 24, No. 2, pp. 147-156
Statistics Canada

Treatment of Nonresponse in Cycle Two of the National Population Health Survey

JEAN-LOUIS TAMBAY, IOANA ŞCHIOPU-KRATINA, JACQUELINE MAYDA, DIANA STUKEL and SYLVAIN NADON[1]

ABSTRACT

The National Population Health Survey (NPHS) is one of Statistics Canada's three major longitudinal household surveys providing an extensive coverage of the Canadian population. A panel of approximately 17,000 people are being followed up every two years for up to twenty years. The survey data are used for longitudinal analyses, although an important objective is the production of cross-sectional estimates. Each cycle panel respondents provide detailed health information (H) while, to augment the cross-sectional sample, general socio-demographic and health information (G) are collected from all members of their households. This particular collection strategy presents several observable response patterns for Panel Members after two cycles: GH-GH, GH-G*, GH-**, G*-GH, G*-G* and G*-**, where "*" denotes a missing portion of data. The article presents the methodology developed to deal with these types of longitudinal nonresponse as well as with nonresponse from a cross-sectional perspective. The use of weight adjustments for nonresponse and the creation of adjustment cells for weighting using a CHAID algorithm are discussed.

KEY WORDS: Longitudinal surveys; Treatment of nonresponse; CHAID algorithms.

1. INTRODUCTION

In 1996-97, Statistics Canada completed data collection for Cycle 2 of the National Population Health Survey (NPHS). This longitudinal survey was launched in 1994 to provide comprehensive information on the health status of the Canadian population and on the determinants of health. The in-scope population covers residents of households and health institutions throughout the country. In the provinces the household questionnaire has two main components which are administered using computer-assisted interviewing. The General component collects socio-demographic and basic health information for each member of the household. The Health component obtains more detailed health information about the household member selected to participate in the longitudinal panel.

Although the NPHS is a longitudinal survey, its objectives also include the production of periodic cross-sectional estimates (Catlin and Will 1992). The data collection methodology reflects both longitudinal and cross-sectional needs. Panel Members, chosen in Cycle 1, are followed-up every two years for up to 20 years. Persons residing with the Panel Members at those times provide General component information for use in cross-sectional estimation. As the cross-sectional coverage of the sample deteriorates over time, the sample needs to be "topped-up" periodically. The first top-up is planned for Cycle 3, in 1998.

This paper presents the methodology developed in Cycle 2 to deal with nonresponse at the household and person levels (flagging will be used for item nonresponse). The methodology is based on reweighting respondents within sub-populations called weighting cells to account for nonresponse. Reweighting is a common approach for the treatment of item nonresponse. The bias and variance of this approach have been considered by Thomsen (1973), Oh and Scheuren (1983), Kalton and Kasprzyk (1986) and Little (1986), among others. If weighting cells are defined such that nonresponse occurs almost completely at random within each cell then the bias due to nonresponse can become negligible. In a similar vein David, Little, Samuhel and Triest (1983) extended to nonresponse the theory developed by Rosenbaum and Rubin (1983) in the context of propensity score matching in observational studies. Their results imply that reweighting can adjust for nonresponse bias when the weighting cells are formed based on the propensity to respond.

An overview of the NPHS sample design and outputs for the first two Cycles is given in Section 2. Section 3 presents the nonresponse treatment strategies and their results. Concluding remarks are given in Section 4. Note that the methodology presented pertains to the provincial household samples; it does not cover the samples in the territories and in institutions.

2. OVERVIEW OF THE NPHS DESIGN AND OUTPUTS

2.1 Cycle 1 Sample Design

The initial sample of households was selected in 1994 using the sample selection vehicle built for the Canadian Labour Force Survey (LFS), and, in the province of Quebec, using dwellings that had participated in a health survey

[1] Jean-Louis Tambay, Ioana Şchiopu-Kratina, Jacqueline Mayda, Diana Stukel and Sylvain Nadon, Household Survey Methods Division, Statistics Canada, 16th floor, R.H. Coats Building, Tunney's Pasture, Ottawa, Ontario, Canada, K1A 0T6.

conducted by Santé Québec the previous year. In both cases the households or dwellings were selected at random within stratified samples of clusters selected using probability proportional to size. The clusters were organized into replicates and collection period to capture seasonality and for variance estimation purposes. There were two "summer" collection periods (June and August) and two "winter" collection periods (November and March, 1995).

Figure 1 illustrates the Panel selection mechanism applied outside the province of Quebec. Sample households were randomly designated as "Adult" or "Children" households, and as eligible for screening or not, prior to collection. Screening increased the presence in the panel of inhabitants of larger households who would be under-represented with the selection of only one member per household, particularly children and youths. Households eligible for screening were rejected from the sample if they had no member aged under 25. Screening was not used in Quebec as information from the provincial health survey allowed the application of different sub-sampling rates by household type and size.

Sample Unit Type	Household Characteristic	Panel selection restricted to:
"Children" household Eligible for Screening (EFS)	No member under 25	N/A – hhld rejected
	No children, some members under 25	Any member
	Children present	Child members
"Children" household not EFS	No children present	Any member
	Children present	Child members
"Adult" hhld	All	Members over 12

Figure 1. Panel Selection Mechanism Outside Quebec

The classification into "Adult" and "Children" households was done for an operational reason: the Health questionnaire for children, would not be available before the winter collection periods. In "Adult" households, which could be interviewed any time, children under 12 were not eligible for the panel. "Children" households, even those in "summer" clusters, were interviewed in a winter collection period. If children were present in those households then the panel selection was restricted to them. To diminish the seasonal distortions to the data collection workload and the panel representability brought about by these procedures, fewer households were classified as "Children" households in summer clusters, and, with one minor exception, screening was applied only to "Children" households.

Provinces wishing to improve sub-provincial estimates could fund additional sample sizes. In three provinces this was done by augmenting the sample size in targeted regions. In British Columbia an additional sample of about 800 households was selected in a local health unit using Random Digit Dialling (RDD). The expected total sample size in the provinces was approximately 23,000 households after screening.

The above gives a general indication of the 1994 sample design which is sufficient for the needs of this paper. Readers wishing a more precise presentation of the 1994 sample should see Tambay and Catlin (1995), or Statistics Canada (1995).

2.2 Cycle 1 Weighting and Outputs

The major output of the NPHS consists of person-level anonymized Public Use Microdata Files (PUMFs) of survey responses (internal versions of those files that include information suppressed for confidentiality reasons are also created). For 1994 a General PUMF (58,400 records) and a Health PUMF (17,600 records) were released containing the General and Health information collected from every household member and from the selected non-child Panel Members, respectively (Statistics Canada 1995).

The sample weights attached to every record on the PUMF were calculated by applying a series of adjustments to a basic weight representing the household inverse sampling rates (ISR). The ISRs are calculated by multiplying the weights of the original LFS or Santé Québec samples by the inverse of the sub-sampling rates applied by the NPHS. For the sake of brevity we will only describe the main adjustments used outside of Quebec.

Adjustments to the weights for the General PUMF include: (1) a household nonresponse adjustment; (2) an adjustment for the rejective method; (3) an adjustment for person nonresponse [within responding households] and, finally; (4) a simple post-stratification adjustment. Adjustment (2) was applied only to households with no member under 25. It was $1/(1 - r_s)$, where r_s was the sub-sampling rate for the screening applied in the stratum. The post-stratification adjustment was done separately for each province-age group-sex cross-class. Weights resulting from all earlier steps are multiplied by the ratio of known to estimated population sizes within the cross-class. The known population sizes are in fact Census-based projections.

The adjustments for household and person level nonresponse (at 11.3% and 1.4%, respectively) were applied to respondent units as the nonrespondents were excluded from the PUMFs. If w_i is the sample weight of a unit i, the nonresponse-adjusted weight, $w_{\text{adj},i}$ is defined as $w_{\text{adj},i} = w_i\,(\sum_{\text{all}} w_i)/(\sum_{\text{resp}} w_i)$, where the sums are taken over all sample units and all respondent units, respectively, within nonresponse adjustment weighting cells. Due to a lack of information on nonrespondent households the weighting cells for household level nonresponse were simply cross-classes of NPHS strata and season (*i.e.*, "summer" *vs.* "winter" clusters). Weighting cells for the person level nonresponse, which was very low, were the province-age-sex cross-classes that were used for the post-stratification adjustment.

Adjustments to the weights for the Health PUMF included: (1) a household nonresponse adjustment; (2) an adjustment for the rejective method; (3) an adjustment for the "Adult/Children" household sub-sampling; (4) an adjustment for the longitudinal Panel Member selection; (5) an adjustment for Panel Member nonresponse; and (6) a post-stratification adjustment. The first two adjustments were exactly those for the General PUMF. As the Health PUMF did not include Panel Members who were children, adjustment (3) compensated for those sample households where non-children were ineligible for panel membership. The adjustment thus applied only to households with children and was equal to $1/r$, where r was the proportion of "Adult" households in the sample. Adjustment (4) was the inverse of the probability of having selected the Panel Member. The adjustments for Panel Member nonresponse (at 3.9%) and for post-stratification were similar to those for the General PUMF, and used the same province-age-sex cross-classes. Although child Panel Members were not included in the Health PUMF, for longitudinal purposes their sample weights were obtained as above using $1/(1-r)$ instead of $1/r$ in step (3).

2.3 Cycle 2 Sample Design

In Cycle 2 the focus of the survey was more on longitudinal estimation: no sample "top-up" was planned until the following cycle. The "Core" sample thus consisted of about 17,000 Panel Members and their current households. Panel Members were traced and administered the General and Health questionnaire components, while other members of their household were administered the General component only. No follow-up was done for 1994 nonresponding households. In Alberta, Manitoba and Ontario large (non-Core) additional samples were obtained, using RDD, to allow the production of cross-sectional estimates at sub-provincial levels. In every RDD household one member aged over 12 was selected to complete the Health component. In Alberta and Manitoba, RDD households with children also had a child selected to complete the Health component.

We note that, for cross-sectional purposes, the Core sample does not cover very well arrivals in the population such as newborns and recent immigrants. The population administered the General questionnaire consists of residents of households where at least one member was in-scope in Cycle 1; households made up entirely of recent immigrants (and their newborns) are thus missed. The population administered the Health questionnaire consists of persons who were in-scope in Cycle 1: recent immigrants and children under 2 years old are excluded from the Core target population (they are included in the RDD target population). For both the General and the Health questionnaires post-stratification is done using population figures that do not exclude the recent immigrants. The result is that the population of recent immigrants is implicitly being estimated for by the population of non-immigrants because the latter's Core weights are adjusted upwards to account for the former's numbers. This is a limitation that is acknowledged in the PUMF documentation. Alternative methods would have been to post-stratify using only non-immigrant population projections or to somehow adjust only the weights of less recent immigrants (who are covered) to account for the more recent immigrants (who are not). These methods would have been difficult to apply where, for the General questionnaire, a distinction between immigrants in immigrant-only households and immigrants in mixed households would have been required.

2.4 Cycle 2 Weighting and Outputs

Figure 2 summarizes the survey's three major outputs planned for Cycle 2: a Longitudinal PUMF; a Health Cross-Sectional PUMF and a General Cross-Sectional PUMF. The planned Longitudinal PUMF contains General and Health information for both Cycles for the 17,000 Panel respondents [note: confidentiality requirements may prevent the release of a longitudinal PUMF – in which case only an internal microdata file will be produced]. The Health Cross-Sectional PUMF contains 1996 General and Health information for about 70,000 Panel Members and RDD Selected Members. The General Cross-Sectional PUMF contains 1996 General information for about 220,000 members of the Core and RDD samples. The weighting processes involved for each PUMF, presented below for the Core sample, are described in more detail in Stukel, Mohl and Tambay (1997).

Output File	LONGITUDINAL PUMF	HEALTH CROSS-SECTIONAL PUMF	GENERAL CROSS-SECTIONAL PUMF
Contents	General & Health	General & Health	General only
Samples	Core only	Core & RDD (3 provs.)	Core & RDD (3 provs.)
Units	Panel Member (PM)	PM/RDD Sel. Mem.	All Hhld. Members
Size	≈ 17,000 records	≈ 70,000 records	≈ 220,000 records
Weighting Strategy (for Core Sample)	1.Base Year Weight 2.PM Nonresp. Adjustment 3.Post-stratification	1.Base Year Weight 2.PM Nonresp. Adj. 3.Core/RDD integration 4.Post-stratification	1.Base Year Weight 2.Hhld. Nonresp. Adj. 3.Weight Share Adj. 4.Hhld. Mem. NR Adj. 5.Core/RDD integration 6.Post-stratification

Figure 2. Description of Output Files for Cycle 2

Respondent survey weights on the Longitudinal PUMF are obtained by adjusting a base year weight first for 1996 panel nonresponse and then for post-stratification. The base year weight represents the inverse sampling rate for 1994 including all Health PUMF adjustments described in section 2.2 up to adjustment (4) for panel selection (a correction is needed for the "removal" of the 1994 provincial sample additions). The weight adjustment for

nonresponse is the focus of the following section and will be described there. Post-stratification is done to reproduce 1994 provincial population counts by age-sex categories.

For the Health Cross-Sectional PUMF, the weighting process for (Core) Panel Members involves three or four steps. Usually, the base year weight is adjusted for Panel Member nonresponse, as explained in the following section, and for post-stratification (to match 1996 provincial or regional population counts by age-sex categories). In provinces with RDD samples the extra step is the integration with the RDD sample. The integrated estimate is obtained by a somewhat degenerate adaptation of the Skinner-Rao dual frame estimator (Skinner and Rao 1996).

For the General Cross-Sectional PUMF, the weighting process for the core sample involves five or six steps. First, once more, is the calculation of the base year weight. Then comes an adjustment for nonresponse at the household level, discussed in the following section. The next step is the application of the "weight share method". The method was described by Ernst (1989) and developed further by Lavallée (1995). The Panel Member's weight, divided by the number of persons in his/her household who were in-scope in Cycle 1, is assigned to all household members including those who were not in-scope in Cycle 1 (*e.g.*, births, immigrants). The method is unbiased for estimates of totals for the population of households where at least one member was in-scope in Cycle 1. The next step is a household member nonresponse adjustment. In RDD provinces this is followed by integration of the Core sample with the RDD sample (this time for all ages). Post-stratification is done in a similar fashion to that for the Health Cross-Sectional PUMF.

3. CYCLE 2 CORE SAMPLE NONRESPONSE STRATEGY

This section presents the strategy adopted for the treatment of Cycle 2 nonresponse in the Core (non-RDD) sample. Adjusting for nonresponse was done once again using the weighting cell approach except that, this time, Cycle 1 data were available to create weighting cells that are more homogeneous with respect to the propensity to respond, and thus more apt to remove nonresponse bias. Section 3.1 identifies nonrespondents in the NPHS. Section 3.2 discusses two general approaches for the creation of weighting cells, giving the one chosen for the NPHS. The strategy for the adjustment for nonresponse is explained in section 3.3 while section 3.4 describes the creation of the nonresponse weighting cells.

3.1 Definitions of Nonrespondent and Out-of-scope Units

The first step in the treatment of nonresponse consisted of its definition or identification. In Cycle 2, questionnaires were fully completed for 89% of the Core sample and partially completed for another 3%. The rest of the sample consisted of refusals (3.1%), of cases where the Panel Member could not be traced (1.7%), had died (1.7%), had left Canada (0.5%), or was institutionalized (0.4%), and of other types of nonresponse such as temporary absences and special circumstances (0.7%). Within responding households person level nonresponse was very low: 1.8% for the General questionnaire and 1.1% for the Health questionnaire. We first identify cases that are not nonresponses for longitudinal and cross-sectional purposes.

For longitudinal purposes a death is considered a valid survey response. Panel Members who had died before Cycle 2 had their status recorded as such and the 1996 portion of their data coded as "Not Applicable" on the Longitudinal microdata file. Panel Members who moved to an institution or to the Territories were followed-up and their responses were used for longitudinal purposes. Panel Members who left the country were not followed-up but were treated as longitudinal nonrespondents even though it would have been more accurate for some analyses to have considered them as having left the scope of the study. This treatment was chosen because such persons would fall back in-scope should they move back to Canada.

For cross-sectional purposes all the cases presented in the preceding paragraph were treated as out-of-scope situations. This was acceptable because the separate Institutional and Territorial survey vehicles assumed the cross-sectional coverage of these particular in-scope populations. Out-of-scope units were not on the PUMFs but, as they represented other such units, they were treated for weighting purposes like respondents in all the weight adjustment steps except the integration and post-stratification steps.

Refusals and cases where questionnaires were missing for reasons other than those given in the preceding paragraphs were defined as nonresponses. As will be seen, a distinction was later made between "full" and "partial" longitudinal nonrespondents to accommodate different users.

3.2 Approach for Creating Nonresponse Adjustment Weighting Cells

Two statistical approaches for creating response weighting cells involve segmentation modelling and logistic regression. An example of the latter is given in Czajka, Hirabayashi, Little and Rubin (1992). The authors obtained advance taxation estimates from early tax filer returns using adjustment weighting cells that were based on ranges of propensities to file early. Logistic regression was used to estimate tax filers' propensities to file early. The longitudinal Survey of Labour and Income Dynamics (SLID) provides another example involving logistic regression (Grondin 1996). Sample units' response indicators were regressed on known (dichotomous) characteristics. Adjustment cells for nonresponse were generated by cross-classifying the sample units using all the significant covariates. In order to respect minimum cell sizes and response rates some collapsing was done starting with cells sharing all but the least significant covariates.

In the segmentation modelling approach a decision tree structure is generated from the data by successively splitting the data set such that, at each node, the most significant predictor for the response variable is used to define the following split. The splitting continues until one cannot find any significant variable for the split or minimum cell size requirements cannot be respected. An early application of segmentation modelling for nonresponse adjustment was with respect to the Panel Study of Income Dynamics (Institute for Social Research 1979). Because of its advantages, given below, the NPHS adopted the segmentation modelling approach using the CHAID algorithm developed by Kass (1980). The CHAID (Chi-square Automatic Interaction Detection) algorithm uses χ^2 tests to define splits for categorical predictors and retains the most significant split at each stage. The splitting, into two or more categories, is done differently for ordered and unordered predictors. CHAID was applied using the Knowledge Seeker software program (ANGOSS Software 1995). Note that Knowledge Seeker applies CHAID to continuous predictors by first transforming them into ordered discrete variables.

Advantages and disadvantages of the logistic and CHAID approaches are known and documented (for example see Kalton and Kasprzyk 1986). The logistic regression approach is based on theory familiar to many analysts, and can be programmed using a number of standard statistical packages. It also provides individual estimates of response propensity that can be used directly to adjust the weights of respondents. However, to ensure reasonable program execution times the number of variable and interaction terms used must usually be limited. Collapsing cells can also become complicated, as in the case of SLID above. The CHAID algorithm offers the advantages of accepting a large number of covariates and, by its decision tree structure, easily accommodating interactions among them. Moreover, minimum cell size requirements can easily be incorporated as program execution parameters. Its main disadvantages are a less familiar theoretical underpinning (Knowledge Seeker is advertised as a "data mining" tool) and the limited documentation and software available for its implementation. It should also be mentioned that, while some statistical packages such as SUDAAN and PC CARP can incorporate the sample design when fitting logistic models to survey data, this is not the case with CHAID. The NPHS tried to address this limitation by including as predictor variables characteristics that were related to the sample design (see Section 3.4).

Two empirical studies comparing the logistic and CHAID approaches for the treatment of nonresponse obtained different results. Rizzo, Kalton and Brick (1996) did not find much of a difference between the two approaches for the Survey of Income and Program Participation. On the other hand Dufour, Gagnon, Morin, Renaud and Särndal (1998), in a simulation study for SLID, obtained a lower bias after nonresponse adjustment with the CHAID approach.

3.3 Adjusting for Nonresponse in the Core Sample

Nonresponse adjustments had to be developed for each PUMF: Longitudinal, General (Cross-Sectional) and Health (Cross-Sectional). We will deal with the General PUMF first.

As Figure 2 showed, the weighting strategy for the General PUMF required separate adjustments for nonresponse at the household and at the person levels. In creating adjustment cells for household level nonresponse, characteristics of the Panel Member as well as those of the household were considered as nonresponse predictors. This was done for three reasons. Firstly, as the Panel Member was the link to the household in Cycle 2, his or her characteristics may be related to finding the household and obtaining a response (the first contact will often be through him or her). Secondly, a few personal characteristics of the Panel Member, such as race, are in some sense household characteristics. Finally, using Panel Member characteristics was not incompatible with our need to use a variety of information for the construction of weighting cells. If Panel Member characteristics are not significant, then CHAID simply does not retain them.

Person level nonresponse to the General component occurred when the information was available for some but not all of the household members, perhaps due to members' refusals or temporary absences. Given the low 1.8% nonresponse rate at the person level, it was felt that the creation of weighting cells based on province-age-sex categories (as in Cycle 1) would be sufficient for our needs.

In contrast to the General PUMF, the adjustments for household and person level nonresponse for both the Longitudinal and the Health PUMFs could be combined into a single adjustment as they concerned only one person – the Panel Member. A single set of adjustment cells thus needed to be created.

For the Longitudinal PUMF it was noted that the data items came from both the General and Health components but that response rates for the two components were different. This difference produced data with different Cycle 1-Cycle 2 reporting patterns: GH-GH, GH-G*, G*-GH, G*-G*, not to mention longitudinal nonresponse patterns GH-** and G*-**, where the letters stand for the component reported each Cycle ("*" if not reported). To maximize the utility of the data it was decided to do two adjustments for longitudinal nonresponse. One adjustment would be for the "Full Longitudinal Response" pattern GH-GH. In other words, all other response patterns would be considered as nonresponses. The other adjustment would be for the "Partial Longitudinal Response" pattern which included cases where, at minimum, General information was available for each cycle (patterns GH-GH, GH-G*, G*-GH and G*-G*). The Full Response data set could be used by researchers who would like to analyse a full longitudinal data set covering the entire questionnaire contents. The Partial Response data set could be of use to researchers primarily interested in the types of variables that are on the

General questionnaire. As the counts in Table 1 below show, the Partial Longitudinal Response data set is only about 3% larger than the Full Longitudinal Response data set.

Table 1
Longitudinal Response Patterns

Response Type		Cycle 1-2 Response Pattern	Number of records
Full	Partial		
■	■	GH-GH	15,670
	■	GH-G*	110
	■	G*-GH	366
	■	G*-G*	22
		GH-**	1,014
		G*-**	94
		Total	17,276

Based upon the above, adjustment cells must be built for five types of responses (or nonresponses) in Cycle 2:

- General PUMF – household response
- General PUMF – person response
- Health PUMF – combined response
- Longitudinal PUMF – full response
- Longitudinal PUMF – partial response

Only three sets of adjustment cells were created for those response types. Adjustment cells created for the General PUMF household level responses were also used for the Longitudinal PUMF partial responses because getting a response from a household led almost always to obtaining a partial response for the longitudinal member (there were 53 exceptions). Likewise, adjustment cells generated for full respondents on the Longitudinal PUMF were used for the Health PUMF responses. Although there were 366 more cases of responses of the latter type (pattern G*-GH) it was considered that the same response mechanism was at work in both instances. The third set of adjustment cells was for person level responses on the General PUMF. Province-age-sex categories were used, as was done in Cycle 1.

Note that, although the same adjustment cells would serve for different data sets, the nonresponse weight adjustments would be calculated separately for each data set type. Thus, the 366 records with response pattern G*-GH would be treated as respondents when adjusting weights for the Health PUMF, but as nonrespondents when adjusting weights for full respondents on the Longitudinal PUMF.

3.4 Creation of Weighting Adjustment Cells

Separate sets of weighting adjustment cells were created for each province. The first step consisted of identifying the variables to consider. With CHAID the number of variables that could be considered was not really an issue, and different types were considered. Characteristics of the household, or dwelling, as well as personal characteristics of the Panel Member would of course be considered. In an effort to incorporate the design of the survey into the analysis some characteristics that were related to the design of the survey or to the sampling weight were also considered. These included geographical variables such as the Census Metropolitan Area code or the Urban/Rural indicator, special Cycle 1 design variables such as the flag identifying households for screening and the "Adult/Children" household type, and characteristics related to the application of those design variables, such as the presence in the household of a member aged under 25 or of a child. The household size was used as it was a household characteristic and was also related to the sample weight. From experience, it was also decided to include, in addition to the household income characteristic, a dummy characteristic that identified if household income had been reported in Cycle 1 or not. As a change of address can lead to an unable-to-trace nonresponse situation we would have liked to use a change-of-address identifier. However, in some nonresponse and no contact situations it was difficult to ascertain whether the Panel Member had indeed moved. In the end a "Mover" variable, which identified whether the Panel Member had changed provinces between Cycles, was used in the analysis even though this was far from ideal because the default value would be "no". Personal characteristics from the Health questionnaire component such as Smoker/Drinker status, Health Index Level and Mental Health/Distress Scale were not used because they were not available for almost 500 Panel Members.

The variables used are listed below. The nonresponse indicator, which was the dependent variable, had its values assigned according to the definition of nonresponse being used.

DESIGN/GEOGRAPHICAL VARIABLES

PROVINCE	The analysis was done at the provincial level
CMA	Census Metropolitan Area (0 if not a CMA)
URBAN	Urban/Rural Indicator
REJECT	Flag if the unit (household) was eligible for screening
ACFLAG	"Adult/Children" design classification for the unit

DWELLING/HOUSEHOLD CHARACTERISTICS

DWELL	Dwelling type (10 categories)
OWNER	Owner/Renter Indicator
FAMTYP	Family Type (unattached individual, single parent hhld., married couple hhld., other)
INC	Household Income Adequacy (5 levels)
INCNR	Nonresponse flag for INC
INCSRC	Main source of income (6 categories)
*HHSIZE	Household size
UND25	Indicator of members under 25 years old
KIDS	Indicator of children under 12 years old

PERSONAL CHARACTERISTICS OF PANEL MEMBER

SEX	Sex
AGE	Age in years
AGE16	Indicator if aged 16 or older
MARIT	Marital Status
FAMID	Family Identifier within household (A, B, C, ...)
RACE	White, Black, Aboriginal or Other
BORN	Place of birth (Canada, USA/Mexico, S. America/Africa, Europe/Australia, Asia)
AGIMM	Age at immigration (for immigrants)
*MOVED	Changed province indicator (see text)
*EDUC	Highest level of education (12 categories)
*STUDNT	Student Indicator
MACT	Main Activity (8 categories)
*NUMJOB	Number of jobs held last year (in Cycle 1)
RESTR	Restriction of Activity Flag
*CAUSE	Main Cause of Restriction (12 categories)
CONSUL	Number of consultations with a Medical Doctor
INHOSP	Overnight Hospital Patient Flag
*CHRONIC	Number of Chronic Conditions

* Indicates the variable was never significant when forming classes.

Figure 3 presents the variables chosen by CHAID to build nonresponse adjustment cells for Household Level Response and for the Full Longitudinal Response in each province. For reasons of confidentiality detail is not given on the individual cell sizes and response rates (some of the variables used are considered sensitive and are not on the PUMFs). However, summary information on the cell construction is given in Tables 2 and 3.

Table 2
Response Adjustment Cell Characteristics
(for Household Level Response)

Prov.	#Units	#NR	Cell Sizes			Cell % NR rates		
			min.	max.	avg.	min.	max.	avg.
Nfld.	1,082	40	354	728	541	1.4	4.8	3.7
P.E.I.	1,037	51	81	478	259	3.0	13.6	4.9
N.S.	1,085	55	46	374	217	0.7	10.9	5.1
N.B.	1,125	59	32	986	281	2.6	34.4	5.2
Que.	3,000	133	123	2363	750	1.8	12.1	4.4
Ont.	4,307	315	44	1,038	308	0.9	25.8	7.3
Man.	1,205	50	1,205	1,205	1,205	4.1	4.1	4.1
Sask.	1,168	59	37	626	167	1.6	35.3	5.1
Alta.	1,544	116	32	837	221	3.9	36.7	7.5
B.C.	1,723	149	82	678	246	5.2	29.0	8.6

The results vary by province. As expected, provinces with larger sample sizes such as Ontario, Quebec, British Columbia and Alberta yield "richer" decision trees. Cell sizes and response rates also vary considerably. In Table 2 on household-level response Manitoba has only one cell, and 88% of New Brunswick's sample is located in one cell. Likewise, in Table 3 almost all of Newfoundland's sample is placed in one of its two cells. Cell nonresponse rates approaching 40% are observed in a few provinces.

Table 3
Response Adjustment Cell Characteristics
(for Full Longitudinal Response)

Prov.	#Units	#NR	Cell Sizes			Cell % NR rates		
			min.	max.	avg.	min.	max.	avg.
Nfld.	1,082	73	35	1,047	541	6.2	22.9	6.7
P.E.I.	1,037	80	41	453	207	4.1	26.8	7.7
N.S.	1,085	96	236	555	362	6.5	14.3	8.8
N.B.	1,125	86	59	819	281	4.8	16.8	7.6
Que.	3,000	211	42	2,202	375	2.5	37.8	7.0
Ont.	4,307	470	34	619	196	0.0	38.0	10.9
Man.	1,205	91	186	763	402	5.6	15.1	7.6
Sask.	1,168	83	90	339	195	0.0	28.9	7.1
Alta.	1,544	148	41	866	172	1.1	39.0	9.6
B.C.	1,723	192	33	408	191	4.5	37.3	11.1

Figure 3 shows a variety in the characteristics of weighting classes both between provinces and between the two types of nonresponse within provinces. In all provinces except Alberta the CHAID algorithm uses different characteristics for the two nonresponse types as early as at the first or second level of branching. A few characteristics figure prominently in the early stages of branching in many of the trees for both types of nonresponse. They are: household income adequacy level (INCNR), income non-response flag (INCNR), Race (RACE) and Place of Birth (BORN).

In Figure 3a household income and its related variables (INCNR and INCSRC), Owner/Renter status (OWNER), Race, Place of Birth and Dwelling Type (DWELL) all were used three or more times in forming weighting classes for Household Level nonresponse. It is also remarked that in five out of nine provinces a personal characteristic of the Panel Member was selected at the first stage of branching by CHAID. This supports the decision to consider personal characteristics when adjusting for household level nonresponse.

In Figure 3b for Full Longitudinal nonresponse Census Metropolitan Area (CMA), Marital Status (MARIT) and SEX, although not as important at first as Income, Race and Place of Birth, were used the most often (5 times each).

As mentioned earlier, design variables such as the rejection flag (REJECT) and the "Adult/Children" flag (ACFLAG) were considered in an attempt to incorporate the sample design in the CHAID analyses. Although these variables were selected only once each, household characteristics used by the design, such as the presence of children (KIDS) and under 25 year-olds (UND25) did get selected occasionally. Household size was not used but

3A. HOUSEHOLD LEVEL RESPONSE CLASSES

3B. FULL LONGITUDINAL RESPONSE CLASSES

(TYPES OF CHARACTERISTICS: **DESIGN/GEOGRAPHICAL**, DWELLING/HOUSEHOLD, *PERSONAL*)

Figure 3. Provincial Response Classes Obtained for Cycle 2 Nonresponse

Family Type (FAMTYP), which is related to the household size, did get selected twice.

The adjustment cells produced by CHAID were reviewed but only in rare cases were they manually altered. Within each cell, the weights of responding units were prorated to add up to the total weight for responding and nonresponding units. The magnitude of the nonresponse weight adjustments never exceeded 1.83.

4. CONCLUSION

This paper presented the strategy developed for the treatment of both longitudinal and cross-sectional nonresponse to Cycle 2 of the NPHS. The approach adopted took into account practical considerations such as the need for an easy-to-use, yet statistically valid, way of defining weight adjustment cells and the need to provide a more useful data set (by having separate adjustments for "Full" and "Partial" Longitudinal Responses) while keeping the additional effort required at a reasonable level (*e.g.*, by using weight adjustment cells for more than one purpose). Having chosen the CHAID algorithm approach rather than logistic regression allowed us more freedom in the number and choice of variables to consider: many design variables and personal variables could thus easily be considered – and some were retained. This did seem to offer some promise about the usefulness of those characteristics in the treatment of nonresponse.

On the other hand, a tight production schedule meant that some analysis that we wished to have carried out was not performed. It would have been interesting to pursue the possibilities offered by the CHAID algorithm, for example, as CHAID allows the use of a categorical response variable we could have classified sample units into three groups: live respondents, dead or out-of-scope units, and nonrespondents. We would have liked to do our own comparison of CHAID with a logistic regression approach. We could also have attempted to use Health questionnaire variables such as the Health Index or Smoker/Drinker status in defining weight adjustment cells, although their usefulness would have been reduced by the fact that they were not present for all units (they are missing in response patterns G*-GH, G*-G* and G*-**). Decisions to use the same weight adjustment cells for different types of nonresponse should be revisited. For example, could the adjustment cells built for household level response have been more suitable for the Health cross-sectional nonresponse? An attempt to compare the efficiency of various nonresponse adjustment strategies would involve evaluating their impact on the variance of estimators. We could also evaluate the impact of our Cycle 2 nonresponse adjustment on the nonresponse bias by using the Cycle 1 data available for all panel members. Estimates using the full sample would be compared to nonresponse-adjusted estimates generated from the responding units.

Cycle 3 itself will present new problems. A global sample "top-up" is planned in that year, which will have an impact on our cross-sectional estimation strategy and therefore on the treatment of nonresponse. As longitudinal nonresponse is increasing we will have to consider side effects of the weighting adjustment such as the possible creation of outlier weights. Providing sets of weights for different types of longitudinal analyses will become cumbersome as the number of "partial" response patterns will increase. How many patterns can reasonably be treated, and which ones? The choice of additional information, such as Mover status, for the treatment of nonresponse should be reconsidered. Some imputation for nonresponse will likely be used in Cycle 3: the question is how to reconcile imputation with the weight adjustment approach to nonresponse. As can be seen, a lot of work remains to be done for the NPHS. One hopes that we will have time to investigate many of those issues before Cycle 3 processing is finished.

ACKNOWLEDGEMENTS

The authors would like to thank Douglas Yeo and the referees for their helpful comments.

REFERENCES

ANGOSS SOFTWARE (1995). *Knowledge SEEKER IV for Windows – User's Guide*. ANGOSS Software International Limited.

CATLIN, G., and WILL, P. (1992). The National Population Health Survey: Highlights of initial developments. *Health Reports*, Statistics Canada, 4, 313-319.

CZAJKA, J.L., HIRABAYASHI, S.M., LITTLE, R.J.A., and RUBIN, B.R. (1992). Projecting from advance data using propensity modelling: an application to income and tax statistics. *Journal of Business & Economic Statistics*, 10, 117-131.

DAVID, M.H., LITTLE, R., SAMUHEL, M., and TRIEST, R. (1983). Imputation models based on the propensity to respond. *Proceedings of the Business and Economic Statistics Section, American Statistical Association*, 168-173.

DUFOUR, J., GAGNON, F., MORIN, Y., RENAUD, M., and SÄRNDAL, C.-E. (1998). Measuring the impact of alternative weighting schemes for longitudinal data. *Proceedings of the Section on Survey Research Methods, American Statistical Association*. To appear.

ERNST, L. (1989). Weighting issues for household and family estimates. In *Panel Surveys*, (Eds. D. Kasprzyk, G. Duncan, G. Kalton and M.P. Singh). New York: John Wiley and Sons, 135-159.

GRONDIN, C. (1996). Pondération longitudinale – Première vague de l'EDTR. Internal note, Social Survey Methods Division, Statistics Canada.

INSTITUTE FOR SOCIAL RESEARCH (1979). A Panel Study of Income Dynamics: Procedures and Tape Codes – 1978 Interviewing Year – Wave XI – A Supplement. The University of Michigan.

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 categorical data. *Applied Statistics*, 29, 119-127.

LAVALLÉE, P. (1995). Cross-sectional weighting of longitudinal surveys of individuals and households using the weight share method. *Survey Methodology*, 21, 25-32.

LITTLE, R.J.A. (1986). Survey nonresponse adjustments for estimates of means. *International Statistical Review*, 54, 137-157.

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

RIZZO, L., KALTON, G., and BRICK, J.M. (1996). A comparison of some weighting adjustment methods for panel nonresponse. *Survey Methodology*, 22, 43-53.

ROSENBAUM, P.R., and RUBIN, D.B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70, 41-55.

SKINNER, C.J., and RAO, J.N.K. (1996). Estimation in dual frame surveys with complex designs. *Journal of the American Statistical Association*, 91, 433, 349-356.

STATISTICS CANADA (1995). National Population Health Survey 1994-95 public use microdata file. Catalogue No. 82F0001XCB.

STUKEL, D.M., MOHL, C.A., and TAMBAY, J.L. (1997). Weighting for cycle two of Statistics Canada's National Population Health Survey. *Proceedings of the Survey Methods Section, Statistical Society of Canada*, 111-116.

TAMBAY, J.L., and CATLIN, G. (1995). Sample design of the National Population Health Survey. *Health Reports*, Statistics Canada, 7, 29-38.

THOMSEN, I. (1973). A note on the efficiency of weighting subclass means to reduce the effects of nonresponse when analyzing survey data. *Statistik Tidskrift*, 11, 278-283.

Survey Methodology, December 1998
Vol. 24, No. 2, pp. 157-169
Statistics Canada

Estimates of the Errors in Classification in the Labour Force Survey and Their Effect on the Reported Unemployment Rate

MICHAEL D. SINCLAIR and JOSEPH L. GASTWIRTH[1]

ABSTRACT

This paper studies response errors in the Current Population Survey of the U.S. Bureau of the Census and assesses their impact on the unemployment rates published by the Bureau of Labour Statistics. The measurement of these error rates is obtained from reinterview data, using an extension of the Hui and Walter (1980) procedure for the evaluation of diagnostic tests. Unlike prior studies which assumed that the reconciled reinterview yields the true status, the method estimates the error rates in both interviews. Using these estimated error rates, we show that the misclassification in the original survey creates a cyclical effect on the reported estimated unemployment rates. In particular, the degree of underestimation increases when true unemployment is high. As there was insufficient data to distinguish between a model assuming that the misclassification rates are the same throughout the business cycle, and one that allows the error rates to differ in periods of low, moderate and high unemployment, our findings should be regarded as preliminary. Nonetheless, they indicated that the relationship between the models used to assess the accuracy of diagnostic tests, and those measuring misclassification rates of survey data, deserves further study.

KEY WORDS: Misclassification errors; Unemployment rates; Diagnostic tests; Reconciliation; Reinterview surveys; Response errors.

1. INTRODUCTION

Several articles, Poterba and Summers (1986 and 1995) and Abowd and Zellner (1985) used the data from the U.S. Bureau of the Census' reinterview program to estimate the misclassification rates of the Current Population Survey (CPS) and assessed their impact on estimates of labour market transition rates. The estimated misclassification rates were based on the assumption, that a particular reinterview method, reconciliation, yields the "truth." Biemer and Forsman (1992), Forsman and Schreiner (1991) and unpublished research of the U.S. Bureau of the Census (1963), have questioned this assumption. The purpose of this paper, is to provide estimates of the misclassification rates, from response errors in all interviews and reinterviews and to explore their impact on the reported unemployment rates. In contrast to the earlier papers that were concerned with gross flow, we emphasize the accuracy of the labour force estimates themselves. Our approach is based on extending the Hui and Walter (1980) paradigm, for estimating error rates of medical diagnostic tests to trinomial classifications. An advantage of this method is that, no single interview needs to be considered as perfect.

Under certain assumptions, Hui and Walter (1980) developed a method for estimating the error rates associated with a new diagnostic screening test, using a confirmatory test with an unknown low error rate. By treating the reinterview as the confirmatory test, and the original survey as the screening test, this methodology can be used to estimate the error rates in the original survey, and the reinterview and the prevalence rates of the trait screened for. The Hui and Walter (1980) method requires two subpopulations with different prevalence rates of the characteristic. While the two tests may have different error rates, the error rates for each test are assumed equal in the two subpopulations. Furthermore, the model (described in more detail in the appendix) assumes that the errors from the two tests conditioned on the subject's true status, are independent.

The Hui and Walter method was developed for dichotomous test outcomes, and was adapted by Sinclair and Gastwirth (1996) to study misclassification of labour force participation rates. Here, we extend the approach to account for three classifications: unemployed, employed and not in the labour force (NLF), and assess the effect of the misclassification on the reported unemployment rates. The basic model is presented in section two. The reinterview program data, to which the model will be fitted, are described in section three. The resulting error rates are given in section four, along with the "adjusted" unemployment rates, which account for the estimated classification errors. In addition, a measure of accuracy, the predictive value, used in the medical screening literature, is applied to the unemployment rate in section four. It shows that the probability an individual classified as unemployed in the CPS is actually unemployed, varies with the true level of unemployment.

2. THE DATA AND THE MODEL

Labour force reinterview data consists of trinomial responses from both the original survey and a subsequent reinterview. This data for a given subpopulation and year,

[1] Michael D. Sinclair, Senior Statistician, Mathematical Policy Research, 101 Morgan Lane, Plainsboro, N.J. 08536, U.S.A.; Joseph L. Gastwirth, Professor of Statistics and Economics, George Washington University, 2201 G Street Rm. 315, Washington, D.C. 20052, U.S.A.

is summarized in a 3 × 3 table, where the observed frequency counts of persons in the table, is denoted by, n_{ygij}. With this notation:

- y denotes the year;
- g denotes subpopulation membership, $g = 1$ or 2;
- i denotes the subject's classification by the original survey, $i = 1$ for unemployed, $i = 2$ for employed and $i = 3$ for NLF; and
- j denotes the same subject's classification by the reinterview, $j = 1, 2$ and 3.

We denote the true prevalence rate for each labour force status, $i = 1, 2$ and 3, by π_{ygi}, for subpopulation g and year y. Throughout this paper, we will use the term prevalence rate, to refer to the proportion of persons in one of the three labour force categories (*e.g.*, π_{yg1}). Note that the fraction, π_{yg3} of the population in the NLF category equals $(1 - \pi_{yg1} - \pi_{yg2})$, and that the true unemployment rate in year y for subpopulation g, is equal to π_{yg1} divided by $(\pi_{yg1} + \pi_{yg2})$.

Each classification rate, β_{ygrij}, is defined as the probability that the r-th data collection process, $r = 1$ for the original survey, and $r = 2$ for the reinterview, will classify a person in year y from subpopulation g, to be in category i, $i = 1$, 2 and 3 when the true status of the individual is category j. For example, β_{11131} denotes the probability that in the first year ($y = 1$), a person from the first subpopulation ($g = 1$), was classified by the original survey ($r = 1$) as NLF ($i = 3$) when the person's true status is unemployed ($j = 1$). The classification rates can be divided into two groups, corresponding to those associated with a correct classification, and those associated with an erroneous classification. For each y, g and r, the probability that survey method r, classifies a truly unemployed person in year y from subpopulation g correctly as unemployed, is equal to $\beta_{ygr11} = (1 - \beta_{ygr21} - \beta_{ygr31})$. The corresponding probabilities for employed and NLF are respectively, $\beta_{ygr22} = (1 - \beta_{ygr12} - \beta_{ygr32})$, and $\beta_{ygr33} = (1 - \beta_{ygr13} - \beta_{ygr23})$. With conditional independence of the original survey and the reinterview classification rates, the expected observed frequencies, as expressed in terms of the given notation, for each of the nine cells associated with a particular year y and subpopulation g are:

$$E(n_{yg11}) = n_{yg..}(\pi_{yg1}(1-\beta_{yg121}-\beta_{yg131})(1-\beta_{yg221}-\beta_{yg231}) + \pi_{yg2}\beta_{yg112}\beta_{yg212} + (1-\pi_{yg1}-\pi_{yg2})\beta_{yg113}\beta_{yg213})$$

$$E(n_{yg12}) = n_{yg..}(\pi_{yg1}(1-\beta_{yg121}-\beta_{yg131})\beta_{yg221} + \pi_{yg2}\beta_{yg112} * (1-\beta_{yg212}-\beta_{yg232}) + (1-\pi_{yg1}-\pi_{yg2})\beta_{yg113}\beta_{yg223})$$

$$E(n_{yg13}) = n_{yg..}(\pi_{yg1}(1-\beta_{yg121}-\beta_{yg131})\beta_{yg231} + \pi_{yg2}\beta_{yg112}\beta_{yg232} + (1-\pi_{yg1}-\pi_{yg2})\beta_{yg113}(1-\beta_{yg213}-\beta_{yg223}))$$

$$E(n_{yg21}) = n_{yg..}(\pi_{yg1}\beta_{yg121}(1-\beta_{yg221}-\beta_{yg231}) + \pi_{yg2}(1-\beta_{yg112}-\beta_{yg132})\beta_{yg212} + (1-\pi_{yg1}-\pi_{yg2})\beta_{yg123}\beta_{yg213})$$

$$E(n_{yg22}) = n_{yg..}(\pi_{yg1}\beta_{yg121}\beta_{yg221} + \pi_{yg2}(1-\beta_{yg112}-\beta_{yg132}) * (1-\beta_{yg212}-\beta_{yg232}) + (1-\pi_{yg1}-\pi_{yg2})\beta_{yg123}\beta_{yg223})$$

$$E(n_{yg23}) = n_{yg..}(\pi_{yg1}\beta_{yg121}\beta_{yg231} + \pi_{yg2}(1-\beta_{yg112}-\beta_{yg132})\beta_{yg232} + (1-\pi_{yg1}-\pi_{yg2})\beta_{yg123}(1-\beta_{yg213}-\beta_{yg223}))$$

$$E(n_{yg31}) = n_{yg..}(\pi_{yg1}\beta_{yg131}(1-\beta_{yg221}-\beta_{yg231}) + \pi_{yg2}\beta_{yg132}\beta_{yg212} + (1-\pi_{yg1}-\pi_{yg2})(1-\beta_{yg123}-\beta_{yg113})\beta_{yg213})$$

$$E(n_{yg32}) = n_{yg..}(\pi_{yg1}\beta_{yg131}\beta_{yg221} + \pi_{yg2}\beta_{yg132}(1-\beta_{yg212}-\beta_{yg232}) + (1-\pi_{yg1}-\pi_{yg2})(1-\beta_{yg123}-\beta_{yg113})\beta_{yg223})$$

$$E(n_{yg33}) = n_{yg..}(\pi_{yg1}\beta_{yg131}\beta_{yg231} + \pi_{yg2}\beta_{yg132}\beta_{yg232} + (1-\pi_{yg1}-\pi_{yg2})(1-\beta_{yg123}-\beta_{yg113})(1-\beta_{yg213}-\beta_{yg223})),$$

where, the total sample size for year y and subpopulation g is denoted by $n_{yg..}$.

The model has 14 parameters (six error rates for the original survey, $r = 1$, six error rates for the reinterview, $r = 2$, and two unique prevalence rates) for each subpopulation and year. On the other hand, the 3 × 3 table for a given year and subpopulation has only 8 independent frequencies, or degrees of freedom. As a result, the model is overparameterized and the number of parameters must be reduced for estimation purposes. The Hui and Walter paradigm enables us to accomplish this.

3. APPLICATION OF THE MODEL AND THE CPS REINTERVIEW PROGRAM

The U.S. Bureau of the Census' Current Population Survey Reinterview Program (U.S. Bureau of the Census 1963) is conducted approximately two weeks after the initial survey, to measure response errors, and to evaluate interviewer performance. The sample design for the reinterview, consists of the self-weighting random sample of households (Levy and Lemeshow 1980) among the selected interviewer assignments. The sample size is about 1/18 of the monthly CPS sample of 50,000 to 60,000 household interviews. Two reinterview procedures are conducted. Three-fourths to four-fifths of the sample cases participate in a response-bias study. Here, an initial reinterview is conducted and after this interview is

completed, the reinterviewer reconciles disagreements with the respondent, between the original and the initial reinterview responses. Hence, in the response-bias study, up to two reinterview responses may be obtained from each subject; the first unreconciled reinterview response and a reconciled reinterview response. The remaining one-fifth to one-fourth of the sample households receive a reinterview without reconciliation.

In the response bias study, the reinterviewer is instructed not to look at the original survey responses until the initial reinterview is completed. Forsman and Schreiner (1991) and Schreiner (1980) suggested that the reinterviewers may change the initial reinterview responses to match the original response, as they observed that the rate of disagreement between the original responses and the initial reinterview responses were greater in the unreconciled sample. Sinclair (1994) and Sinclair and Gastwirth (1996) showed that these differences were statistically significant. As a result, the reconciliation process creates a correlation between the original and unreconciled reinterview responses, in the reconciled sample. Hence, we decided to limit our analysis to the original and unreconciled reinterview data from the unreconciled study sample. For the purposes of this study, we will assume that in the unreconciled sample, the errors from the original survey and the unreconciled reinterview conditioned on the respondent's true status, are independent.

To apply the Hui and Walter approach, one needs two subpopulations with different prevalence rates. As males and females are known to have different labour force participation rates, we use them. We also need to assume, that the classification error rates are equal in the two subpopulations, males and females, *i.e.*, $\beta_{y1rij} = \beta_{y2rij}$. At this stage, we assume that the classification error rates for the original survey and the unreconciled reinterview, may be different, and that they may differ by year. With this reduction, for the two subpopulations, in a given year, we now have a total of 12 error rate parameters and 4 prevalence rates, yielding 16 parameters. Since two 3×3 tables contain a total of 16 degrees of freedom, estimation is possible. In this paper, we have analyzed the CPS unreconciled reinterview sample data for the period 1981 through 1990. Complete yearly data for 1987 as well as more recent data, were not available from the U.S. Bureau of the Census.

The CPS estimates of the unemployment rate are published regularly by the Bureau of Labour Statistics (BLS) (see Bureau of Labour Statistics 1992). Since the reinterview is a sub-sample of the full CPS sample, the original survey estimates of the unemployment rate from the reinterview sample, will differ from the BLS published results. Data processing procedures are used on the full sample CPS, that are not applied to the reinterview data. For example, the full CPS sample is weighted, based on the sample selection probabilities, and nonresponse adjustment factors are applied to the data. Given these differences, the estimated prevalences from our model, based solely on the reinterview data, are not directly comparable to the BLS reported values. We have used the CPS reinterview data, primarily to estimate the error rates in the original survey. Furthermore, we have treated the unreconciled reinterview data as a simple random sample of the population, for analysis and hypothesis testing purposes, throughout this paper. Using these error rate estimates, we estimate adjusted Bureau of Labour Statistics (BLS) unemployment rates, where the term adjusted, means that the reported values have been modified to account for the misclassification in the survey. The formula for estimating the true unemployment rate as a function of the reported BLS prevalences from the full CPS sample, and the estimated classification error rates as obtained from the unreconciled reinterview data, is given in the appendix.

4. DATA ANALYSIS AND RESULTS

The first step in preparing our final estimates, was to obtain the parameter estimates, for each of nine yearly data tables, using the SAS NLIN procedure with the Gauss-Newton weighted least squares method. As the reinterview procedures remained constant during the period, we decided to test the hypothesis, that each of the error rates remained equal across the years studied, *i.e.*, $\beta_{ygrij} = \beta_{y'grij}$ for all years $y \neq y'$. In conjunction with the basic assumption, that the error rates for males and females are equal, *i.e.*, $\beta_{y1rij} = \beta_{y2rij}$, this implies, $\beta_{ygrij} = \beta_{y'g'rij}$ for all $y \neq y'$ and $g \neq g'$.

From the two sets of results, we conducted a likelihood ratio test under the assumption, that the reinterview sample is a simple random sample of the population, to test the assumption that each of the error rates was the same for all years. The likelihood ratio statistic, $-2 \log \lambda$ with 96 degrees of freedom (144 parameters in the full model less 48 parameters in the reduced model) yielded a value of 84.06 with a p-value of 0.8027. Hence, the data is consistent with the reduced model, enabling us to use the reduced model estimates and to simplify the notation. We will now use β_{rij} to denote β_{ygrij} for all g and y.

The estimated error rates for the original survey and for the unreconciled reinterview, are presented in Tables 1 and 2, respectively, with their estimated standard errors. The estimated reinterview error rates in Table 2, are similar to corresponding error rate estimates for the original survey. This similarity indicates that the U.S. Bureau of the Census unreconciled reinterview serves as an effective replication. The error rate estimates show that the CPS survey procedures are able to classify the employed, and those not in the labour force, quite accurately. On the other hand, these procedures do not perform well for classifying the unemployed, as the proportion of truly unemployed persons who are classified as unemployed, $(1 - \beta_{121} - \beta_{131})$, is only 0.8397.

For comparative purposes we conducted an analysis of the 75% sample reconciled reinterview data, for the same

1981-1990 period, under the assumption that the reconciled responses were error-free. We created a 3 × 3 table for the number of persons classified by the original interview, in each labour force category, by the number of persons classified by the reconciled reinterview, in each labour force category. The data is given in Table 3. The table frequencies report aggregate data, by year and sex, so that the error rates derived from this table, are comparable to our model. Using the column status, as the true status, one computes an estimate of the error rates. For example, the estimate of β_{121}, the probability that an unemployed person will be classified in the original survey as employed, is 332/17,681 = 0.0188. These error rates are presented in Table 1, to illustrate how the estimated error rates from our method, based on the unreconciled data, differ from those relying on the assumption that the reconciled reinterview is perfect.

Table 1 also presents the estimates of the original survey error rates, as obtained by Poterba and Summers (1986), using reinterview data (combined for both sexes) for the first half of 1981. The Poterba and Summers' method uses both the data from the unreconciled and reconciled samples to estimate the error rates. These authors assume that in the reconciled sample, the interviewers use the original survey data provided, to influence the initial reinterview response. As a result, they assume that a reconciled value is only obtained for a portion of persons, that should have had a discrepancy between the original survey and the initial reinterview. When a reconciled value is obtained, Poterba and Summer assume that the reconciled data is error-free. With these assumptions, they use the unreconciled sample to estimate the incidence of the error, and the reconciled data to provide the information on the true labour force status. In summary, both the Poterba and Summers method, and the reconciled reinterview estimates, rely on the reconciled reinterview data being perfect.

Table 4 presents the reported BLS yearly unemployment rates among those in the labour force, for males and females combined, in comparison to the estimated adjusted unemployment rates based on: (1) our error rate estimates, (2) Poterba and Summers (1986) error rates, and (3) error rates assuming the reconciled reinterview is perfect. If the results in Table 4, are sorted by the value of the BLS reported unemployment rate, an apparent trend is observed in the bias in the original CPS estimates. Figure 1 shows that the reported values, tend to overestimate the actual unemployment rate of persons in the labour force in low unemployment years (1989, 1988 and 1990), and to underestimate the unemployment rate in high unemployment years (1982-1983). Furthermore, the bias associated with our method is shifted upward from the two other approaches. All three methods indicate cyclical effect, the smallest of which is obtained when the reconciled reinterview is assumed perfect.

Table 1
Estimated Error Rates in the Original CPS Estimates

Error Rate Parameter	Description		Estimated Value β_{1ij}			Estimated Standard Error
	Classified as	True Status	Our Method	P&S (1986)	Recon. Reint. Perfect	Our Method
β_{121}	Employed	Unemployed	0.0407	0.0378	0.0188	0.01892
β_{131}	NLF	Unemployed	0.1196	0.1146	0.0838	0.01463
β_{112}	Unemployed	Employed	0.0049	0.0054	0.0017	0.00124
β_{132}	NLF	Employed	0.0100	0.0172	0.0098	0.00154
β_{113}	Unemployed	NLF	0.0110	0.0064	0.0034	0.00155
β_{123}	Employed	NLF	0.0205	0.0116	0.0053	0.00247

Table 2
Estimated Error Rates in the Unreconciled Reinterview CPS Estimates

Error Rate Parameter	Description		Estimated Value Our Method β_{2ij}	Estimated Standard Error
	Classified as	True Status		
β_{221}	Employed	Unemployed	0.0333	0.01772
β_{231}	NLF	Unemployed	0.1128	0.01360
β_{212}	Unemployed	Employed	0.0057	0.00135
β_{232}	NLF	Employed	0.0145	0.00160
β_{213}	Unemployed	NLF	0.0157	0.00171
β_{223}	Employed	NLF	0.0248	0.00238

Table 3
Cross-tabulation of the Aggregated 1981-1990 Original/Reconciled Reinterview Responses
75% Reconciled CPS Reinterview Data

Survey Result	Reconciled Reinterview			
Original CPS	Unemployed	Employed	NLF	Total
Unemployed	15,868	372	480	16,720
Employed	332	213,987	744	215,063
NLF	1,481	2,123	138,077	141,681
Total	17,681	215,482	139,301	373,464

Table 4
Implications of the Error Rate Estimates

Year y	BLS Reported Unemployment Rate UE_y^{BLS}	Prob. Unemp. Given Classified Unemp.	Adjusted Estimate of BLS Reported Unemployment Rate AUE_y^{BLS}			Difference in Reported *vs.* Adjusted	Estimated Standard Error in Difference
			Our Method	Poterba and Summers (1986)	Reconciled Data (1981-1990) Perfect	Our Method	Our Method
1990	5.44%	.8135	5.27%	5.36%	5.63%	0.17%	.27%
1989	5.20%	.8052	4.99%	5.09%	5.37%	0.21%	.26%
1988	5.43%	.8113	5.25%	5.35%	5.62%	0.18%	.27%
1986	6.89%	.8503	6.97%	7.04%	7.22%	-0.08%	.33%
1985	7.09%	.8531	7.20%	7.27%	7.44%	-0.11%	.34%
1984	7.41%	.8581	7.56%	7.63%	7.79%	-0.15%	.36%
1983	9.47%	.8894	9.99%	10.00%	10.04%	-0.52%	.48%
1982	9.54%	.8902	10.08%	10.09%	10.12%	-0.54%	.49%
1981	7.50%	.8581	7.66%	7.72%	7.88%	-0.16%	.36%

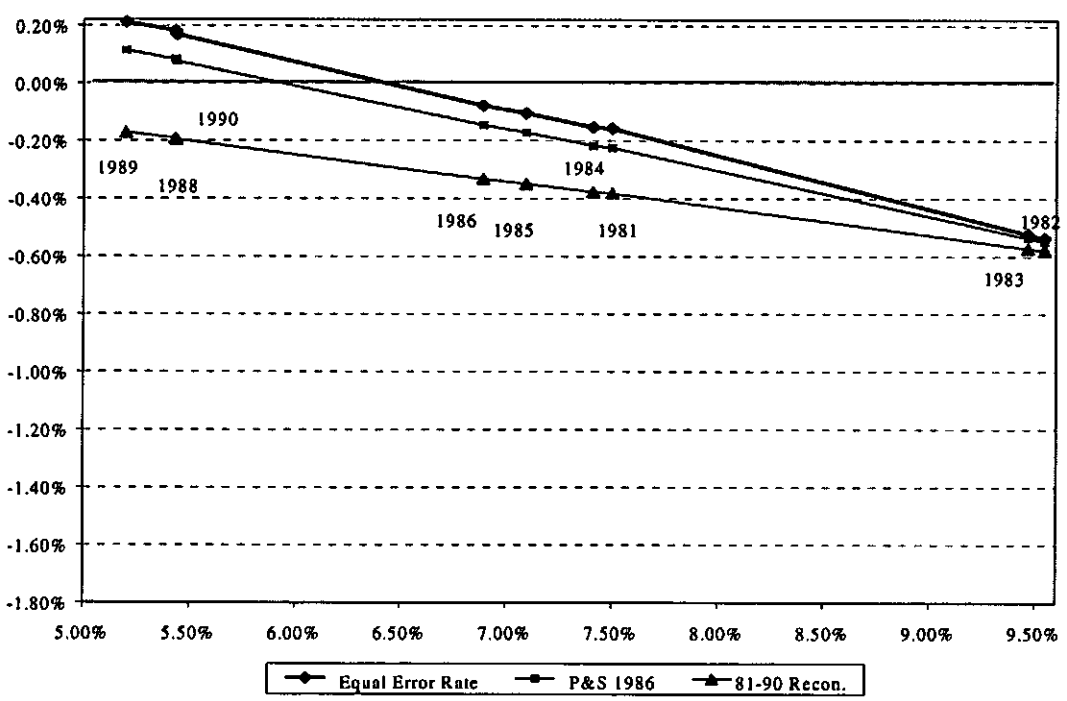

Figure 1. A Comparison of the Bias in the Reported Unemployment Rates as Computed Using Three Methods

In the screening test literature (Gastwirth 1987), the fraction of positive classifications which are correct, called the predictive value of a positive test, is known to vary directly with the prevalence of the characteristic. This is why, quite accurate diagnostic tests can have unacceptably high misclassification rates when populations with a low prevalence of a disease, are screened with them. The analog of this measure in our context, is the proportion of individuals classified as unemployed who are truly unemployed. This proportion is given in the third column of Table 4. Even though the range of reported unemployment rates is fairly narrow, a similar relationship with the unemployment rate can be seen.

While the results of the likelihood ratio test indicated, that the error rates were constant throughout the period, the referees suggested a further analysis to explore this assumption. We divided each of the nine survey years into three groups, according to the year's reported unemployment rate. Survey years, 1990, 1989 and 1988 were classified as having low unemployment, with reported rates from 5.20% to 5.44%. Similarly, survey years 1982 and 1983 were classified as having high unemployment, with reported rates of 9.54% and 9.47%, respectively. The remaining years with rates ranging from 6.89% to 7.5%, were classified as having moderate unemployment rates. With this three group structure, we developed an alternative model that assumed that the error rates were constant within each of the three rate size groups, but allowed each of these groups to have different error rates. The estimated error rates for the original interview are presented in Table 5. The error rates from Table 1, using the equal error rate model, are presented for comparative purposes.

We conducted a likelihood ratio test, to test the assumption that each of the error rates was the same, within each of these three groups, in comparison to the initial nine year model. The likelihood ratio statistic, $-2 \log \lambda$ with 72 degrees of freedom (144 parameters in the full model less 72 parameters in the three-group model), yielded a value of 69.25 with a p-value of 0.5697.

In general, the error rate estimates for the three unemployment rate classes, appear to be similar. Because the standard errors of the estimated error rates are quite large, a formal homogeneity test would have insufficient power to detect any variation in an error rate over the three periods.

To assess the sensitivity of the adjusted unemployment rate estimates in Table 4, we recomputed them using the error rates from the three-group model. The results are given in Table 6, which also provides the standard error of the unemployment rate estimates, ranging from a low of about 1.4% to a high of about 2.6%.

Figure 2 presents a graph of the bias in the unemployment using the three group model, and for comparison, the original equal error rate model. The results in Figure 2 are quite interesting. While the cyclical effect is still apparent, the estimated bias is shifted downward and shows a consistent negative bias throughout the business cycle.

Table 5
Error Rates in the Original CPS Data Estimated for Three Unemployment Rate Classes

Error Rate Parameter	Description		Error Rate Estimates							
			Model in Table 1 Assumes Constant Error Rates Across Years		Estimates Using Three Group Model					
	Classified as	True Status			Low Years 1990,1989, & 1988		Moderate Years 1981, 1984-1986		High Years 1982, 1983	
			Est.	STE	Est.	STE	Est.	STE	Est.	STE
β_{121}	Employed	Unemployed	0.0407	0.0189	0.0635	0.1061	0.1113	0.1258	0.0974	0.0717
β_{131}	NLF	Unemployed	0.1196	0.0146	0.1680	0.0538	0.1000	0.0246	0.1084	0.0221
β_{112}	Unemployed	Employed	0.0049	0.0012	0.0000	0.0047	0.0000	0.0098	0.0000	0.0069
β_{132}	NLF	Employed	0.0100	0.0015	0.0080	0.0038	0.0096	0.0025	0.0096	0.0031
β_{113}	Unemployed	NLF	0.0110	0.0015	0.0096	0.0040	0.0109	0.0024	0.0103	0.0029
β_{123}	Employed	NLF	0.0205	0.0025	0.0187	0.0065	0.0202	0.0034	0.0227	0.0044

Table 6
Implications of the Error Rate Estimates Using Three Group Model

Year *y*	BLS Reported Unemployment Rate	Prob Unemp. Given Classified Unemp. Three Group Model	Adjusted Estimate of BLS Reported Unemployment Rate		Difference in Reported *vs.* Adjusted		
			Original Equal Error Rate Model	Three Group Model	Original Equal Error Rate Model	Three Group Model	Estimate Standard Error of the Difference Three Group Method
1990	5.44%	0.9124	5.27%	6.43%	0.17%	-0.99%	1.40%
1989	5.20%	0.9088	4.99%	6.12%	0.21%	-0.93%	1.35%
1988	5.43%	0.9105	5.25%	6.41%	0.18%	-0.98%	1.41%
1986	6.89%	0.9170	6.97%	8.01%	-0.08%	-1.12%	2.35%
1985	7.09%	0.9178	7.20%	8.25%	-0.11%	-1.16%	2.42%
1984	7.41%	0.9199	7.56%	8.64%	-0.15%	-1.23%	2.53%
1983	9.47%	0.9400	9.99%	11.18%	-0.52%	-1.71%	2.05%
1982	9.54%	0.9404	10.08%	11.27%	-0.54%	-1.73%	2.08%
1981	7.50%	0.9191	7.66%	8.74%	-0.16%	-1.24%	2.56%

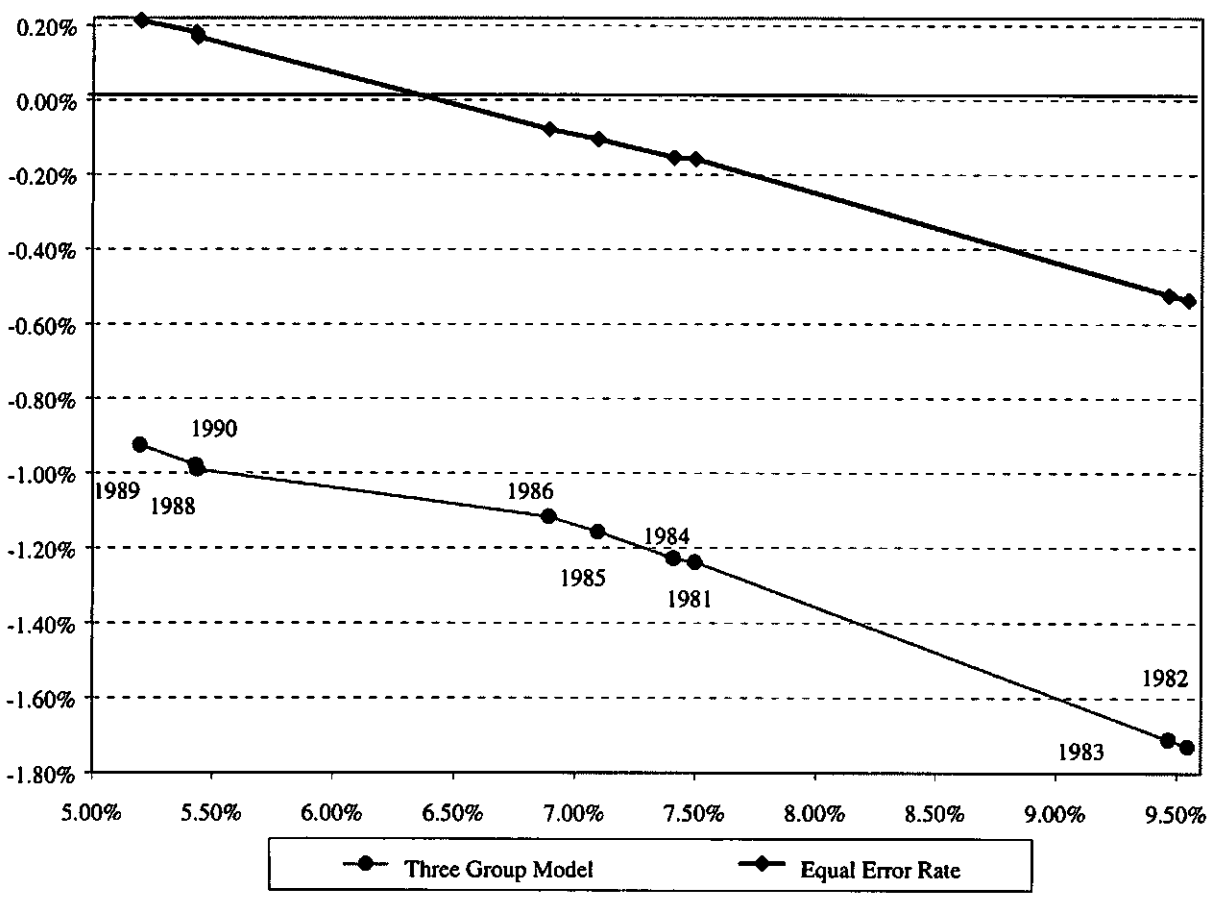

Figure 2. A Comparison of the Bias in the Reported Unemployment Rates as Computed Using the Equal Error Rate Model and the Three Group Model

5. IMPLICATIONS OF THE ADJUSTED ESTIMATES

The results in Figure 1 and 2 show that, all methods for adjusting the unemployment rate for misclassification error, indicate that the degree of bias in the reported rate varies over the business cycle. Given the differences in the estimated bias yielded by the two approaches, it is difficult to determine the magnitude of the bias. Unfortunately, the estimates are sensitive to the model specification, due to the small unreconciled reinterview sample size. This is reflected in the large standard errors of the estimated error rates, and consequently, the estimated bias.

Our approach using the assumption that the error rates remained constant throughout, suggests that bias in the survey estimates is small in years when the unemployment rate is between 5.5% and 7.5%. With this model, the reported unemployment rate appears to be unbiased when the true unemployment rate is around 6.3%, and yields an underestimate when the true rate is above this level, and an overestimate when the true rate is below it. The underestimation bias becomes quite noticeable when unemployment reaches 9%, while the overestimation bias could be meaningful when unemployment is less than 5%.

Using the three-group model results, implies that the reported unemployment rates are underestimates. If the finding is accurate, these results show that the bias in low unemployment years is still about -0.7%, but can be as high as -1.7% in high unemployment years. This contrasts the results obtained from the equal error rate model.

The fact that both the magnitude and direction of the bias in the reported unemployment rate change over the business cycle, may affect the use of that rate in studies of the "natural rate" of unemployment, and the trade-off between inflation and unemployment. Specifically, our results indicate that the range of the true unemployment rate over the business cycle, is larger than the range of the reported rate (see Table 4). Hughes and Perlman (1984) survey the literature on the "natural rate" of unemployment, and the trade-off between inflation and unemployment, as well as the role of search theory in explaining why unemployment is not that low at "full" employment. McKenna (1985) provides a more advanced treatment of job search theory, and its relationship to the duration of unemployment, and the degree to which unemployment is voluntary. Resolving the issue of which model underlies the misclassification error rates in the CPS survey, has important economic implications. If the equal error rate model were correct, in periods of low unemployment, the reported rate would be a slight overestimate. Hence, there would be less true unemployment to explain, by job search and related theories. On the other hand, if the three group model is the correct one, then even at low levels of reported unemployment, there are more persons really unemployed.

6. DISCUSSION

In this paper, we have presented an alternative method for estimating the error rates in the CPS survey. Our study differs from prior work, as we follow the Hui and Walter (1980) approach to estimate the error rates, by assuming that males and females will have the same error rates, and that the errors in the original survey are independent of those in the unreconciled reinterview. While the errors could be slightly correlated, the assumption of independence is standard in data analysis of this type, (see Bailar 1968, Chua and Fuller 1987, and Singh and Rao 1995). A discussion of the bias in the H&W method with dependent errors is given in Vacek (1985). As for the equal error rate assumption, several of the authors cited in this paper (*e.g.*, Poterba and Summers 1986), have noted minor to moderate differences in the error rates between males and females, under the assumption that the reconciled reinterview is perfect. However, this assumption has been questioned. For example, consider the estimate of β_{121}, the probability that an unemployed person, will be classified in the original survey as employed. From Table 3, we estimate this value under the assumption that the reconciled reinterview is unbiased, by dividing n_{21}, divided by $n_{.1}$ (332/17,681 = 0.0188), where n_{ij} is defined previously, with j now corresponding to the classification status in the reconciled reinterview. Using the expected value of these two frequencies from section 2, we can write an expression for the expectation of the estimate in large samples as follows:

$$
\begin{aligned}
&E(n_{21}/n_{.1}) \\
&= \frac{\pi_1\beta_{121}(1-\beta_{221}-\beta_{231})+\pi_2(1-\beta_{112}-\beta_{132})\beta_{212}+(1-\pi_1-\pi_2)\beta_{123}\beta_{213}}{\pi_1(1-\beta_{221}-\beta_{231})+\pi_2\beta_{212}+(1-\pi_1-\pi_2)\beta_{213}} \\
&= \beta_{121} + \beta_{121}\left[\frac{\pi_1(1-\beta_{221}-\beta_{231})}{\pi_1(1-\beta_{221}-\beta_{231}) + \pi_2\beta_{212} + (1-\pi_1-\pi_2)\beta_{213}} - 1\right] \\
&+ \left[\frac{\pi_2(1-\beta_{112}-\beta_{132})\beta_{212}+(1-\pi_1-\pi_2)\beta_{123}\beta_{213}}{\pi_1(1-\beta_{221}-\beta_{231})+\pi_2\beta_{212}+(1-\pi_1-\pi_2)\beta_{213}}\right]. \qquad (1)
\end{aligned}
$$

From (1) it follows that, if the reconciled reinterview error rates, β_{2ij} are equal to zero, that this estimator is unbiased. However, if the reconciled reinterview is not perfect, then the bias in the estimator depends on the prevalence rates in the population studied. As a result, if the actual original survey error rates are in fact equal in the two subpopulations studied, and the reconciled survey classifications are not perfect, the estimated original survey error rates for the two populations will differ. Therefore, one cannot use the similarities or differences in the estimated error rates for males and females from earlier

papers, to justify or to contradict the assumptions used here.

We have also conducted a sensitivity analysis of the Hui and Walter (1980) method for dichotomous responses (Sinclair 1994), that indicates that the procedure is sensitive to a violation in the equal error rate assumption, in some circumstances, but the procedure is quite robust in others. Further research is needed to develop reinterview procedures and analytical techniques, to relax the restrictive assumptions currently required in the analysis of the reinterview data.

It should be noted that Chua and Fuller (1987) also obtained estimates of the 3-outcome classification errors in the 1977-1980 CPS 25% sample reinterview data. Analogous to our results, their study found that the largest error rates were associated with classifying the truly unemployed. Poterba and Summers (1995) and Singh and Rao (1995) also found this group to be the hardest to classify. Because all models examined, indicated that the overall misclassification rate of an unemployed individual is around 20%, future reinterviews might focus on understanding why these rates are so high. Hopefully, this will lead to an improved survey.

A potential use of the "adjusted" estimates in Table 4, is in a sensitivity analysis of the literature (*e.g.*, Abowd and Zellner 1985; Poterba and Summers 1995) on gross flows, and labour market dynamics, which assumed that the reconciled interview was perfect. This is equivalent to their adoption of the estimates in the next to the last column of Table 3. Similarly, estimates of the classification errors may be incorporated in procedures, for estimating probit and logit models with misclassified response variables (Hausman and Morton 1994), and in the development of formal statistical procedures for survey data (Rao and Thomas 1991). It should be emphasized, that all the estimates adjusting for misclassification, are still in the research phase, and that the error rates are not yet estimated with sufficient accuracy, to adjust the regular survey data, especially as a new questionnaire and new interviewing procedures were introduced as of January 1994 (Bureau of Labour Statistics 1993).

ACKNOWLEDGEMENTS

The authors wish to thank Irv Schreiner of the U.S. Bureau of the Census for the data used in this research, and helpful discussion. We also wish to thank John Thompson, Henry Woltman and Jon Clark of the U.S. Bureau of the Census, and the referees, and the Associate Editor, for their many helpful comments. This research was supported in part by a grant from the National Science Foundation.

TECHNICAL APPENDIX A

A Review of the Hui and Walter Method

The Hui and Walter method was developed for the evaluation of diagnostic tests. The advantage of the technique is that, it allows the researcher to measure the error rate in a given test, without requiring the comparison test to be error-free. To accomplish this task, the procedure uses two populations (or subpopulations) with different prevalences, to estimate the parameters. The data from such a study, can be summarized in a 2×2 table as given in Figure A below. This Table for a specific subpopulation, is indexed by the letter g. We will denote the frequency of cases from subpopulation g, that have a classification from the first test, of status i ($i = 1$ for those having the trait, and $i = 2$ for those not having the trait), and from the second test of status j ($j = 1$ or 2), by n_{ij}. Let π denote the true unknown prevalence rate of the trait, and let α_r and β_r denote the unknown false positive and false negative rates. These error rates are indexed by the letter r, where $r = 1$ corresponds to the outcome from the first test, and $r = 2$ for the second test, (which, in our context, $r = 1$ corresponds to the original interview, and $r = 2$ to a reinterview). The false positive rate, α_r refers to the probability, that the evaluation from the r-th test, will classify the person as positive when in truth the person should have been classified as negative. Similarly, the false negative rate, β_r, is the probability that evaluation from the r-th test will classify the case as negative, when the case has the trait. One (1) minus each of these parameters, reflects to the specificity and sensitivity of the test (or survey) classification procedures, respectively.

Test 1 Outcome (Original Survey)	Test 2 Outcome (Reinterview)		
	Positive	Negative	Total
Positive	Cell 1	Cell 3	$n_{1.}$
Negative	Cell 2	Cell 4	$n_{2.}$
Total	$n_{.1}$	$n_{.2}$	$n_{..}$

Figure A. Cross-classification of Test 1 and Test 2 Outcomes

Assuming the errors of the first and second tests are independent of each other (given the true state), the expected probabilities, denoted by P_{ij} associated with the cell frequencies given in Figure A, for a given subpopulation g are as follows:

For

$$\begin{aligned}
&\text{Cell 1} \quad P_{g11} = \pi_g(1-\beta_{1,g})(1-\beta_{2,g}) + (1-\pi_g)(\alpha_{1,g}\alpha_{2,g}) \\
&\text{Cell 2} \quad P_{g21} = \pi_g(\beta_{1,g})(1-\beta_{2,g}) + (1-\pi_g)(1-\alpha_{1,g})(\alpha_{2,g}) \\
&\text{Cell 3} \quad P_{g12} = \pi_g(1-\beta_{1,g})\,\beta_{2,g} + (1-\pi_g)(\alpha_{1,g})(1-\alpha_{2,g}) \\
&\text{Cell 4} \quad P_{g22} = \pi_g(\beta_{1,g}\beta_{2,g}) + (1-\pi_g)(1-\alpha_{1,g})(1-\alpha_{2,g}).
\end{aligned} \tag{A.1}$$

From (A.1), we observe that we have a total of five parameters, but only three independent cell entries (or degrees of freedom), from which to estimate them. Therefore, the number of parameters must be reduced.

To reduce the parameters, Hui and Walter first, assume that, the proportion of cases with the trait, differs by subpopulation, which implies that, $\pi_1 \neq \pi_2$. Secondly, they require that two subpopulations can be found, such that the error rates for each test are the same for both subpopulations. The error rates associated with the two tests are allowed to differ. For two subpopulations, this implies that in (A.1), $\beta_r = \beta_{r,1} = \beta_{r,2}$, and $\alpha_r = \alpha_{r,1} = \alpha_{r,2}$, with $\beta_1 \neq \beta_2$, and $\alpha_1 \neq \alpha_2$. Under these conditions, the number of parameters reduces to six, (two prevalence rates, one for each subpopulation, and two error rates each for test 1 and test 2). Given that the two 2×2 tables contain six degrees of freedom, estimation is possible. Notice that if $\pi_1 = \pi_2$, and the error rates were the same in both subpopulations, then the probabilities in (A.1) would be the same for both subpopulations, so we would really have one table, and estimation would not be possible. Weighted nonlinear least squares estimates under the Hui and Walter model, can be computed using the Gauss Newton algorithm from the SAS Nonlinear Regression (NLIN) procedure. With this approach, one can express the observed frequencies, n_{ij}, in terms of the total sample size, $n_{..}$, multiplied by the probabilities in expression (A.1). Hui and Walter also present the closed formed estimators given in (A.2), expressed in terms of the observed cell probabilities denoted by p_{gij}.

$$\hat{\alpha}_r = \frac{(p_{r1\cdot}p_{\bar{r}\cdot 1} - p_{r\cdot 1}p_{\bar{r}1\cdot} + p_{211} - p_{111} + D)}{2E_r}$$

$$\hat{\beta}_r = \frac{(p_{r\cdot 2}p_{\bar{r}2\cdot} - p_{r2\cdot}p_{\bar{r}\cdot 2} + p_{122} - p_{222} + D)}{2E_r} \qquad \text{(A.2)}$$

where,

$\bar{r} = 2$ if $r = 1$, $\bar{r} = 1$ if $r = 2$

$$p_{g\cdot j} = \sum_{i=1}^{2} p_{gij}, \; p_{gi\cdot} = \sum_{j=1}^{2} p_{gij};$$

$$\hat{\pi}_g = \frac{1}{2} + \frac{[p_{g1\cdot}(p_{1\cdot 1} - p_{2\cdot 1}) + p_{g\cdot 1}(p_{1\cdot 1} - p_{21\cdot}) + p_{211} - p_{111}]}{2D}$$

where,

$$D = \pm\,[(p_{11\cdot}p_{21\cdot} - p_{1\cdot 1}p_{111} + p_{111} - p_{211})^2 - 4(p_{11\cdot} - p_{21\cdot})(p_{111}p_{2\cdot 1} - p_{211}p_{1\cdot 1})]^{\frac{1}{2}}$$

with,

$$E_1 = p_{2\cdot 1} - p_{1\cdot 1}, \; E_2 = p_{21\cdot} - p_{11\cdot}.$$

Note that two distinct points exist in the solution set, for either a positive or a negative value of D; however, only one of the values will yield reasonable estimates. Variances for the estimators, derived from the estimated asymptotic information matrix, are given in Hui and Walter's (1980) paper.

TECHNICAL APPENDIX B

Adjusting the Reported Unemployment Rates

To evaluate the implications of the estimated error rates, we needed an expression for estimating the actual prevalence rates (the four π parameters), in terms of the estimated error rates and the observed prevalence rates (or sample frequencies), from a given survey. In this section, we present the formula for these computations. With this expression, we can use the BLS reported unemployed and employed prevalence rates, as the observed values to estimate the adjusted BLS prevalence rates. Such an expression is given in (B.1).

Note that in expression (B.1), we have deleted the g-th subscript from the π parameters, so that the expression represents the prevalence rates among the general population, males and females combined. Note that, in this study, we have assumed that the estimated error rates are equal for males and females.

$$\begin{bmatrix} \hat{\pi}_{y1} \\ \hat{\pi}_{y2} \end{bmatrix} = \begin{bmatrix} 1-\hat{\beta}_{121}-\hat{\beta}_{131}-\hat{\beta}_{113} & \hat{\beta}_{112}-\hat{\beta}_{113} \\ \hat{\beta}_{121}-\hat{\beta}_{123} & 1-\hat{\beta}_{112}-\hat{\beta}_{132}-\hat{\beta}_{123} \end{bmatrix}^{-1} \begin{bmatrix} \dfrac{n_{y1.}}{n_{y..}} - \hat{\beta}_{113} \\ \dfrac{n_{y2.}}{n_{y..}} - \hat{\beta}_{123} \end{bmatrix}. \qquad \text{(B.1)}$$

In this paper, we have three sets of observed values. We have two observed prevalence rates from the reinterview sample (which is a sub-sample of the full CPS sample), including the unreconciled reinterview sample data, and the reconciled reinterview data, from the response-bias study sample, and BLS reported prevalence rates, as observed from the full CPS original survey. We will concentrate our efforts on the first and last of these three sets of statistics, the unreconciled reinterview sample data, and the published BLS estimates. To keep these two sets separate, we will define,

$$U_y^R = \frac{n_{y1.}}{n_{y..}}$$

$$E_y^R = \frac{n_{y2.}}{n_{y..}} \qquad \text{(B.2)}$$

as the observed unemployed and employed prevalence rates, obtained from the CPS unreconciled reinterview sample data. The corresponding BLS reported prevalence rates based on the full CPS original survey weighted data, are defined by U_y^{BLS} and E_y^{BLS}.

Similarly, the observed unemployment rate among those in the labour force, from the unreconciled reinterview sample data, is denoted by UE_y^R, equal to U_y^R divided by $(U_y^R + E_y^R)$, and the observed BLS reported unemployment rate, is defined as UE_y^{BLS}.

Simplifying expression (B.1) in terms of the observed reinterview prevalence rates, U_y^R and E_y^R we find:

$$\hat{\pi}_{y1} = \frac{\{U_y^R - \hat{\beta}_{113} - \hat{\beta}_{112}U_y^R + \hat{\beta}_{113}\hat{\beta}_{112} - \hat{\beta}_{113}\hat{\beta}_{132} - \hat{\beta}_{123}U_y^R - \hat{\beta}_{112}E_y^R + \hat{\beta}_{123}\hat{\beta}_{112} + \hat{\beta}_{113}E_y^R\}}{\{1 - \hat{\beta}_{112} - \hat{\beta}_{132} - \hat{\beta}_{123} - \hat{\beta}_{121}(1 + \hat{\beta}_{132} + \hat{\beta}_{123} + \hat{\beta}_{113}) - \hat{\beta}_{131}(\hat{\beta}_{112} + \hat{\beta}_{132} - 1) - \hat{\beta}_{113}(\hat{\beta}_{112} + \hat{\beta}_{132} - 1) + \hat{\beta}_{123}\hat{\beta}_{112}\}}$$

$$\hat{\pi}_{y2} = \frac{\{-\hat{\beta}_{121}U_y^R + \hat{\beta}_{121}\hat{\beta}_{113} + \hat{\beta}_{123}U_y^R + E_y^R - \hat{\beta}_{123} - \hat{\beta}_{121}E_y^R + \hat{\beta}_{122}\hat{\beta}_{123} - \hat{\beta}_{131}E_y^R + \hat{\beta}_{131}\hat{\beta}_{123} - \hat{\beta}_{123}E_y^R\}}{\{1 - \hat{\beta}_{112} - \hat{\beta}_{132} - \hat{\beta}_{123} - \hat{\beta}_{121}(1 + \hat{\beta}_{132} + \hat{\beta}_{123} + \hat{\beta}_{113}) - \hat{\beta}_{131}(\hat{\beta}_{112} + \hat{\beta}_{132} - 1) - \hat{\beta}_{113}(\hat{\beta}_{112} + \hat{\beta}_{132} - 1) + \hat{\beta}_{123}\hat{\beta}_{112}\}}. \quad \text{(B.3)}$$

Using expression (B.3), we can compute estimates of the adjusted unemployment rate among those in the labour force from the reinterview survey, denoted by AUE_y^R equal to $\hat{\pi}_{yg1}$ divided by $(\hat{\pi}_{yg1} + \hat{\pi}_{yg2})$. Note the AUE_y^R can be expressed as follows:

$$A\hat{U}E_y^R = \frac{\{-U_y^R + E_y^R + \hat{\beta}_{112}(U_y^R - \hat{\beta}_{113} + E_y^R) + \hat{\beta}_{132}(U_y^R - \hat{\beta}_{113}) + \hat{\beta}_{123}(U_y^R - \hat{\beta}_{112}) - \hat{\beta}_{113}E_y^R\}}{\{U_y^R + \hat{\beta}_{113}(1 + \hat{\beta}_{112} - \hat{\beta}_{121} - \hat{\beta}_{123}) + \hat{\beta}_{112}(U_y^R + E_y^R - \hat{\beta}_{113}) + \hat{\beta}_{121}(U_y^R + E_y^R - \hat{\beta}_{123}) - E_y^R + \hat{\beta}_{123} + \hat{\beta}_{131}(U_y^R - \hat{\beta}_{123})\}}. \quad \text{(B.4)}$$

Finally, to obtain the adjusted estimate of the BLS unemployment rate, denoted by, AUE_y^{BLS}, we substitute the values of U_y^{BLS} for U_y^R and E_y^{BLS} for E_y^R, into expression (B.4). Note that the estimated standard errors of the estimates for AUE_y^{BLS}, presented in section four, were computed using a Taylor series approximation method, (Wolter 1985). As a first step in this process, we assumed the variance in the published estimates of U_y^{BLS} and E_y^{BLS} were negligible. While this is not true, this assumption greatly simplifies the computation of the variances, and captures the majority of the total variation. This assumption is supported by the fact, that the size of the variance of these estimates, given the large full CPS yearly sample sizes is negligible in comparison to the sampling error associated with error rate estimates, which are based on the small unreconciled reinterview sample sizes. In summary, once the substitution of U_y^{BLS} for U_y^R, and E_y^{BLS} for E_y^R into expression (B.4) is completed, we assume that U_y^{BLS} and E_y^{BLS} are fixed known values in this equation. Finally, the sampling variance associated with the difference between the adjusted value and the published value, which defines the bias in the original estimate, is computed from the sum of the variances. Hence, by assuming the published value is sampling variance-free, the sampling variability associated with the difference or bias, is simply equal to the sampling variability associated with the adjusted value.

TECHNICAL APPENDIX C

Estimating Standard Errors of the Adjusted Unemployment Rates

For a complex function of several estimated parameters, the estimates of the variances associated with this function, can be computed using a Taylor series approximation as discussed by Wolter (1985). Suppose that the population parameter of interest is $Y = G(\Theta)$. Where Θ represents a n dimensional vector of population parameters, $\Theta = \{\theta_1, ..., \theta_n\}$. If G possesses continuous second derivatives, in an admissible range for Θ and Θ-hat, then Wolter (1985) presents the relationship:

$$\hat{Y} - Y = A + R(\hat{\Theta}, \Theta)$$

where,

$$A = \sum_{k=1}^{n} \frac{\partial G(\Theta)}{\partial \theta_k}(\hat{\theta}_k - \theta_k)$$

$$R(\hat{\Theta}, \Theta) = \sum_{k=1}^{n}\sum_{i=1}^{n} (1/2!)\frac{\partial^2 G(\Lambda)}{\partial \theta_k \partial \theta_i}(\hat{\theta}_k - \theta_k)(\hat{\theta}_i - \theta_i)$$

$$\hat{\Theta} \le \Lambda \le \Theta. \quad \text{(C.1)}$$

The remainder term is often regarded of little consequence, and is eliminated from the relationship. Given the first order approximation, Wolter (1985) presents,

$$\begin{aligned} \text{MSE}(\hat{Y}) &= E[G(\hat{\Theta}) - G(\Theta)]^2 \\ &= \text{Var}(A) \\ &= \sum_{k=1}^{n}\sum_{i=1}^{n} \frac{\partial G(\Theta)}{\partial \theta_k}\frac{\partial G(\Theta)}{\partial \theta_i} \text{Cov}(\hat{\theta}_k, \hat{\theta}_i) \\ &= d\, \Sigma_{\hat{\Theta}}\, d^T \end{aligned} \quad \text{(C.2)}$$

where d is a row vector of dimension n with the elements,

$$d_k = \left[\frac{\partial G(\ominus)}{\partial \theta_k}\right]. \qquad \text{(C.3)}$$

Wolter calls this estimator, the first order approximation to the mean square error (equal to the sampling variance + the bias of the estimator squared). Higher order approximations can be developed, by retaining additional terms in the expansion. For purposes of variance estimation, we substitute the estimated covariance matrix for $\sum_\theta$, and evaluate d at the estimated values of θ. Specifically, in our problem, we wish to estimate the variance associated with the function of the estimates in expression (C.4), given below.

$$G(\ominus)=G(\hat{\beta}_{121},\hat{\beta}_{131},\hat{\beta}_{112},\hat{\beta}_{132},\hat{\beta}_{113},\hat{\beta}_{123},U_y^{BLS},E_y^{BLS})=$$

$$\frac{\left\{-U_y^{BLS}+E_y^{BLS}+\hat{\beta}_{112}(U_y^{BLS}-\hat{\beta}_{113}+E_y^{BLS}) +\hat{\beta}_{132}(U_y^{BLS}-\hat{\beta}_{113})+\hat{\beta}_{123}(U_y^{BLS}-\hat{\beta}_{112})-\hat{\beta}_{113}E_y^{BLS}\right\}}{\left\{U_y^{BLS}+\hat{\beta}_{113}(1+\hat{\beta}_{112}-\hat{\beta}_{121}-\hat{\beta}_{123})+\hat{\beta}_{112}(U_y^{BLS}+E_y^{BLS}-\hat{\beta}_{113}) +\hat{\beta}_{121}(U_y^{BLS}+E_y^{BLS}-\hat{\beta}_{123})-E_y^{BLS}+\hat{\beta}_{123}+\hat{\beta}_{131}(U_y^{BLS}-\hat{\beta}_{123})\right\}} \qquad \text{(C.4)}$$

To create the estimates, we have assumed that the values of U_y^{BLS} and E_y^{BLS} are fixed (*i.e.*, have a negligible sampling variance). Taking the partial derivatives of equation (C.4) with respect to the six error rates, and evaluating these expressions at the estimated values of the error rates, yield a vector d which depends on the values of the error rate estimates and the published BLS unemployed and employed proportions for each year of the study. With our original model, that assumes the error rates are fixed across each year, this d vector for the period of study, only varies from year-to-year for the published values. For illustrative purposes the estimated vector d for 1989 using the BLS published unemployed and employed prevalence rates of .0347 and .6329 is equal to:

$$\hat{d} = \begin{bmatrix} \hat{\beta}_{121} & .07851 \\ \hat{\beta}_{131} & .07558 \\ \hat{\beta}_{112} & -1.2918 \\ \hat{\beta}_{132} & -.04813 \\ \hat{\beta}_{113} & -.64214 \\ \hat{\beta}_{123} & .03884 \end{bmatrix}.$$

The estimated covariance matrix from our SAS NLIN analysis, which, based on the original model that assumes the error rates are fixed by year, and as such, is the same for all years under study, is given below.

$\sum$	β121	β131	β112	β132	β113	β123
β_{121}	0.000358	-4.7E-05	-3.5E-07	-2.6E-08	-3.9E-07	2.9E-07
β_{131}	-4.7E-05	0.000214	-1.7E-07	-5.2E-07	-1.4E-06	-2.8E-07
β_{112}	-3.5E-07	-1.7E-07	1.54E-06	2.14E-07	-2.3E-08	9.9E-10
β_{132}	-2.6E-08	-5.2E-07	2.14E-07	2.37E-06	-1.5E-08	-6.1E-08
β_{113}	-3.9E-07	-1.4E-06	-2.3E-08	-1.5E-08	2.4E-06	-8E-08
β_{123}	2.9E-07	-2.8E-07	9.9E-10	-6.1E-08	-8.0E-08	6.1E-06

Pre and post multiplying the vector d, by the estimated covariance matrix, yields an estimated variance for AUE^{BLS} for 1989 of 6.72 E-6 and a standard error of the estimate equal to .0026 (.26%) as given in Table 4.

REFERENCES

ABOWD, J., and ZELLNER, A. (1985). Estimated gross labor force flows. *Journal of Economic and Business Statistics*, 3, 253-283.

BAILAR, B.A. (1968). Recent research in reinterview procedures. *Journal of the American Statistical Association*, 63, 41-63.

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

BUREAU OF LABOR STATISTICS (1993). Overhauling the population survey. *Monthly Labor Review*, 116, 9. Washington DC: U.S. Government Printing Office.

BUREAU OF LABOR STATISTICS (1992). *Employment and Earnings*. 38, 8. Washington DC: U.S. Government Printing Office.

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

FORSMAN, G., and SCHREINER, I. (1991). The design and analysis of reinterview: an overview. In *Measurement Errors in Surveys*, (Eds. Paul Biemer *et al.*). New York: John Wiley and Sons.

GASTWIRTH, J.L. (1987). The statistical precision of medical screening procedures: Application to polygraph and AIDs antibodies test data. *Statistical Science*, 2, 213-238.

HAUSMAN, J.A., and MORTON, S. (1994). Misclassification of a Dependent Variables in a Discrete Response Setting. Working paper, Department of Economics, MIT, Cambridge.

HUI, S.L., and WALTER, S.D. (1980). Estimating the error rates of diagnostic tests. *Biometrics*, 36, 167-171.

HUGHES, J.J., and PERLMAN, R. (1984). *The Economics of Unemployment*. New York: Cambridge University Press.

LEVY, P., and LEMESHOW, S. (1980). *Sampling for Health Professionals*. California: Lifetime Learning Publications.

McKENNA, C.J. (1985). *Uncertainty and the Labor Market: Recent Developments in Job Search Theory.* New York: St. Martins Press.

POTERBA, J.M., and SUMMERS, L.H. (1986). Reporting errors and labor market dynamics. *Econometrica*, 54, 1319-1338.

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

RAO, J.N.K., and THOMAS, D.R. (1991). Chi-squared tests with complex survey data subject to misclassification error. In *Measurement Errors in Surveys*, (Eds. Paul Biemer *et al.*). New York: Wiley.

SCHREINER, I., (1980). Reinterview Results From the CPS Independent Reconciliation Experiment (second quarter 1978 through third quarter 1979). Unpublished U.S. Bureau of the Census memorandum, May 7, 1980.

SINCLAIR, M.D. (1994). Evaluating Reinterview Survey Methods for Measuring Response Errors. Phd. Dissertation, George Washington University, September.

SINCLAIR M.D., and GASTWIRTH, J.L. (1996). On procedures for evaluating the effectiveness of reinterview survey methods: application to labor force data. *Journal of the American Statistical Associatio*n, 91, (435) 961-969.

SINGH, A.C., and RAO, J.N.K (1995). On the adjustment of gross flow estimate for classification error with application to data from the Canadian Labour Force Survey. *Journal of the American Statistical Association*, 90, 478-488.

U.S. BUREAU OF THE CENSUS (1963). *The Current Population Survey Reinterview Program: Some Notes and Discussion.* Technical Paper No. 6. Washington, D.C.: U.S. Government Printing Office.

U.S. BUREAU OF THE CENSUS (1985). *Evaluating Censuses of Population and Housing.* Statistical Training Document #ISP-TR-5. Washington, D.C.: U.S. Government Printing Office, 1985.

VACEK, P.M. (1985). The effect of conditional dependence on the evaluation of diagnostic tests. *Biometrics*, 41, 959-968.

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

Survey Methodology, December 1998
Vol. 24, No. 2, pp. 171-183
Statistics Canada

Use of Statistical Matching Techniques in Calibration Estimation

ROBBERT H. RENSSEN[1]

ABSTRACT

This article deals with an attempt to cross-tabulate two categorical variables, which were separately collected from two large independent samples, and jointly collected from one small sample. It was assumed that the large samples have a large set of common variables. The proposed estimation technique can be considered a mix between calibration techniques and statistical matching. Through calibration techniques, it is possible to incorporate the complex designs of the samples in the estimation procedure, to fulfill some consistency requirements between estimates from various sources, and to obtain fairly unbiased estimates for the two-way table. Through the statistical matching techniques, it is possible to incorporate a relatively large set of common variables in the calibration estimation, by means of which the precision of the estimated two-way table can be improved. The estimation technique enables us to gain insight into the bias generally obtained, in estimating the two-way table, by sole use of the large samples. It is shown how the estimation technique can be useful to impute values of the one large sample (donor source) into the other large sample (host source). Although the technique is principally developed for categorical variables Y and Z, with a minor modification, it is also applicable for continuous variables Y and Z.

KEY WORDS: Consistency between estimates; General regression estimator; Imputation; Multivariate auxiliary information; Two-way table.

1. INTRODUCTION

Most statistical surveys are conducted to obtain estimates of simple descriptive finite population parameters. The estimates are often presented in tabular form, with cells containing estimates of population totals or subgroup totals. Often, data are collected on an extensive set of variables, producing numerous results for these variables and their relationships. In order to save resources and decrease response burden, statistical bureaus wish to reduce sample sizes and shorten questionnaires. They resort to administrative data sources and existing large-scale sample surveys, or applying splitting questionnaire survey designs (see Raghunathan and Grizzle 1995). As a consequence, methods for combining distinct data sources have become a popular tool in the production of statistics. Combining data sources can be done in many different ways; two well-known techniques in survey sampling are statistical matching and calibration estimation.

Singh, Mantel, Kinack and Rowe (1993) describe statistical matching as a special case of imputation in which there are two distinct micro-data sources containing different information on different units. One data source serves as a host or recipient file to which new information is imputed for each record, using data from the other source, which is the donor file. More specifically, they consider a host file A, containing information on variables (X, Y) and a donor file B containing information on variables (X, Z). The common variable X can be used to identify similar units in the two files. In general, statistical matching deals with the problem of completing the records in file A, by imputing values for Z using the information on the (X, Z) relationship in file B. These imputed Z-values suffer from a serious limitation in that, the real relationship between Y and Z may be completely lost in the enriched host file. This limitation amounts to the so-called assumption of conditional independence between Y and Z given X. In order to get rid of this conditional independence assumption, Singh *et al.* (1993) consider a third data set (file C) representing auxiliary information about the full set (X, Y, Z). For example, this data set could come from a small-scale specially conducted survey. They discuss several imputation methods to complete file A, by adding Z from file B using information from A, B, and C, on the joint relationships of X, Y, and Z. Singh *et al.* (1993) give many relevant references on statistical matching techniques. We only mention Rodgers (1984), Rubin (1986) and Paass (1986).

In Deville and Särndal (1992), calibration estimation is derived as a general technique to weight sample surveys, taking into account the complex design of the sample and auxiliary information obtained from external sources (see also Deville, Särndal, and Sautory 1993). The use of auxiliary information, *i.e.*, control variables, primarily aim at three goals: namely, reducing sampling variance, reducing bias due to non-response, and ensuring consistency between estimates from various sources with respect to the used control variables. There is an extensive body of literature on weighting methods in sample surveys. We refer to Bethlehem and Keller (1987), Alexander (1987), Lemaître and Dufour (1987), and Zieschang (1990).

[1] Robbert H. Renssen, Department of Statistical Methods, Statistics Netherlands, P.O. Box 4481, 6401 CZ Heerlen, Netherlands.

This article deals with the specific problem of how to estimate the cross-product between Y and Z (*e.g.*, the two-way table between Y and Z in case these variables are categorical or the covariance between Y and Z in case these variables are continuous), using statistical matching techniques as well as calibration estimation. We assume that two data files A and B represent two large-scale sample surveys, possibly both obtained by a complex design. In order to weight the specially conducted small sample (file C), auxiliary information is derived from these large samples. It might be difficult to judge whether the large samples should be considered as suppliers of auxiliary information for the small sample, or vice versa. Through the statistical matching, it is possible to incorporate a large set of X-variables in the estimation procedure, despite the sample size of the small sample. The use of calibration estimation makes it possible to take account of the complex design of all samples in the estimation procedure, and to fulfill some consistency requirements. Most of the article is devoted to categorical Y and Z, because of the specific properties of these variables. For example, it is shown that the marginal counts of the estimated YZ-table, always coincide with estimates for the population totals of Y and Z, when the ordinary calibration estimator is applied with the X-variables as control variables, on the first and second large sample respectively. Nevertheless, the proposed method is also applicable for continuous Y and Z. Throughout this article it will be assumed that X may consist of several variables, which may be categorical and/or continuous. It is argued that when the X-variables are highly correlated with either Y or Z, then our estimation method gives relatively precise estimates for the cross-product between Y and Z, *e.g.*, for the complete YZ-table when Y and Z are categorical.

The proposed estimation procedure closely resembles a method presented in Singh *et al.* (1993, Section 2) to estimate a correlation coefficient between Y and Z. These variables are assumed to be univariate in this article. Our method, however, differs from theirs in that it incorporates the complex designs of all data sources in the estimation procedure and that it uses the large data sources more efficiently in estimating population parameters from the small data source. When Y and Z are categorical, and there is no linear correlation between X and Y as well as between X and Z, then our method corresponds to incomplete post-stratification (Deville and Särndal 1992, Bethlehem and Keller 1987). On the other hand, if Y is perfectly correlated with X, then our method gives an estimated two-way table between Y and Z which corresponds to an estimated two-way table that would have been obtained from file B if first the Y-values were imputed. A similar result holds if Z and X are perfectly correlated.

Although combining distinct data sources across common variables may be fruitful from a theoretical point of view, in practice, complications may arise because common variables in the strict sense are not easily found, mainly due to discrepancies between definitions, methods of observation, and reference period. These complications may be reduced if the survey processes involved, are harmonized at an early stage. A promising application of the use of common variables, lies in integrated survey designs, such as the Dutch Household Survey on Living Conditions, see van Tuinen (1995), Bakker and Winkels (1998), Winkels and Everaers (1998), and Hofmans (1998). The questionnaire design of this survey has a three-shell structure. The first shell contains questions on demographic and socioeconomic issues, and level of education. The second shell contains a few easy to answer core questions, on every relevant aspect of living conditions. The questions in the third shell also concern living conditions, but they are more exhaustive than the questions in the second shell. In order to shorten the time it takes to answer, the third shell questionnaire is split. Each respondent has to fill in the complete questionnaire of the first and second shell and one sub-questionnaire of the third shell. On account of the third shell, the sample is split into sub- samples associated with each sub-questionnaire. The sampling design of each sub-sample can be described as two-phase sampling for the general regression estimator.

The organization of this article is as follows. The theoretical framework is developed in Section 2. For this purpose it is convenient to discuss a calibration estimator for the small sample, obtaining auxiliary information from two distinct registrations instead of two distinct large samples. One registration contains values on X and Y and the other registration on X and Z. Sections 2.1 to 2.4 deal with categorical Y- and Z-variables. In Section 2.1, the registrations are used to obtain a first synthetic estimate of the YZ-table by regression methods of imputation. It is shown that this synthetic two-way table has some interesting properties. In Section 2.2 we propose a set of calibration equations to weight the small sample, based on these properties. We briefly discuss its relationship to complete and incomplete post-stratification. A numerical illustration is given in Section 2.3. The linkage to statistical matching techniques as discussed in Singh *et al.* (1993) is given in Section 2.4. The treatment of categorical Y and Z is unnecessary and restrictive. In Section 2.5, it is shown that the proposed weighting technique is also applicable for continuous Y and Z or for continuous Y and categorical Z. In Section 3, the technique is modified, using auxiliary information from two distinct large samples instead of two registrations. By means of a simulation study, the modified weighting method is compared to the traditional incomplete two-way stratification. Finally, Section 4 contains some concluding remarks.

2. COMBINING REGISTRATIONS ACROSS COMMON VARIABLES

Consider a finite population $\Omega = \{1, ..., N\}$ of N persons and suppose there are two registrations available of these

persons. The first registration contains of each person k, a record with scores y_k and x_k of the variables Y and X respectively, and the second registration of each person k, a record with scores z_k and x_k of the variables Z and X respectively, $k = 1, ..., N$. Obviously, the variable X is present in both registrations. We note that the records from both registrations correspond to the same finite population. The process of merging these registrations, would be like exact matching if X is used to compare the records in the one registration with those in the other registration, in an effort to determine which pairs of records relate to the same population unit (see Fellegi and Sunter 1969). In this article we will proceed differently.

2.1 Formulating the Synthetic Population Totals

Let Y denote education with p categories and Z denote employment with q categories. Then y_k is a vector of order p, representing p dummy variables. Each dummy variable corresponds to a specific category; it equals 1 if person k belongs to that category, otherwise it equals 0. Analogously defined, z_k is a vector of order q. Further, X may be the result of a complete or incomplete crossing (stratification) of a number of characteristics (*e.g.*, sex, age, region, marital status, *etc.*). The scores x_k are vector valued, of order r. In case X consists of a complete stratification, x_k represents r dummy variables. In the remaining of this article, r should be considered large in comparison with $p \times q$. The population totals for Y and Z are the marginal frequency distributions with respect to education and employment. Using the common variable X, predictions for Y and Z can be defined with a multiple linear regression model:

$$\hat{y}_k = B^t x_k, \; k = 1, ..., N,$$

and

$$\hat{z}_k = A^t x_k, \; k = 1, ..., N,$$

where B and A are the ordinary least squares regression coefficients satisfying the normal equations

$$\left(\sum_{k=1}^{N} x_k x_k^t \right) B = \sum_{k=1}^{N} x_k y_k^t \tag{1}$$

and

$$\left(\sum_{k=1}^{N} x_k x_k^t \right) A = \sum_{k=1}^{N} x_k z_k^t. \tag{2}$$

The superscript 't' denotes transposition. This model is called a linear probability model, (see Maddala 1983, chap. 2). There are more elegant models, such as probit and logit models, to predict binary variables. However, we are not interested in the predictions themselves, but in the synthetic population totals of these predictions. These totals appear to have nice properties if the linear prediction model is used, and for this reason the model can be justified. Note that B is calculated from the first registration and A from the second one. By means of the common variable X and the regression coefficients B and A, we construct a synthetic registration, which contains a record of each person k with scores x_k, $B^t x_k$, and $A^t x_k$. In fact, either y_k or z_k may be added to this registration, but for our purposes this addition appears to be superfluous (see next paragraph). If there exists a vector a of order r of fixed numbers such that $a^t x_k = 1$ for all k, then the population totals of the new variables $B^t x_k$ and $A^t x_k$ equal the population totals of the corresponding original variables (see *e.g.*, Bethlehem and Keller 1987). This can be shown easily by first premultiplying the normal equations (1) and (2) by a^t and subsequently substituting $a^t x_k = 1$ into the resulting equations.

From the synthetic registration, a synthetic two-way table can be defined by $\sum_{k=1}^{N} (B^t x_k)(A^t x_k)^t$. This synthetic two-way table can be considered as an approximation of the (simultaneous) frequency distribution $\sum_{k=1}^{N} y_k z_k^t$. Using the normal equations (1) and (2), the following identities can be derived:

$$\sum_{k=1}^{N} (B^t x_k)(A^t x_k)^t = \sum_{k=1}^{N} y_k (A^t x_k)^t$$

$$= \sum_{k=1}^{N} (B^t x_k) z_k^t.$$

Clearly, the crossings between $B^t x_k$ and $A^t x_k$, y_k and $A^t x_k$, or $B^t x_k$ and z_k, all result in identical synthetic two-way tables. Therefore, it suffices to consider only $\sum_{k=1}^{N} (B^t x_k)(A^t x_k)^t$, and delete either y_k or z_k in the synthetic registration. The difference between the real frequency distribution between Y and Z and its synthetic "approximation", can be obtained from the following decomposition

$$\sum_{k=1}^{N} y_k z_k^t = \sum_{k=1}^{N} (B^t x_k)(A^t x_k)^t + \sum_{k=1}^{N} (y_k - B^t x_k)(z_k - A^t x_k)^t. \tag{3}$$

Note the strong resemblance with the ordinary variance decomposition in regression analysis (see *e.g.*, Searle 1971). If either $B^t x_k = y_k$ or $A^t x_k = z_k$ for all k, then the two-way table derived from the synthetic registration, equals the real simultaneous frequency distribution between Y and Z.

Let l be a vector of appropriate order consisting of ones, and note that $l^t y_k = 1$ and $l^t z_k = 1$ for all k. If there exists a constant a such that $a^t x_k = 1$ for all k, then we also have

$$l'\hat{y}_k = l'B'x_k = l'\left(\sum_{k=1}^{N} y_k x_k'\right)\left(\sum_{k=1}^{N} x_k x_k'\right)^{-1} x_k =$$

$$a'\left(\sum_{k=1}^{N} x_k x_k'\right)\left(\sum_{k=1}^{N} x_k x_k'\right)^{-1} x_k = a'x_k = 1$$

for all k, and similarly $l'\hat{z}_k = l'A'x_k = 1$ for all k. It follows that

$$l'\sum_{k=1}^{N} (B'x_k)(A'x_k)' = \sum_{k=1}^{N} (A'x_k)' = \sum_{k=1}^{N} z_k' \qquad (4)$$

and

$$\sum_{k=1}^{N} (B'x_k)(A'x_k)'l = \sum_{k=1}^{N} (B'x_k) = \sum_{k=1}^{N} y_k. \qquad (5)$$

So, the row and column totals of the synthetic two-way table, equal the corresponding marginal population counts with respect to Y and Z.

What remains to consider, is the condition $a'x_k = 1$ for all k, for some constant a. This condition is satisfied if X represents a categorical variable. More generally, the condition is always satisfied if the vector X can be partitioned into two sub-vectors, one of which represents a categorical variable.

2.2 Formulating the Constraints in Calibration Estimation

Suppose a probability sample s of size n is drawn from the finite population $\Omega = \{1, ..., N\}$ according to a sampling design $p(s)$ such that the first and second order inclusion probabilities $\Pr(k \in s) = \pi_k$ and $\Pr(k, l \in s) = \pi_{kl}$ are strictly positive. For each $k \in s$ the vector of scores (x_k, y_k, z_k) is observed. Two distinct registrations are available to provide auxiliary information. The first registration contains for each $k \in \Omega$, records with scores on x_k and y_k, the second registration contains for each $k \in \Omega$, scores on x_k and z_k. The objective is to estimate the YZ- table from the sample s, using auxiliary information from both registrations. There exists a wide range of weighting type estimators in the presence of multivariate auxiliary information. In Särndal, Swensson and Wretman (1992), the general regression estimator is extensively discussed. It implicitly defines sample weights, which reproduce the known population totals of the auxiliary variables, used as control variables in the estimator. Such a consistency property is attractive if the auxiliary information is used both for publication and for weighting. As a generalization of the general regression estimator, the calibration estimator is developed (Deville and Särndal 1992 and Deville *et al.* 1993).

To be specific, let G be a real valued function as defined in Deville *et al.* (1993) and consider the following weighting type estimator for our YZ-table:

$$\hat{T} = \sum_{k=1}^{n} w_k (y_k z_k'), \qquad (6)$$

where w_k is a scalar, representing a weight assigned to person $k \in s$. Denote $d_k = \pi_k^{-1}$. A calibration estimator for the YZ-table uses weights which are obtained by minimizing $\sum_{k=1}^{n} d_k G(w_k/d_k)$ with respect to w_k subject to a set of constraints on w_k for any particular sample s. We first consider the following set of constraints:

$$\sum_{k=1}^{n} w_k y_k = \sum_{k=1}^{N} y_k \text{ and } \sum_{k=1}^{n} w_k z_k = \sum_{k=1}^{N} z_k. \qquad \text{(I)}$$

This (first) set of constraints only uses the (marginal) counts with respect to Y and Z. No use is made of the common variable X. One of the $p + q$ equations is redundant, so to solve the minimization problem, one equation can be deleted. For $G(w_k/d_k) = (w_k/d_k - 1)^2$, the resulting calibration estimator corresponds to incomplete two-way stratification as defined in Bethlehem and Keller (1987). By taking $G(w_k/d_k) = 1 + w_k/d_k(\log(w_k/d_k) - 1)$, the classical raking ratio estimator is obtained (see *e.g.*, Oh and Scheuren 1987). Copeland, Peitzmeier and Hoy (1987) have compared these methods, based on data of the Current Population Survey. They conclude that the estimates produced by the two methods are very similar. In Deville *et al.* (1993), two other distance functions are discussed, which are especially interesting in view of the problem of extreme weights. Estimating two-way tables with constraints on the marginal counts, is frequently performed in sample surveys. Often, the constraints on the marginal counts are required for two reasons. The first reason is to reduce sampling error and sampling bias, and the second reason is to meet consistency requirements with published population counts.

Suppose that x_k is categorical with r categories. Since population information about the crossings between Y and X, and the crossings between Z and X are available, we may also consider the following set of constraints:

$$\sum_{k=1}^{n} w_k (y_k x_k') = \sum_{k=1} y_k x_k' \quad \text{and}$$

$$\sum_{k=1}^{n} w_k (z_k x_k') = \sum_{k=1}^{N} z_k x_k'.$$

The number of non-redundant constraints in this set equals $r(p + q - 1)$. For large r, this set may be not feasible because it contains too many constraints in comparison with

the sample size. Only if r is small, the set may be of practical interest. In the remaining of this article, this set of constraints will be disregarded.

In view of incorporating a large set of common variables in the weighting procedure, we consider a set of constraints, which exploits the bivariate population information that we have in the synthetic table:

$$\sum_{k=1}^{n} w_k(B'x_k)(A'x_k)' = \sum_{k=1}^{N} (B'x_k)(A'x_k)'. \qquad \text{(II)}$$

This (second) set of constraints is a straightforward application of the theory of calibration estimators. Population totals of the crossing between $B'x_k$ and $A'x_k$ are known, so these crossings are taken as auxiliary variables to formulate the set of constraints. Evidently, for large r, the number of non-redundant constraints remains bounded by $p \times q$. A major disadvantage of the resulting calibration weights is that, they do not necessarily reproduce the (marginal) population counts with respect to Y and Z, when applying these weights to y_k and z_k respectively. In other words, the resulting calibration weights do not necessarily satisfy the first set of constraints. Especially, if this set of constraints is formulated in view of consistency requirements, this is a serious drawback.

Therefore, as an alternative, we consider a third set of constraints:

$$\sum_{k=1}^{n} w_k\left(y_k z_k' - (y_k - B'x_k)(z_k - A'x_k)'\right) = \sum_{k=1}^{N} (B'x_k)(A'x_k)' \qquad \text{(III)}$$

Assuming that there exists a constant a, such that $a'x_k = 1$ for all k, this set of constraints meets the consistency objective. Let l denote a vector of ones of appropriate order and recall that $l'y_k = l'B'x_k = l'z_k = l'A'x_k = 1$ for all k, $B'\sum_{k=1}^{N} x_k = \sum_{k=1}^{N} y_k$, and $A'\sum_{k=1}^{N} x_k = \sum_{k=1}^{N} z_k$. By pre-multiplying the third set of equations on both sides with l', we obtain the first set of constraints with respect to Z, and post-multiplying the third set on both sides with l gives the first set of constraints with respect to Y. The resulting calibration estimator can be expressed as

$$\hat{T} = \sum_{k=1}^{n} w_k(y_k z_k') = \sum_{k=1}^{N} (B'x_k)(A'x_k)' + \sum_{k=1}^{N} w_k(y_k - B'x_k)(z_k - A'x_k)'.$$

Clearly, this estimator obeys the decomposition given by (3). It equals the synthetically defined two-way table plus an adjustment term. This adjustment term is a calibration estimate for the difference between the real frequency distribution between Y and Z and the synthetically defined two-way table. Similarly to the second set of constraints, the number of non-redundant constraints in the third set is bounded by $p \times q$.

An important special case is $G(w_k/d_k) = (w_k/d_k - 1)^2$. Then each estimated cell is a general regression estimate with $(y_k z_k)$, $\mathrm{vec}(B'x_k x_k' A)$, and $\mathrm{vec}(y_k z_k' - (y_k - B'x_k)(z_k - A'x_k)')$ as control variables in case of the first, second, and third set of constraints respectively. Analytical formulas for the design variance of the general regression estimator, are given in *e.g.*, Särndal *et al.* (1992, chap. 6). In fact, these formulas are approximations for large sample sizes. In Deville and Särndal (1992), sufficient conditions are given under which these approximations are valid for calibration estimators in general.

In Deville *et al.* (1993), complete post-stratification is described as a calibration method for which all population counts with respect to the cross-classifications, are used in the set of constraints. An elaboration of complete post-stratification, results in the ordinary post-stratification estimator, regardless of the distance function G. As an alternative, incomplete post-stratification is described as a calibration method, in which less detailed than a complete knowledge of all cell counts, is used in the constraint set. The calibration estimator defined under the first set of constraints, is a commonly used example of incomplete post-stratification. Several cases are discussed, in which incomplete post-stratification is preferable to complete post-stratification. Two of them are, lack of population information and, some zero or extremely small cell counts (see also Oh and Scheuren 1987). The calibration estimator defined under the second and third set of constraints, corresponds to complete post-stratification in the sense that, all crossings are used as auxiliary information. Except when a perfect linear relationship exists either between Y and X, or between Z and X, the method differs from complete post-stratification in using synthetic population totals instead of real population counts. Complete post-stratification gives unstable results, if some sample cells have only few observations. In such situations, incomplete post-stratification is of practical interest. Similarly, the calibration estimator under the second and third set of constraints may be unstable. Analogously to incomplete post-stratification, one might consider using an incomplete crossing in the constraints instead.

2.3 A Numerical Illustration

We illustrate the calibration estimator under the three different sets of constraints by means of a hypothetical example. The example is based on real data from a sample on behalf of the Dutch National Travel Survey (1994). The sampling design is roughly a self-weighted cluster sample of addresses. All persons living in a selected address, are included in the sample. The net sample size is approximately 80,000 persons within 34,000 addresses. From this sample, two hypothetical registrations of approximately

N = 80,000 persons are constructed. In the one registration, age is registered (in six categories), and in the other registration, car ownership (in two categories). The common variable between the registrations is a key number for addresses, resulting in r = 34,000 categories for the X-variable. For this particular example the synthetic two-way table simplifies to

$$\sum_{k=1}^{N} (B^t x_k)(A^t x_k)^t = \sum_{j=1}^{r} N_j \bar{y}_j \bar{z}_j^t,$$

where N_j denotes the size of the j-th address, $\bar{y}_j$ the mean of the six age categories of the j-th address, and $\bar{z}_j$ the mean of the two car ownership categories of the j-th address.

In order to calculate the synthetic two-way table, both registrations are combined as follows. Firstly, they are sorted according to the key number for addresses. Secondly, the address counts of the six age categories and the two car ownership categories are calculated. Thirdly, each address count of age, is linked with its corresponding address count of car ownership. By means of this synthetic registration of r = 34,000 addresses, the synthetic two-way table can be calculated. The result is shown in Table 1. This table can be considered as a first approximation of the real frequency distribution between age and car ownership. A sufficient condition for a close approximation, is homogeneity with respect to either age or car ownership within all addresses, *i.e.*, all persons at the same address should either be in the same age category or in the same car ownership category. For most (multiple) person addresses, this seems to be an unlikely proposition. It follows from equations (4) and (5) that the row and column totals in table 1 coincide with the real (marginal) population counts of age and car ownership respectively.

By means of a simple random sample of n = 1000 persons, the population cell counts are estimated using a general regression estimator. Three sets of auxiliary variables are used, in accordance with the three sets of constraints mentioned in the previous section. The estimated tables are given below (for convenience we have taken the quadratic distance measure: $G(w_k/d_k) = (w_k/d_k - 1)^2$). The corresponding estimated standard deviations are within parenthesis. These estimates are based on the usual variance formulas of the general regression estimator, see Särndal *et al.* (1992, chap. 6).

Table 1
Synthetic Population Totals for Crossings Between Age and Car Ownership

	1	2	3	4	5	6	total
yes	3461	1659	5739	10770	6536	3334	31499
no	9827	4692	7902	17102	6424	5389	51336
total	13288	6351	13641	27872	12960	8723	82835

Table 2
Estimated Population Totals for Crossings Between Age and Car Ownership, Satisfying the First Set of Constraints

	1	2	3	4	5	6	total
yes	0 (0)	0 (0)	4968 (423)	15414 (543)	7518 (458)	3599 (375)	31499
no	13288 (0)	6351 (0)	8673 (423)	12458 (543)	5422 (458)	5124 (375)	51336
total	13288	6351	13641	27872	12960	8723	82835

Table 3
Estimated Population Totals for Crossings Between Age and Car Ownership, Satisfying the Second Set of Constraints

	1	2	3	4	5	6	total
yes	0 (0)	0 (0)	4791 (435)	13826 (811)	6887 (494)	3421 (321)	28923 (1005)
no	14385 (782)	7012 (595)	8118 (563)	12893 (796)	5853 (464)	5654 (306)	53912 (1005)
total	14385 (782)	7012 (595)	12908 (603)	26718 (958)	12739 (419)	9074 (177)	82835

Table 4
Estimated Population Totals for Crossings Between Age and Car Ownership, Satisfying the Third Set of Constraints

	1	2	3	4	5	6	total
yes	0 (0)	0 (0)	5501 (226)	15647 (227)	6898 (177)	3453 (78)	31499
no	13288 (0)	6351 (0)	8139 (226)	12224 (227)	6062 (177)	5270 (78)	51336
total	13288	6351	13641	27872	12960	8723	82835

In Table 2 the population counts are estimated according to the ordinary incomplete two-way stratification (Bethlehem and Keller 1987). There are no young people (age category 1 and 2) owning a car, observed in the sample, which is likely to be representative for the population, so these cells are estimated by zero. Due to the consistency requirements, *i.e.*, the first set of constraints, it follows that the estimated cell counts of young people without a car equal the corresponding marginal cell counts. An attempt to improve Table 2, is to use the common variable address in the weighting procedure. In Table 3, the cell estimates are given according to the second set of constraints. As already mentioned in the previous section, the estimated row and column totals may differ from the real population counts. A comparison between Table 2 and Table 3 shows that these differences can be considerable. In addition, almost all estimated cell counts in Table 2 have smaller estimated standard deviations than the corresponding estimated cell counts in Table 3. So, the second set of constraints gives quite unsatisfactory results. The third set of constraints covers the first set of constraints. This implies 1) consistency of the estimated marginal cell counts with respect to the corresponding known population cell counts, and 2) smaller asymptotic variances of all estimated cell counts. The results are shown in Table 4. Indeed, the estimated marginal cell counts are consistent, and the estimated standard deviations are at most half of the corresponding standard estimates given in Table 2.

2.4 Imputing Values of the one Registration into the Other Registration

Until now, we have developed a weighting method to estimate a two-way table between two variables, which are registered in two distinct registrations. Often, one is interested not only in estimated two-way tables, or more generally, estimated linear relations, but in complete registrations in which both variables are simultaneously registered. Users of statistics find such complete data-bases easy to analyze. The creation of such enriched registrations can be seen as a special case of imputation. One registration serves as a host or recipient source, and the other as a donor source. Assuming the second registration to be the donor source, the problem is imputing Z-values from the second registration, into the first registration using the estimated two-way table discussed in Section 2.2, as auxiliary information. Statistical matching problems using data from a third data source, have already been considered by Rubin (1986) and Paass (1986). Singh *et al.* (1993) gives a review of their methods. In addition, they propose some modifications to Rubin's (1986) and Paass's (1986) methods. Our imputation method is based on the regression method suggested by Rubin (1986) and Singh *et al.* (1993).

After having defined predictors for the Z-variables by means of the regression model

$$\hat{z}_k = A'x_k,\ k = 1,\ldots,N,$$

where A is given by (2), we define new predictions for these variables by means of the enlarged regression model

$$\tilde{z}_k = \alpha_1' x_k + \alpha_2' y_k,\ k = 1,\ldots,N,$$

with

$$\begin{pmatrix} \alpha_1 \\ \alpha_2 \end{pmatrix} = \left[\sum_{k=1}^{N} \begin{pmatrix} x_k x_k' & x_k y_k' \\ y_k x_k' & y_k y_k' \end{pmatrix}^{-1}\right] \left[\sum_{k=1}^{N} \begin{pmatrix} x_k z_k' \\ y_k z_x' \end{pmatrix}\right].$$

Using well-known results about partial regression coefficients in the general linear model (see *e.g.*, Seber 1977), α_1 and α_2 can be expressed as

$$\alpha_1 = A - B\alpha_2$$

and

$$\alpha_2 = \left[\sum_{k=1}^{N} (y_k - B'x_k)(y_k - B'x_k)'\right]^{-1} \times \left[\sum_{k=1}^{N} (y_k - B'x_k)(z_k - A'x_k)'\right],$$

where B and A are given by (1) and (2) respectively. They can be calculated from the first and second registration. The partial regression coefficients should be estimated from the third source. We suggest

$$\hat{\alpha}_1 = A - B\hat{\alpha}_2$$

and

$$\hat{\alpha}_2 = \left[\sum_{k=1}^{N} (y_k - B'x_k)(y_k - B'x_k)'\right]^{-1} \times \left[\sum_{k=1}^{n} w_k (y_k - B'x_k)(z_k - A'x_k)'\right],$$

where w_k are calibration weights which are discussed in Section 2.2. Based on these estimates we define new predictions for the Z-values:

$$\hat{z}_k = \hat{\alpha}_1' x_k + \hat{\alpha}_2' y_k = A'x_k + \hat{\alpha}_2'(y_k - B'x_k),\ k = 1,\ldots,N. \quad (7)$$

These new predictions equal the old predictions (see Section 2.1) plus an adjustment term. This adjustment term depends on the difference between the Y-value and its (old) prediction. It can be viewed as an attempt to improve the prediction for Z, however, and more important, it is a means to reconstruct the weighting type estimator under the third set of constraints (Section 2.2). Indeed, the following equality holds:

$$\sum_{k=1}^{N} y_k \hat{z}_k' = \sum_{k=1}^{N} (B'x_k)(A'x_k)' + \sum_{k=1}^{n} w_k (y_k - B'x_k)(z_k - A'x_k)'.$$

This is just the weighting type estimator under the third set of constraints, if the corresponding calibration weights are used to estimate α_2. It is easy to show that

$$\sum_{k=1}^{N} x_k \hat{z}_k' = \sum_{k=1}^{N} x_k \hat{z}_k' = \sum_{k=1}^{N} x_k z_k'.$$

So, also the XZ-table can be reconstructed. At the beginning of this section, we assumed the second registration to be the donor source. This choice was arbitrary. If the Y-values were imputed instead of the Z-values, we would have obtained an identical estimate for the YZ-table. In addition, the XY-table could have been reconstructed.

The new predictions for the Z-values can be used for imputation. Singh *et al.* (1993) give algorithms for imputation using regression models. These Z-values can be imputed in the first registration in two steps. In the first step, the predictions given by (7) are calculated for each

(x_k, y_k) in the first registration. We have shown that the crossings between the Y-values and these predicted Z-values, can be considered as weighting type estimators. However, the calculated predictions have in general no realistic values, and therefore the first step is followed by a second step. In the second step, each predicted Z-value in the first registration is replaced by a live Z-value from the second registration, which is nearest under some Euclidean distance in (X, Z).

2.5 Estimating Cross-Products for Continuous Y- and Z-Variables

The consistency property of the third set of constraints (Section 2.2) also hold with respect to continuous Y- and Z-variables, provided that there exist constants a_y and a_z of proper order, such that $a_y' y_k = 1$ and $a_z' z_k = 1$ for all k. To see this, we slightly extend the results of Section 2.1. First note that

$$a_y' B' x_k = a_y' \sum_{k=1}^{N} y_k x_k' \left(\sum_{k=1}^{N} x_k x_k' \right)^{-1} x_k =$$

$$a' \sum_{k=1}^{N} x_k x_k' \left(\sum_{k=1}^{N} x_k x_k' \right)^{-1} x_k = a' x_k = 1$$

(it is still assumed that there exists a constant a such that $a' x_k = 1$ for all k). Similarly, it holds that $a_z' A' x_k = 1$. The equivalent equations of (4) and (5) for the continuous case are readily obtained. Consequently, pre-multiplying both sides of (III) with a_y' gives $\sum_{k=1}^{n} w_k z_k' = \sum_{k=1}^{N} z_k'$ and post-multiplying both sides of (III) with a_z yields $\sum_{k=1}^{n} w_k y_k = \sum_{k=1}^{N} y_k$. So, the third set of constraints meets the consistency objective, *i.e.*, the calibration equation of the first set of constraints, for quite general Y- and Z-variables. We will give two examples.

In the first example we take $y_k = (1, y_{2k})'$ and $z_k = (1, z_{2k})'$, where both y_{2k} and z_{2k} are assumed to be continuous. By taking $a_y = a_z = (1, 0)'$ we see that $a_y' y = a_z' z = 1$ for all k. The cross-product between Y and Z equals

$$\sum_{k=1}^{N} y_k z_k' = \begin{pmatrix} N & \sum_{k=1}^{N} z_{2k} \\ \sum_{k=1}^{N} y_{2k} & \sum_{k=1}^{N} y_{2k} z_{2k} \end{pmatrix},$$

from which the covariance between y_{2k} and z_{2k} is easily derived. This cross-product can be estimated using the third set of constraints. An elaboration of this set gives the following four constraints for this particular example:

$$\sum_{k=1}^{n} w_k = N, \sum_{k=1}^{n} w_k y_{2k} = \sum_{k=1}^{N} y_{2k}, \sum_{k=1}^{n} w_k z_{2k} = \sum_{k=1}^{N} z_{2k},$$

and

$$\sum_{k=1}^{n} w_k (y_{2k} z_{2k} - (y_{2k} - B_2' x_k)(z_{2k} - A_2' x_k)) = \sum_{k=1}^{N} (B_2' x_k)(A_2' x_k),$$

where the regression coefficients are given by

$$B_2 = \left(\sum_{k=1}^{N} x_k x_k' \right)^{-1} \sum_{k=1}^{N} x_k y_{2k}$$

and

$$A_2 = \left(\sum_{k=1}^{N} x_k x_k' \right)^{-1} \sum_{k=1}^{N} x_k z_{2k}.$$

If one is specially interested in the correlation coefficient between y_{2k} and z_{2k}, then following constraints may be considered in addition:

$$\sum_{k=1}^{n} w_k y_{2k}^2 = \sum_{k=1}^{N} y_{2k}^2 \text{ and } \sum_{k=1}^{n} w_k z_{2k}^2 = \sum_{k=1}^{N} z_{2k}^2.$$

In the second example, we suppose that $y_k = (1, y_{2k})'$, where y_{2k} may be continuous, and z_k is categorical with q categories. By taking $a_y = (1, 0)'$ and $a_z = l$, where l is a vector of ones of proper order, we see that $a_y' y_k = a_z' z_k = 1$ for all k. The cross-product between Y and Z is

$$\sum_{k=1}^{N} y_k z_k' = \begin{pmatrix} N_1 & N_2 & \cdots & N_q \\ \sum_{k \in C_1} y_{2k} & \sum_{k \in C_2} y_{2k} & \cdots & \sum_{k \in C_q} y_{2k} \end{pmatrix},$$

where C_h denotes the set of population elements belonging to the h-th category of Z, and N_h the size of C_h. It is ensured that the calibration weights according to the third set of constraints, satisfy the 'marginal' calibration equations $\sum_{k=1}^{n} w_k z_k = \sum_{k=1}^{N} z_k = (N_1 \ldots N_q)'$ and $\sum_{k=1}^{n} w_k y_{2k} = \sum_{k=1}^{N} y_{2k}$, which both may be of interest in view of consistency requirements.

3. COMBINING INDEPENDENT SAMPLES ACROSS COMMON VARIABLES

In the previous section, we have presented a method for combining two registrations across common variables, using auxiliary information from a small sample. In this section, the method is adjusted by combining two independent samples. We consider a complete registration of persons, two large-scale sample surveys, and a small-scale sample survey. The registration contains a limited set of variables such as sex, age, region, and marital status. These

variables are denoted by X. In the one large sample, the variables Y, U, and X are observed, and in the other large sample, the variables Z, U, and X. In the small sample all variables, *i.e.*, Y, Z, U, and X, are observed. The small sample could come from a specially conducted small-scale survey, or from sample overlap of the large-scale surveys. In Figure 1, the data sources are schematically given. For convenience, it is assumed that all samples correspond to different units, *i.e.*, it is assumed that there is no sample overlap.

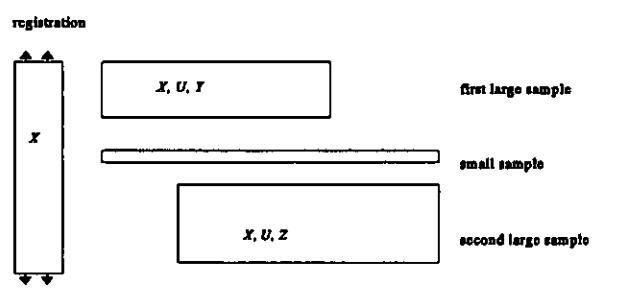

Figure 1. Overview of the Several Data Sources

The common variables X and U are partitioned into $C = (X\ U)$, where X denotes the set of common variables with known population totals, and U denotes the set of common variables with unknown population totals. All samples may be drawn by some complex sampling design. Both Y and Z are assumed to be categorical, however, as in Section 2.5, the suggested weighting methods are also applicable for continuous Y and Z. The purpose is to estimate the two-way table between Y and Z. We consider two estimators. One estimator is based on incomplete two-way stratification (analogous to the first set of constraints of Section 2.2), and the other estimator is based on a mix between statistical matching and calibration (analogous to the third set of constraints of Section 2.2).

3.1 Incomplete Two-Way Stratification

First the population totals of Y and Z are estimated by means of the first and second (large) sample respectively. These population totals are estimated in two phases. In the first phase, both (large) samples are weighted using X as a set of control variables. This implies that both (large) samples are weighted such that they reproduce the known population totals of X, which are denoted by t_x. Based on these weights, a pooled estimate for the population totals of U is

$$\hat{t}_u = \lambda \sum_{k \in n_1} w_{1k} u_k + (1 - \lambda) \sum_{k \in n_2} w_{2k} u_k,$$

where w_{1k} and w_{2k} denote the (first phase) calibration weights of the first and second sample, and $\lambda \in [0, 1]$. In the second phase, both samples are reweighted using simultaneously X and U as control variables. Let v_{1k} and v_{2k} denote these second phase calibration weights. The resulting estimators for the population totals of Y and Z can be considered as calibration estimators in two phases (see Renssen and Nieuwenbroek 1997, Section 6). These estimators are denoted by $\hat{t}_y$ and $\hat{t}_z$ respectively:

$$\hat{t}_y = \sum_{k \in n_1} v_{1k} y_k \text{ and } \hat{t}_z = \sum_{k \in n_2} v_{2k} z_k.$$

We note that both estimators are based on a similar set of control variables. If the common set of variables is large, one may consider using a smaller subset to weight both samples. In general, the subset to weight the first sample may differ from the subset to weight the second sample. However, we shall assume in the sequel that both (large) samples are weighted according to the same set of control variables.

The two-way table between Y and Z can be estimated by weighting the (small) third sample, using simultaneously Y and Z as control variables, *i.e.*,

$$\hat{T} = \sum_{k \in n_3} w_{3k} (y_k z_k^t),$$

where the calibration weights w_{3k} satisfy the constraints

$$\sum_{k \in n_3} w_{3k} y_k = \hat{t}_y \text{ and } \sum_{k \in n_3} w_{3k} z_k = \hat{t}_z.$$

This is incomplete two-way stratification, where the unknown population totals of Y and Z are replaced by their estimates. These sets of constraints ensure precisely estimated marginal counts of the YZ-table if the common variables C are highly correlated with Y and Z.

3.2 Synthetic Two-Way Stratification

In this section, we consider an alternative estimator for the YZ-table, which also uses the (large) samples as a source of auxiliary information. However, instead of using estimated marginal counts as auxiliary information, estimated synthetic cell counts are used. Let B denote the population regression coefficient between Y and C, which is estimated by the first (large) sample:

$$\hat{B} = \left(\sum_{k \in n_1} v_{1k} c_k c_k^t \right)^{-1} \left(\sum_{k \in n_1} v_{1k} c_k y_k^t \right).$$

Similarly, let A denote the population regression coefficient between Z and C, which is estimated by the second (large) sample:

$$\hat{A} = \left(\sum_{k \in n_2} v_{2k} c_k c_k^t \right)^{-1} \left(\sum_{k \in n_2} v_{2k} c_k z_k^t \right).$$

Note that these estimated regression coefficients are based on the second phase calibration weights instead of the inclusion weights. If there exists a constant a, such that $a'c_k = 1$ for all k, then we still have $l'\hat{B}'c_k = l'\hat{A}'c_k = 1$ for all k. Now, inspired by the decomposition given by (3), *i.e.*,

$$\sum_{k=1}^{N} y_k z_k' = B' \sum_{k=1}^{N} (c_k c_k') A + \sum_{k=1}^{N} (y_k - B'c_k)(z_k - A'c_k)',$$

we suggest estimating the two-way table in two steps. In the first step the first term on the right-hand side is estimated by substituting the population regression coefficients B and A by their estimates $\hat{B}$ and $\hat{A}$. Furthermore, we suggest to estimate $\sum_c = \sum_{k=1}^{N} c_k c_k'$ by the pooled estimate:

$$\hat{\sum}_c = \gamma \sum_{k \in n_1} v_{1k}(c_k c_k') + (1-\gamma) \sum_{k \in n_2} v_{2k}(c_k c_k'),$$

where v_{1k} and v_{2k} denote the (second phase) weights of the first and second sample and $\gamma \in [0, 1]$. Eventually, the first term is estimated by $\hat{B}'\hat{\sum}_c \hat{A}$. Until now, no use of the third (small) sample has been made. If desired, estimates for B, A, and $\sum_c$ can be improved slightly by also using the small sample.

In the second step, the complete two-way table between Y and Z is estimated by weighting the third (small) sample according to the calibration estimator subject to the third set of constraints (see Section 2.2), where B, A, and $\sum_c$ are replaced by their estimates $\hat{B}$, $\hat{A}$, and $\hat{\sum}_c$. The resulting estimator equals

$$\sum_{k=1}^{n_3} w_3(y_k z_k') = \hat{B}'\hat{\sum}_c \hat{A} + \sum_{k=1}^{n_3} w_{3k}(y_k - \hat{B}'c_k)(z_k - \hat{A}'c_k)'. \quad (8)$$

The first term on the right-hand side is an estimate for the synthetic two-way table. This estimate is approximately unbiased for the YZ-table, if the conditional independence assumption holds. We note that, this type of estimator is essentially obtained by applying the constrained statistical matching method (see *e.g.*, Barr and Turner 1980, Rodgers 1984, or Rubin 1986). The second term is an adjustment term to obtain an approximately unbiased estimate for the YZ-table, without this assumption. If there exists a constant a such that $a'c_k = 1$ for all sampled elements, then we obtain by pre-multiplying both sides of (8) with l', the following estimator for the population total of Z:

$$\sum_{k \in n_3} w_{3k} z_k' = \left(\gamma \sum_{k \in n_1} v_{1k} c_k' + (1-\gamma) \sum_{k \in n_2} v_{2k} c_k'\right) \hat{A} = \left(\sum_{k \in n_2} v_{2k} c_k'\right) \hat{A} = a'\left(\sum_{k \in n_2} v_{2k} c_k c_k'\right) \hat{A} = \hat{t}_z'.$$

Similarly, we have by post-multiplying both sides with l, an estimator for the population total of Y:

$$\sum_{k \in n_3} w_{3k} y_k = \hat{B}'\left(\gamma \sum_{k \in n_1} v_{1k} c_k + (1-\gamma) \sum_{k \in n_2} v_{2k} c_k\right) = \hat{B}'\left(\sum_{k \in n_1} v_{1k} c_k\right) = \hat{B}'\left(\sum_{k \in n_1} v_{1k} c_k c_k'\right) a = \hat{t}_y.$$

It follows that the marginal cell counts of the estimated two-way table, are the two-phase calibration estimators for the population totals of Y and Z as defined Section 3.1.

3.3 A Simulation Study; Integration of Household Surveys

In this subsection, we wish to compare the weighting techniques incomplete two-way stratification as discussed in subsection 3.1, and synthetic two-way stratification as discussed in subsection 3.2, by means of a simulation study. To that purpose, we use a data set, which stems from a pilot study of the Dutch Household Survey on Living Conditions, (see van Tuinen 1995). The data set consists of 1,085 records of which the following variables are observed: age (six categories: 15-24, 25-34, 35-44, 45-54, 55-64, 65+), sex (two categories: male or female), ownership of house (two categories: yes or no), occupation (five categories: work, housekeeping, education, voluntary, other), and health (two categories: yes or no). On behalf of the simulation study, this data set is considered as a finite population. The population totals of age and sex are assumed to be known.

In order to simulate the weighting techniques, we have carried out a Monte Carlo algorithm. Namely, we have drawn 500 samples, independently of each other, according to a two-phase sampling design. In the first phase, a simple random sample of size 20,500 is drawn with replacement. In this sample, age, sex, and ownership of house, are observed. In the second phase, the (first phase) sample is randomly divided into two large sub-samples of sizes 10,000 and one small sub-sample of size 500; in the one large sub-sample, occupation is observed (denoted by Y), in the other large sub-sample, health (denoted by Z), and in the small sub-sample, both occupation and health are observed. At each run, we have estimated the two-way table between Y and Z, according to four weighting methods which are discussed next.

The first phase sample is weighted with a crossing between sex and age as control variables. This is just post-stratification with twelve post-strata. Based on these weights, population totals can be estimated for all observed variables in the first phase sample, and for crossings between them. In particular, we may reproduce the population totals for the crossing between age and sex, and obtain estimated population totals for the crossings between age, sex, and ownership of house. Now, we distinguish two sets of common variables to weight the large sub-samples, as well as to obtain an estimate for the synthetic two-way table between Y and Z. The first set is a crossing between age and sex (12 categories) and the second set is a crossing between age, sex, and ownership (24 categories). For each simulation, this gives two different estimates for the marginal counts, *i.e.* two different estimates for the population totals of Y and Z – note that both estimates are based on post-stratification – and two different estimates for the synthetic two-way table. In order to weight the small sub-sample, we distinguish between the weighting method based on incomplete two-way stratification, and the weighting method based on synthetic two-way stratification. Since two different sets of common variables are used to weight the large sub-samples, as well as for statistical matching, we obtain four sets of calibration weights for each simulation run with respect to the small sub-sample, which in turn gives for each simulation run, four different estimated two-way tables between Y and Z. For the ease of computation, we have used the quadratic distance measure in the calibration estimation, implying that each estimated cell corresponds to a general regression estimate. Finally, we have taken the averages and variances of these two-way tables over the 500 simulations. The results are shown in tables 5 to 8.

The averages over the 500 simulations are almost identical for the four types of estimators, as can be seen from these tables. Note that the given cell counts are rounded off. We have also calculated the real YZ-table from the finite population. The real counts equal exactly the averages, which are given in Table 5 (or 6). For this particular simulation study, we conclude that all estimators have a very small bias.

The variances over these 500 simulations are given within the brackets. The variances of the estimated marginal counts of Tables 5 and 7 coincide, because these estimates are based on the same estimator. For the same reason it holds that the variances of the estimated marginal counts in tables 6 and 8 coincide. Note that the variances of the estimated marginal counts in tables 6 and 8 are slightly smaller than the variances of the estimated marginal counts in Tables 5 and 7, due to the larger set of common variables. However, for most estimated marginal counts this variance reduction can be considered negligible.

Tables 5 and 6 give identical variances with respect to all estimated cell counts. The variances for most estimated cell counts in Table 7, are plainly smaller than those in tables 5 and 6. In Table 8, this variance reduction is even greater. For this particular example, we conclude that the use of the larger set of common variables, in combination with the first weighting method, slightly reduces the variances of the estimated marginal counts, but leaves the variances of the estimated cell counts unaffected. Naturally, using the larger set of common variables in combination with the second weighting method, also slightly reduces the variances of the marginal cell counts. Finally, given a set of common variables, the weighting method based on synthetic matching, results in smaller variances for the estimated cell counts, than the weighting method based on incomplete two-way stratification.

Table 5
Incomplete Two-way Stratification Combined with the First Set of Common Variables

	1	2	3	4	5	total
yes	$447_{(96)}$	$232_{(97)}$	$89_{(28)}$	$25_{(21)}$	$59_{(49)}$	$852_{(17)}$
no	$61_{(79)}$	$104_{(90)}$	$11_{(21)}$	$11_{(19)}$	$46_{(46)}$	$233_{(17)}$
total	$508_{(23)}$	$336_{(19)}$	$100_{(8)}$	$36_{(3)}$	$105_{(10)}$	1085

Table 6
Incomplete Two-way Stratification Combined with the Second Set of Common Variables

	1	2	3	4	5	total
yes	$447_{(96)}$	$232_{(97)}$	$89_{(28)}$	$25_{(21)}$	$59_{(49)}$	$852_{(17)}$
no	$61_{(79)}$	$104_{(90)}$	$11_{(21)}$	$11_{(19)}$	$46_{(46)}$	$233_{(17)}$
total	$508_{(23)}$	$336_{(19)}$	$100_{(8)}$	$36_{(3)}$	$105_{(9)}$	1085

Table 7
Synthetic Two-way Stratification Combined with the First Set of Common Variables

	1	2	3	4	5	total
yes	$447_{(75)}$	$231_{(74)}$	$89_{(17)}$	$25_{(20)}$	$59_{(42)}$	$851_{(17)}$
no	$61_{(58)}$	$105_{(65)}$	$11_{(12)}$	$11_{(19)}$	$46_{(38)}$	$234_{(17)}$
total	$508_{(23)}$	$336_{(19)}$	$100_{(8)}$	$36_{(3)}$	$105_{(10)}$	1085

Table 8
Synthetic Two-way Stratification Combined with the Second Set of Common Variables

	1	2	3	4	5	total
yes	$447_{(70)}$	$231_{(70)}$	$89_{(16)}$	$25_{(18)}$	$59_{(40)}$	$851_{(17)}$
no	$61_{(52)}$	$105_{(60)}$	$11_{(11)}$	$11_{(16)}$	$46_{(37)}$	$234_{(17)}$
total	$508_{(23)}$	$336_{(19)}$	$100_{(8)}$	$36_{(3)}$	$105_{(9)}$	1085

3.4 Imputing Values of the one Large Sample into the Other Large Sample

By means of the two large samples and the small sample, one may construct a synthetic sample in which the real Y-values and predicted Z-values, and/or the predicted Y-values and the real Z-values are simultaneously recorded.

We define predictions for the Y- and Z-values analogously to (7), namely

$$\hat{y}_k = \hat{B}'c_k + \tilde{\beta}_2' \ (z_k - \hat{A}'c_k), k = 1, ..., n_2, \tag{9}$$

and

$$\hat{z}_k = \hat{A}'c_k + \tilde{\alpha}_2' \ (y_k - \hat{B}'c_k), k = 1, ..., n_1, \tag{10}$$

with

$$\tilde{\beta}_2 = \left[\sum_{k=1}^{n_2} v_{2k}(z_k - \hat{A}'c_k)(z_k - \hat{A}'c_k)'\right]^{-1} \times \left[\sum_{k=1}^{n_3} w_{3k}(y_k - \hat{B}'c_k)(z_k - \hat{A}'c_k)'\right],$$

and

$$\tilde{\alpha}_2 = \left[\sum_{k=1}^{n_1} v_{1k}(y_k - \hat{B}'c_k)(y_k - \hat{B}'c_k)'\right]^{-1} \times \left[\sum_{k=1}^{n_3} w_{3k}(y_k - \hat{B}'c_k)(z_k - \hat{A}'c_k)'\right].$$

For each (c_k, y_k) the Z-values can be imputed in the first large sample by means of (10), $k = 1, ..., n_1$, and similarly for each (c_k, z_k) the Y-values can be imputed in the second large sample by means of (9), $k = 1, ..., n_2$. Based on these imputed values, we may define the following estimates for the two-way table between Y and Z:

$$\sum_{k=1}^{n_1} v_{1k}\, y_k \hat{z}_k' = \hat{B}' \sum_{k=1}^{n_1} v_{1k}\, c_k\, c_k'\, \hat{A} + \sum_{k=1}^{n_3} w_{3k}(y_k - \hat{B}'c_k)(z_k - \hat{A}'c_k)' \tag{11}$$

and

$$\sum_{k=1}^{n_2} v_{2k}\, \hat{y}_k z_k' = \hat{B}' \sum_{k=1}^{n_2} v_{2k}\, c_k\, c_k'\, \hat{A} + \sum_{k=1}^{n_3} w_{3k}(y_k - \hat{B}'c_k)(z_k - \hat{A}'c_k)'. \tag{12}$$

One estimate is based on the first synthetic sample, the other on the second synthetic sample. By pooling the synthetic samples, one obtains a pooled synthetic sample of size $n_1 + n_2$, from which a pooled estimated for the two-way table can be constructed. This pooled estimate shows a close resemblance to (8). Note that if C and Z are perfectly correlated, then the left-hand side of (11) reduces to $\sum_{k=1}^{n_1} v_{1k} y_k z_k'$, *i.e.*, our estimated two-way table corresponds to a weighted estimated two-way table based on the first sample, as if the real values of Z were imputed in this sample. Similarly, if C and Y are perfectly correlated, then (12) reduces to $\sum_{k=1}^{n_2} v_{2k} y_k z_k'$.

An important special case to consider, is when c is categorical. Then the following equalities hold true:

$$\sum_{k \in n_1} v_{1k}(c_k c_k') = \sum_{k \in n_2} v_{2k}(c_k c_k') = \text{diag}\begin{pmatrix} t_x \\ \hat{t}_u \end{pmatrix},$$

so (11) and (12) coincide. Furthermore, we have for categorical c:

$$\sum_{k \in n_1} v_{1k} c_k \hat{z}_k' = \sum_{k \in n_2} v_{2k} c_k z_k'$$

and

$$\sum_{k \in n_2} v_{2k} \hat{y}_k c_k' = \sum_{k \in n_1} v_{1k} y_k c_k'.$$

Obviously, if c is categorical, then it suffices to create a synthetic sample, which is based on either the first synthetic sample or the second synthetic sample. In either case, the weighting type estimates for the CZ-table, the CY-table, and the YZ-table, can be reconstructed. Finally, we note that the imputed values in all synthetic samples may be unrealistic. As described in Section 2.4, the calculated predictions may be replaced by live values according to some algorithm.

4. SUMMARY

In this article we presented a weighting procedure to combine information from distinct sample surveys. The linking pin between these surveys, is a set of common variables, (see Figure 1). It is argued that these samples should be weighted according to a sequential structure. First, both large samples were weighted using X as control variables. Based on these weighted samples, we could obtain a pooled estimate for the population total of U. Then both large samples were reweighted using simultaneously X and U as control variables. This gave an estimate for the population total of Y and Z.

Using statistical matching techniques with X and U as common variables, we may also obtain an estimate for a synthetic two-way table between Y and Z. Eventually, the small sample was weighted according to two different sets of control variables. The first set of control variables corresponded to the estimated population totals of Y and Z, and the second set of control variables to the estimated synthetic two-way table. Using the first set of control variables, is strongly related to incomplete two-way stratification. The theoretical framework needed to develop the second weighting method, was discussed all through this article. By means of both weighting methods, the

YZ-table can be estimated (it is tacitly assumed that *Y* and *Z* are categorical). The marginal counts of the *YZ*-table corresponding to the first weighting method, equal by definition of the calibration equations, the estimated population totals of *Y* (which is based on the first large sample) and *Z* (which is based on the second large sample). It was shown, that this consistency property also holds for the second weighting method. A numerical study was conducted to evaluate the performance of the weighting methods with respect to the cell counts. It was found that both weighting methods yielded nearly (design) unbiased estimated two-way tables. The simulated (design) variances of the second weighting method, appeared to be smaller than the corresponding (design) variances of the first weighting method, with respect to all estimated cell counts. In principle, the *Y*- and *Z*-variables were assumed to be categorical, however, it was argued that the ideas presented were also applicable for continuous *Y* and *Z* or for continuous *Y* and categorical *Z*.

ACKNOWLEDGEMENTS

The author wishes to thank Peter Kooiman, Nico Nieuwenbroek, and Ger Slootbeek for their careful reading and useful remarks. The author also thanks two anonymous referees and an associated editor for their valuable suggestions to improve the article. The views expressed in this article are those of the author and do not necessarily reflect the policy of Statistics Netherlands.

REFERENCES

ALEXANDER, C.H. (1987). A class of methods for using person controls in household weighting. *Survey Methodology*, 13, 183-198.

BAKKER, B.F.M., and WINKELS, J.W. (1998). Why integration of household surveys? – Why POLS?. *Netherlands Official Statistics*, 13, 5-7.

BARR, R.S., and TURNER, J.S. (1980). Merging the 1977 Statistics of Income and the March 1978 Current Population Survey. Technical report, U.S. Department of the Treasury, Office of Tax Analysis.

BETHLEHEM, J.G., and KELLER, W.J. (1987). Linear weighting of sample survey data. *Journal of Official Statistics*, 3, 141-153.

COPELAND, K.R., PEITZMEIER, F.K., and HOY, C.E. (1987). An alternative method of controlling Current Population Survey estimates to population counts. *Survey Methodology*, 13, 173-181.

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

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

FELLEGI, I.P., and SUNTER, A.B. (1969). A theory for record linkage. *Journal of the American Statistical Association*, 64, 1183-1210.

HOFMANS, M.G. (1998). Innovative weighting in POLS. Making use of core questions. *Netherlands Official Statistics*, 13, 12-15.

LEMAÎTRE, G., and DUFOUR, J. (1987). An integrated method for weighting persons and families. *Survey Methodology*, 13, 199-208.

MADDALA, G.S. (1983). *Limited-dependent and Qualitative Variables in Econometrics*. Cambridge: Cambridge University Press.

OH, H.L., and SCHEUREN, F. (1987). Modified raking ratio estimation. *Survey Methodology*, 13, 209-219.

PAASS, G. (1986). Statistical match: Evaluation of existing procedures and improvements by using additional information. In *Microanalytic Simulation Models to Support Social and Financial Policy*. Amsterdam: Elsevier Science.

RAGHUNATHAN, T.E., and GRIZZLE, J.E. (1995). A split questionnaire survey design. *Journal of the American Statistical Association*, 90, 54-63.

RENSSEN, R.H., and NIEUWENBROEK, N.J. (1997). Aligning estimates for common variables in two or more sample surveys. *Journal of the American Statistical Association*, 92, 368-374.

RODGERS, W.L. (1984). An evaluation of statistical matching. *Journal of Business & Economic Statistics*, 2, 91-102.

RUBIN, D.B. (1986). Statistical matching using file concatenation with adjusted weights and multiple imputations. *Journal of Business & Economic Statistics*, 4, 87-94.

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

SEARLE, S.R. (1971). *Linear Models*. New York: John Wiley & Sons.

SEBER, G.A.F. (1977). *Linear Regression Analysis*. New York: John Wiley & Sons.

SINGH, A.C., MANTEL, H.J., KINACK, M.D, and ROWE, R. (1993). Statistical matching: use of auxiliary information as an alternative to the conditional independence assumption. *Survey Methodology*, 19, 59-79.

TUINEN VAN, H.K. (1995). Social indicators, social surveys and integration of social statistics. *Statistical Journal of the United Nations ECE*, 12, 379-394.

WINKELS, J.W., and EVERAERS, P.C.J. (1998). Design of an integrated survey in the Netherlands. The case POLS. *Netherlands Official Statistics*, 13, 6-11.

ZIESCHANG, K.D. (1990). A generalized least squares weighting system for the consumer expenditure survey. *Journal of the American Statistical Association*, 85, 986-1001.

Survey Methodology, December 1998
Vol. 24, No. 2, pp. 185-192
Statistics Canada

Sampling on Two Occasions: Estimation of Population Total

RAGHUNATH ARNAB[1]

ABSTRACT

Two sampling strategies have been proposed for estimating the finite population total for the most recent occasion, based on the samples selected over two occasions involving varying probability sampling schemes. Attempts have been made to utilize the data collected on a study variable, in the first occasion, as a measure of size and a stratification variable for selection of the matched-sample on the second occasion. Relative efficiencies of the proposed strategies have been compared with suitable alternatives.

KEY WORDS: Composite estimator; Matched-sample; Sampling schemes; Sampling strategies; Varying probability sampling schemes.

1. INTRODUCTION

We very often survey the same population at regular time intervals to estimate the same population characteristics which change over time. For example, many countries collect data to estimate total number of unemployed persons, HIV infected people, immigrants *etc.*, on an annual or quarterly basis. In this article, we consider a finite population $U = (U_1, ..., U_i, ..., U_N)$ of N identifiable units, which is supposed to be sampled over two occasions, to estimate the population total of a variable under study for the current (second) occasion. In successive sampling, one utilizes data collected on the previous (first) occasion effectively, to get an efficient strategy in consideration of cost, and providing an efficient estimator of the population total for the current occasion. Extensive literature is now available for this purpose. Singh (1967), and Avadhani and Sukhatme (1970) utilized information, collected on the first occasion as a measure of size, for the selection of the matched sample on the second occasion; while Arnab (1991) utilized such information as a stratification variable, as well as the measure of size, for selection of the sample on the second occasion. Recently, Prasad and Graham (1994) modified Raj's (1965) and Chotai's (1974) sampling strategies, by using information of the first occasion as a measure of size, for the selection of the matched sample in the second occasion. They found empirically, that one of their proposed strategies fares better than that given by Chotai (1974). In this article, two alternative strategies are proposed. One of them utilizes information in the first occasion as a measure of size, and the other utilizes information as a measure of size and also as a stratification variable for selection of the matched sample in the second occasion. In this paper, it is shown that one of the proposed strategies is better than that given by Prasad and Graham (1994) and for the other, we do not have any definite theoretical conclusion. However, empirical evidence shows that the latter is more efficient than that described by Prasad and Graham (1994), as well as the former proposed strategy. This is possible because it utilizes first occasion values in all possible stages *viz.*, stratification, estimation and selection of the matched sample in the second occasion.

The general methods of selection of samples and estimation over two occasions are described below.

1.1 Sampling Schemes

On the first occasion, a sample s_1, of size n, is selected by some suitable sampling design, say $\boldsymbol{P}_1$, and the data y_{1i} $i\varepsilon s_1$, is obtained where y_{1i} (y_{2i}) is the value of the variate y under study, for the i-th unit on the first (second) occasion. On the second occasion, a matched sample (sub-sample) s_m of size $m\,(=n\lambda$, assumed to be an integer, $0 \le \lambda \le 1)$ is selected from s_1 by some suitable sampling scheme P_m, and it is supplemented by an un-matched sample s_u of size $u\,(=n\mu = n - m, \mu = 1 - \lambda)$ either from the entire population U or from U/s_1, the set of units not selected in the first occasion, by some suitable sampling design $\boldsymbol{P}_u$, and information $y_{2i}\,(i\varepsilon s_m, i\varepsilon s_u)$ on the second occasion is obtained. It is obvious that the cost of survey for the matched sampled units is expected to be much lower than that of the un-matched units, but for the sake of simplicity, we assume that the cost of the survey remains the same for all the units in the second occasion.

1.2 Method of Estimation

From the data $y_{1i}, i\varepsilon s_1$, and $y_{2i}, i\varepsilon s_m$ collected through the initial sample s_1, and the matched sample s_m, an unbiased estimator $\hat{Y}_{2m}$ for Y_2, the population total for the second occasion, is formed by treating the y_{1i}'s, $i\varepsilon s_1$, as auxiliary information. Thus $\hat{Y}_{2m}$ is normally a difference, ratio or regression estimator. From the un-matched sample s_u, an unbiased estimator $\hat{Y}_{2u}$ is also constructed for Y_2. Finally, a composite estimator, a combination of $\hat{Y}_{2m}$ and

[1] Raghunath Arnab, Department of Statistics, University of Durban-Westville, Private Bag-X54001, Durban - 4000, South Africa.

$\hat{Y}_{2u}$, is obtained by using a suitable weight of φ $(0 \le \varphi \le 1)$, as

$$\hat{Y}_2 = \varphi \hat{Y}_{2m} + (1 - \varphi) \hat{Y}_{2u}. \quad (1)$$

The optimum value of $\varphi = \varphi(\lambda)$ is obtained by minimizing $V(\hat{Y}_2)$, the variance of $\hat{Y}_2$ with respect to φ, for a given value of m (*i.e.*, λ). The expressions for $\varphi(\lambda)$ and $V(\hat{Y}_2 I\lambda)$, the variance of $\hat{Y}_2$ with $\varphi = \varphi(\lambda)$ are obtained as follows, when $\hat{Y}_{2m}$ and $\hat{Y}_{2u}$ are independent:

$$\varphi(\lambda) = (1/V_m)[1/V_m + 1/V_u]^{-1},$$

$$V(\hat{Y}_2 I\lambda) = [1/V_m + 1/V_u]^{-1},$$

where V_m and V_u are variances of $\hat{Y}_{2m}$ and $\hat{Y}_{2u}$ respectively. The optimum proportion of matched sample $\lambda = \lambda_0$, is obtained by minimizing $V(\hat{Y}_2 I\lambda)$ with respect to λ. Finally, putting $\lambda = \lambda_0$ in the expression for $V(\hat{Y}_2 I\lambda)$, the minimum variance of $\hat{Y}_2$ is obtained, and it will be denoted by $V_{\min}(\hat{Y}_2) = V(\hat{Y}_2 I\lambda_0)$. Our object is to find a suitable strategy, which is a combination of $P = (P_1, P_m, P_u)$ and $\hat{Y}_2$, to control the magnitude of $V_{\min}(\hat{Y}_2)$ to a minimum.

1.3 A Few Sampling Strategies

1.3.1 Avadhani and Sukhatme (1970)

On the first occasion, the initial sample s_1 of size n was selected by simple random sampling without replacement (SRSWOR) method, assuming that no auxiliary information is available prior to this survey. On the second occasion, the matched sample s_m of size m was selected from s_1 by the Rao, Hartley and Cochran (RHC, in brief, 1962) sampling scheme using y_{1i} as a measure of size for the i-th unit $i\varepsilon s_1$, assuming y_{1i}'s are positive. Under the RHC sampling scheme, the selected n units of s_1, are divided at random into m groups, each of size n/m, which is assumed to be an integer. From each of the selected groups, one unit is selected independently with probability proportional to the measure of size. Thus if the i-th unit, U_i, belongs to the j-th group G_j $(j = 1, ..., m)$ then U_i will be selected with the probability $q_i^*(i\varepsilon s_1) = y_{1i}/\sum_{i\varepsilon s_1} y_{1i}$. The un-matched sample s_u was selected from U/s_1 by SRSWOR.

1.3.2 Chotai (1

On the first occasion, the initial sample s_1 of size n was selected by the RHC scheme of sampling (assuming N/n is an integer), as described above with probability proportional to z_i, the size measure for the i-th unit which is, assumed to be positive and known for every $i\varepsilon U$. Let $\Delta_j = \sum_{k\varepsilon G_j} p_k$, the sum of p_k $(= z_k/Z, Z = \sum_{i\varepsilon U} z_i)$ values that belong to the random group G_j $(j = 1, ..., n)$, which is formed in selecting the sample s_1 by the RHC method. The matched sample s_m was selected from s_1 by the RHC scheme, with normed size measure Δ_i, for the i-th unit $i\varepsilon s_1$ $(\sum_{i\varepsilon s_1} \Delta_i = 1)$ assuming n/m is an integer. The un-matched sample, s_u was selected by the RHC sampling scheme with normed size measure p_i for the i-th unit assuming N/u is an integer. Let P_i^+ (P_i') = total of the Δ_i (p_i) values associated with those units that belong to the random group from which the i-th unit was selected in s_m (s_u) by the RHC sampling scheme with $\sum_{i\varepsilon s_m} P_i^+ = 1$ $(\sum_{i\varepsilon s_u} P_i' = 1)$.

The composite estimator for Y_2 is given by

$$\hat{Y}_2 = \varphi \hat{Y}_{2m} + (1 - \varphi) \hat{Y}_{2u}$$

where

$$\hat{Y}_{2m} = \sum_{i\varepsilon s_m} (y_{2i}/p_i) P_i^+ - \gamma\left[\sum_{i\varepsilon s_m} (y_{1i}/p_i) P_i^+ - \sum_{i\varepsilon s_1} (y_{1i}/p_i) \Delta_i\right];$$

$$\hat{Y}_{2u} = \sum_{i\varepsilon s_u} (y_{2i}/p_i) P_i' \quad (2)$$

where γ is a suitably chosen constant to minimize variance of $\hat{Y}_{2m}$. Chotai (1974) derived the expression for the minimum variance of $\hat{Y}_2$ as

$$V_{\min}(\hat{Y}_2) = k[1 - f + \sqrt{(1 - \delta^{*2})}]\,\sigma_2^2/2 = V_c \text{ (say)} \quad (3)$$

where

$$k = N/\{n(N-1)\},\ f = n/N,$$

$$\sigma_t^2 = \sum_{i\varepsilon U} p_i (y_{ti}/p_i - Y_t)^2,\ t = 1, 2$$

$$Y_t = \sum_{i\varepsilon U} y_{ti},\ t = 1, 2$$

$$\delta^* = \sum_{i\varepsilon U} p_i (y_{2i}/p_i - Y_2)(y_{1i}/p_i - Y_1)/(\sigma_1 \sigma_2). \quad (4)$$

1.3.3 Arnab (1991)

Arnab (1991) presented several strategies where the initial sample s_1 was selected by probability proportional to size with replacement (PPSWR) using normed size measure $p_i = z_i/Z$ for the i-th unit. Utilizing the ascertain values y_{1i}'s $(i\varepsilon s_1)$ on the basis of certain criteria, the n sample units are assigned to a suitable number of L strata. Let s_{1h} be the sample of size n_h, belonging to the h-th stratum $(s_1 = U_h s_{1h}$ and $\sum_h n_h = n)$. Here, it is assumed that n is large enough to ensure that n_h is positive for every h in practice. On the second occasion, sub-samples s_{mh}'s

of size m_h's ($= \nu_h n_h$, ν_h is a predetermined fraction and m_h is assumed to be an integer) are selected from s_{1h}'s independently, by suitable sampling schemes involving y_{1i}'s, $i \varepsilon s_1$ in the selection of matched samples s_{mh}'s. The unmatched sample s_u is selected by PPSWR method from the entire population U using z_i' as measure of size.

1.3.4 Prasad and Graham (1994)

Here the initial sample s_1 is selected by the RHC scheme of sampling similar to Chotai (1974) with normed size measure $p_i = z_i / Z$ for the i-th unit. The matched sample s_m is selected from s_1 by the RHC scheme with $p_i^* = (y_{1i}\Delta_i/p_i)/\sum_{i\varepsilon s_1}(y_{1i}\Delta_i/p_i)$ for the i-th unit, $i\varepsilon s_1$; where Δ_i is the sum of the p_j values for the group containing the i-th unit, formed in selecting s_1 by the RHC sampling scheme of sampling. The un-matched sample, s_u was selected from the entire population U by the RHC scheme similar to that presented by Chotai (1974). Here also N/n, n/m and N/u are assumed to be integers. Prasad and Graham (1994) proposed the following composite estimator for Y_2:

$$\hat{Y}_2 = \varphi \hat{Y}_{2m} + (1 - \varphi) \hat{Y}_{2u}$$

where $\hat{Y}_{2m} = \sum_{i\varepsilon s_m}(y_{2i}^*/p_i^*)\tilde{P}_i$; $\hat{Y}_{2u} = \sum_{i\varepsilon s_u}(y_{2i}/p_i)P_i'$; $y_{2i}^* = y_{2i}\Delta_i/p_i$; $\tilde{P}_i(P_i') =$ total of the $p_i^*(p_i)$ values associated with those units that belong to the random group from which the i-th unit was selected in $s_m(s_u)$. The expression for minimum variance of $\hat{Y}_2$, is obtained as:

$$V_{\min}(\hat{Y}_2) = k(1 - f + \sqrt{\zeta})\,\sigma_2^2/2 = V_{PG}\,(\text{say}) \qquad (5)$$

where

$$\zeta = \sigma_3^2/\sigma_2^2,\ \sigma_3^2 = \sum_{i\varepsilon U} q_i(y_{2i}/q_i - Y_2)^2,\ q_i = y_{1i}/Y_1; \qquad (6)$$

k, f, σ_2^2 and Y_1 are defined in (4).

In Prasad and Graham's (1994) expression for $V_{\min}(\hat{Y}_2)$, the divisor 2 was omitted and is obviously a typographical error.

Remark 1.1

From the strategies described in section 1.3, we note that the Avadhani and Sukhatme (1970) scheme does not require information on size measures in the whole frame, and hence is less demanding than the others. Chotai (1974) used the original size measures p_i in selection, but the first survey values y_{1i}'s, $i\varepsilon s_1$ were used additionally in estimation only. The use of additional information, p_i's, for the selection of the initial sample s_1 will make Chotai's (1974) strategy more efficient than that of Avadhani and Suhkatme (1970). But to use the optimal estimator $\hat{Y}_2$ for the Avadhani and Sukhatme (1970) strategy, one needs to estimate φ, the only unknown parameter. However, in Chotai's (1974) strategy, both the parameters φ and γ have to be estimated in order to use the optimum $\hat{Y}_2$. Prasad and Graham (1994) used both these variables in the selection of the matched sample (hence automatically in the estimation) and showed empirically that their strategy fares better than that of Chotai (1974). In addition, to gain in efficiency, Prasad and Graham's (1994) strategy can be used in practice, because $\hat{Y}_2$ involves only one unknown parameter, φ. It should be noted that Arnab (1991) first introduced the principle of stratification using y_{1i}'s, $i\varepsilon s_1$ as a stratification variable. This should always be done in practice whenever the necessary information is available, particularly in the selection of large units with marked size differences of the type considered in the numerical examples in section 3. Arnab's (1991) strategy is expected to be more efficient than the preceding strategies, since it utilizes first occasion values for stratification in addition to estimation. However, the optimal estimator $\hat{Y}_2$ contains the several unknown parameters (for details see Arnab 1991) which may hinder the application of the strategy especially when the sample size is not large enough.

2. PROPOSED STRATEGIES

Here two sampling strategies have been proposed which are modifications of strategies proposed by Prasad and Graham (1994) and Arnab (1991), respectively.

2.1 Strategy 1

The sampling scheme for this strategy is the same as was considered by Prasad and Graham (1994), and described in section 1.3.4. Here, only the estimator based on the matched sample s_m, has been modified by introducing the original size measure into the estimation. The proposed modified estimator $\hat{Y}_{2m}^*$ and the composite estimators for Y_2 are as follows:

$$\hat{Y}_{2m}^* = \sum_{i\varepsilon s_m}(y_{2i}^*/p_i^*)\tilde{P}_i - \beta\left[\sum_{i\varepsilon s_m}(z_i^*/p_i^*)\tilde{P}_i - Z\right] =$$

$$\sum_{i\varepsilon s_m}(r_i^*/p_i^*)\tilde{P}_i + \beta Z$$

where $z_i^* = z_i\Delta_i/p_i, y_{2i}^* = y_{2i}\Delta_i/p_i, r_i^* = r_i\Delta_i/p_i, r_i = y_{2i} - \beta z_i$ and β is a suitably chosen constant to minimize variance of $\hat{Y}_{2m}^*$; $p_i^*, \tilde{P}_i$ and Δ_i are as described in the section 1.3.4;

$$\hat{Y}_2 = \varphi\,\hat{Y}_{2m}^* + (1 - \varphi)\,\hat{Y}_{2u}$$

where $\hat{Y}_{2u}$ is given in (2).

Denoting $E_1(V_1)$ as unconditional expectation (variance) over selection of the sample s_1, and $E_2(V_2)$ the conditional expectation (variance) over s_m when s is fixed, one gets the variance of $\hat{Y}_{2m}^*$ for a given value of β, as

$$V(\hat{Y}_{2m}^* I\beta) = E_1 V_2(\hat{Y}_{2m}^* I\beta) + V_1 E_2(\hat{Y}_{2m}^* I\beta).$$

Following Prasad and Graham (1994), we obtain

$$E_1 V_2(\hat{Y}^*_{2m} I\beta) = k_1 \sigma_3^{*2}(\beta)$$

and

$$V_1 E_2(\hat{Y}^*_{2m}) = k(1-f)\sigma_2^2$$

where

$$k_1 = N(n-m)/\{nm(N-1)\};$$

$$\sigma_3^{*2}(\beta) = \sum_{i\epsilon U} q_i (r_i/q_i - R)^2$$

$$= \sigma_3^2 + \beta^2 \sigma_0^2 - 2\beta\sigma_0\sigma_3\delta;$$

$$R = \sum_{i\epsilon U} R_i = Y_2 - \beta Z, \delta = \sigma_{03}/(\sigma_0\sigma_3),$$

$$\sigma_0^2 = \sum_{i\epsilon U} q_i (z_i/q_i - Z)^2,$$

$$\sigma_{03} = \sum_{i\epsilon U} q_i (y_{2i}/q_i - Y_2)(z_i/q_i - Z) \tag{7}$$

σ_2^2, k and σ_3^2, q_i are as in (4) and (6), respectively. The optimum value of β that minimizes $V(\hat{Y}^*_{2m} I\beta)$ comes out as, opt $\beta = \beta_0 = \delta\sigma_3/\sigma_0$.

Putting the optimum value of $\beta = \beta_0$ in the expression of $V(\hat{Y}^*_{2m} I\beta)$, we get the optimum value of

$$V(\hat{Y}^*_{2m} I\beta) = V(\hat{Y}^*_{2m} I\beta_0) = k[(1-f) + (1-\lambda)\zeta^*/\lambda]\sigma_2^2$$

where $\zeta^* = (1-\delta^2)\zeta$; k, f and ζ are defined in (4) and (6) respectively.

The optimum variance of $\hat{Y}_2$ for a given value of λ is obtained by minimizing the variance of $\hat{Y}_2$ with respect to φ when $\beta = \beta_0$, and is given by

$$V_{opt}(\hat{Y}_2 I\lambda) = [1/V(\hat{Y}^*_{2m} I\beta_0) + 1/(\hat{Y}_{2u})]^{-1}$$

$$= [1/\{k(1-f) + (1-\lambda)\zeta^*/\lambda\} + \mu/\{k(1-f\mu)\}]^{-1}\sigma_2^2.$$

Finally, minimizing $V_{opt}(\hat{Y}_2 I\lambda)$ with respect to λ, the optimum proportion of the matched sample and minimum variance of $\hat{Y}_2$ are obtained respectively as

$$\text{opt } \lambda = \lambda_0 = \sqrt{\zeta^*}/(1+\sqrt{\zeta^*})$$

and

$$V_{min}(\hat{Y}_2) = k(1-f+\sqrt{\zeta^*})\sigma_2^2/2 = M_1 \text{ (say)} \tag{8}$$

Remark 2.1

The estimator $\hat{Y}^*_{2m}$, described in (1) is usable in practice when the optimum value of $\beta = \beta_0$ is known, or a good guess value of β_0 is available from some previous surveys. If instead of the regression estimator $\hat{Y}^*_{2m}$ described above, one uses the difference estimator $\hat{Y}^{**}_{2m} = \sum_{i\epsilon s_m}(y^*_{2i}/p^*_i)\tilde{P}_i - [\sum_{i\epsilon s_m}(z^*_i/p^*_i)\tilde{P}_i - Z]$ based on the matched sample, the expression for the minimum variance of $\hat{Y}_2$ would be as follows:

$$V_{min}(\hat{Y}_2) = k(1-f+\sqrt{\tilde{\zeta}})\sigma_2^2/2 = \tilde{M}_1 \text{ (say)}$$

with

$$\tilde{\zeta} = (1+\tau^2 - 2\tau\delta)\zeta, \ \tau = \sigma_0/\sigma_3.$$

2.1.1 Variance Estimation

To get approximate unbiased estimators for $V_{opt}(\hat{Y}_2)$, we first present the following theorems without proof:

Theorem 1

$$\hat{V}(\hat{Y}^*_{2m}) = \{k/(1-k)\}\left[\left\{\sum_{i\epsilon s_m}(y_{2i}^2\Delta_i/p_i^2)\tilde{P}_i/p_i^* - \hat{Y}^{*2}_{2m}\right\}\right.$$

$$\left. + \{k_2/k\}\sum_{i\epsilon s_m}\tilde{P}_i\left(r_i^*/p_i^* - \sum_{i\epsilon s_m}\tilde{P}_i r_i^*/p_i^*\right)^2\right]$$

is an unbiased estimator of $V(\hat{Y}^*_{2m})$, when β_0 is known, $k = (N-n)/\{n(N-1)\}$ and $k_2 = (n-m)/\{m(n-1)\}$.

Theorem 2

$E_1 V_2[\sum_{i\epsilon s_m}\tilde{r}_i^*/p_1^*] = N(n-m)/\{nm(N-1)\}[\sigma_3^2 + \sigma_0^2 - 2\sigma_{03}]$ can be estimated unbiasedly by

$$\{(n-m)/n(m-1)\}\sum_{i\epsilon s_m}(\tilde{r}_i^*/p_i^* - \sum_{i\epsilon s_m}\tilde{r}_i^*/p_i^*)^2\tilde{P}_i$$

where $\tilde{r}_i^* = \tilde{r}_i\Delta_i/p_i$, $\tilde{r}_i = y_{2i} - z_i$; σ_3^2, σ_0^2 and σ_{03} are given in (4) and (7) respectively.

From the Theorem 2 we note that

$$\hat{\sigma}_0^2 = d\sum_{i\epsilon s_m}\left(z_i/p_i^* - \sum_{i\epsilon s_m} z_i\tilde{P}_i/p_i^*\right)^2\tilde{P}_i,$$

$$\hat{\sigma}_3^2 = d\sum_{i\epsilon s_m}\left(y_{2i}/p_i^* - \sum_{i\epsilon s_m} y_{2i}\tilde{P}_i/p_i^*\right)^2\tilde{P}_i$$

and

$$\hat{\sigma}_{30}^2 = d \sum_{i \in s_m} \left(z_i / p_i^* - \sum_{i \in s_m} z_i \bar{P}_i / p_i^* \right)$$

$$\left(y_{2i} / p_i^* - \sum_{i \in s_m} y_{2i} \bar{P}_i / p_i^* \right) \bar{P}_i$$

are unbiased estimators of σ_0^2, σ_3^2 and σ_{30}, respectively where $d = m(N-1)/\{N(m-1)\}$.

Estimator for $V_{\text{opt}}(\hat{Y}_2 I\lambda)$

Thus for a given value of m (*i.e.*, λ), we can suggest an approximate unbiased estimator of $V_{\text{opt}}(\hat{Y}I\lambda)$ as,

$$V_{\text{opt}}(\hat{Y}I\lambda) = (1/\hat{V}_m + 1/\hat{V}_u)^{-1},$$

where $\hat{V}_m = \hat{V}(\hat{Y}_{2m}^* I\beta_0)$ and $\hat{V}_u$ = an unbiased estimator of $V(\hat{Y}_{2u}) = \{(N-u)/N(u-1)\} \sum_{i \in s_u} P_i' (y_{2i}/p_i - \hat{Y}_{2u})^2$.

Estimator for $V_{\min}(\hat{Y}_2)$

Putting suitable estimators for λ, ζ^* and σ_2^2 in the expression for $V_{\min}(\hat{Y}_2)$, we get an approximate unbiased estimator for $V_{\min}(\hat{Y}_2)$ as,

$$\hat{V}_{\min}(\hat{Y}_2) = k[1 - f + (1-\hat{\lambda})\hat{\zeta}^*/\hat{\lambda}]/\hat{\sigma}_2^2,$$

where

$$\hat{\zeta}^* = (1 - \hat{\delta}^2)\hat{\zeta}, \hat{\lambda} = \sqrt{\hat{\zeta}^*}/(1 + \sqrt{\hat{\zeta}^*}),$$

$$\hat{\delta} = \hat{\sigma}_{03}/(\hat{\sigma}_0^2 \hat{\sigma}_3^2)^{1/2}, \hat{\zeta}^* = \hat{\sigma}_3^2/\hat{\sigma}_2^2,$$

$$\hat{\sigma}_2^2 = \hat{\lambda}\hat{\sigma}_2^2(m) + (1-\hat{\lambda})\hat{\sigma}_2^2(u)$$

$\hat{\sigma}_2^2(m)$ = an approximate unbiased estimator of σ_2^2 based on the matched sample $s_m = \sum_{i \in s_m} (y_{2i}^2 \Delta_i / p_i^2) \bar{P}_i / p_i^* - \{\hat{Y}_{2m_2}^{*2} - \hat{V}_m\}$, $\hat{\sigma}_2^2(u)$ = an approximate unbiased estimator of σ_2^2 based on the un-matched sample $s_u = u(N-1)/\{N(u-1)\} \sum_{i \in s_u} P_i' (y_{2i} P_i'/p_i - \hat{Y}_{2u})^2$; k and f are as in (4).

Remark 2.2

Ideally one should estimate σ_2^2 through the optimum combination of $\hat{\sigma}^2(m)$ and $\hat{\sigma}^2(u)$ and in this case, the optimum combination will involve unknown parameters. To avoid this complexity, the simpler estimator $(\hat{\sigma}^2)$ of σ^2 has been suggested above.

2.2. Strategy 2

The population is supposed to consist of L strata with N_h as the unknown size of the h-th stratum ($h = 1, \ldots, L$; $\sum_h N_h = N$) stipulating that one can identify the stratum to which a unit belongs, as soon as its value is observed on the first occasion. On the first occasion, the initial sample s_1 of size n was selected by PPSWR method with normed size p_i attached to the i-th unit. Let n_h units of s_1, falling in the h-th stratum, be denoted as s_{1h}. Let $y_{1i}(h), y_{2i}(h)$ be respectively the value of the variate under study, of the i-th unit of the h-th stratum for the first and second occasions, and $z_i(h)$ be the corresponding size measure. On the second occasion, independent samples s_{mh}'s of sizes $m_h = m n_h / n$ (assumed an integer for every h), keeping $\sum_h m_h = m$ as fixed, are selected by the RHC sampling scheme with normed size $q_{hi}^* = [y_{1i}(h)/z_i(h)]/\sum_{i \in s_1} [y_{1i}(h)/z_i(h)]$ for the i-th unit of h-th stratum. The unmatched sample s_u was selected from the entire population by the RHC method with normed size measure p_i for the i-th unit as in strategy 1. The proposed estimators for Y_2, based on the matched-sample s_m, and the un-matched sample s_u are respectively as follows:

$$\hat{Y}_{2m} = \sum_h w_h \hat{Y}_{2m}(h); \; \hat{Y}_{2u} = \sum_{s_u} (y_{2i}/p_i) P_i' \tag{9}$$

where

$$\hat{Y}_{2m}(h) = \sum_{s_{mh}} r_i(h) Q_{hi} / (n_{1h} p_{hi} q_{hi}^* + c_h \sum_{s_{1h}} z_j(h) /$$

$$(n_{1h} p_{hj}), w_h = n_{1h}/n, p_{hj} = z_j(h)/Z,$$

$$r_i(h) = y_{2i}(h) - c_h y_{1i}(h),$$

Q_{hi} = sum of q_{hj}^* for the group containing i-th unit of the h-th stratum, that was formed for selection of the matched sample s_{mh} by RHC method. c_h's are constants chosen to minimize variance of $\hat{Y}_{2m}(h)$. Following Arnab (1991), the expression for variance of $\hat{Y}_{2m}$ is obtained as:

$$V(\hat{Y}_{2m}) = k_2 \sum_h \sum_{j=1}^{N_h} q_{hj} (r_{hj}/q_{hj} - R_h)^2 / P(h) + \sigma_2^2/n$$

where $k_2 = (n-m)/n, q_{hj} = y_{1j}(h)/y_1(h), Y_1(h) = \sum_{j=1}^{N_h} y_{1j}(h), N_h$ = population size of the h-th stratum, $P(h) = Z_h/Z, Z = \sum_{j=1}^{N_h} z_j(h)$.

The optimum value of c_h that minimizes $V(\hat{Y}_{2m})$ and the corresponding value of $V(\hat{Y}_{2m})$ comes out respectively as

$$\text{opt } c_h = c_h(0) = \delta_{h3} = \sum_{j=1}^{N_h} q_{hj} \alpha_{hj} \beta_{hj} / (\sigma_{h0} \sigma_{h3})$$

and $[1 + (n-m)\theta/m]\sigma_2^2/n$, where

$$\alpha_{hj} = y_{2j}(h)/q_{hj} - Y_2(h), \; \beta_{hj} = z_{hj}/q_{hj} - Z_h,$$

$$\sigma_{h3}^2 = \sum_{j=1}^{N_h} q_{hj} \alpha_{hj}^2, \sigma_{h0}^2 = \sum_{j=1}^{N_h} q_{hj} \beta_{hj}^2, Y_2(h) = \sum_{j=1}^{N_h} y_{2j}(h)$$

and $\theta = \sum_h (1 - \delta_h^2) \sigma_{h3}^2 / \{P_h \sigma_2^2\}$.

The proposed composite estimator for Y_2, the optimum proportion of matched sample and the expression for the minimum variance of the composite estimator $\hat{Y}_2$ are given respectively by

$$\hat{Y}_2 = \varphi \hat{Y}_{2m} + (1 - \varphi) \hat{Y}_{2u}$$

$$\text{opt}\,\lambda = \lambda_0 = [\theta - (1 - f)\sqrt{\theta}\sqrt{f^*}]/[\theta + f\sqrt{\theta}\sqrt{f^*} - 1]$$

$$V_{\min}(\hat{Y}_2) = k(1/\mu_0 - f)\,\sigma_2^2 / [1 + (\lambda_0/\mu_0)\sqrt{f^*}/\sqrt{\theta}]$$

$$= M_2 \text{ (say)}$$

where $\hat{Y}_{2m}$ and $\hat{Y}_{2u}$ are given in (9), $f^* = N/(N-1)$, $\mu_0 = 1 - \lambda_0$; k, f and σ_2^2 are given in (4).

3. EFFICIENCIES OF THE PROPOSED STRATEGIES

The proposed Strategy 1 is more efficient than the strategy proposed by Prasad and Graham (1994) in the sense of yielding smaller minimum variance, as $\delta^2 \leq 1$. Efficiency of the Strategy 1 increases as δ, the correlation between y_{2i}/q_i and z_i/q_i increases. The efficiency of the Strategy 1 and Prasad and Graham's (1994) strategy increases as ζ decreases. The value of $\zeta = \sigma_3^2/\sigma_2^2$ depends on the magnitudes of σ_3^2 and σ_2^2. σ_3^2 will be smaller (greater) than σ_2^2 if the proportionality of y_{2i} on y_{1i} is higher (lower) than that of y_{2i} on z_i. Obviously, Strategy 1 can be used in practice when a good guess value of β is available from the past surveys. If the difference estimator is used in Strategy 1 instead of the regression estimator mentioned in Remark 2.1, then the proposed Strategy 1 fares better than that of Prasad and Graham (1994) whenever $\delta > \frac{1}{2}\sigma_0/\sigma_3$. Strategy 1 fares better or worse than Chotai's (1974) strategy according to $\zeta^* = (1 - \delta^2)\zeta <$ or $> (1 - \delta^{*2})$. Here, δ^* may be regarded as a correlation coefficient between y_{2i}/p_i and y_{1i}/p_i. In particular, if z_i's, are constant, then δ^* becomes the simple correlation coefficient between y_{1i}'s and y_{2i}'s. The expression for the minimum variance M_2 for Strategy 2 is complex and does not yield any simple comparison with the other strategies described here. However, we note that the efficiency of the Strategy 2 increases as the stratum correlation δ_{h3} increases. Following numerical examples based on the live data reveals that the proposed Strategy 2 fares better than Strategy 1 and also the alternatives proposed by Prasad and Graham (1994) and Chotai (1974).

For numerical comparisons, three data sets are considered. One of them (will be called Population 1) was considered by Prasad and Graham (1994) which relates to the area under wheat in 1937 (y_2) and 1936 (y_1) and cultivated area (z) for a set of 34 villages in India, compiled by Sukhatme and Sukhatme (1970). The population 1 is stratified in two strata in accordance with area under wheat in 1936 less than or more than 200 acres. Parameters for this population are: $N = 34$, $N_1 = 20$, $N_2 = 14$, $\delta^* = .7635$, $\delta = .3638$, $\zeta = .3811$, $\theta = .2436$. The Population 2 comprises of production of cereals in South America for the years 1980 (z), 1988 (y_1) and 1989 (y_2), compiled from The Statistical year book, United Nations (1988/89). The population is stratified in two strata considering 1988 production of more or less than 570 (thousand metric tons). The parameters for this population 2 are: $N = 19$, $N_1 = 7$, $N_2 = 12$, $\delta^* = -.6939$, $\delta = .7666$, $\zeta = 1.1478$, $\theta = .3681$. The population 3 compiled by Singh and Chaudhuri (1986) relates to the area under wheat in hector during 1979-80 (y_2) and 1978-79 (y_1) and total cultivated area in 1978-79 (z) of 16 villages of Meerut District. The parameters for the population 3 are: $N = 16$, $N_1 = 9$, $N_2 = 7$, $\delta^* = .7729$, $\delta = .1057$, $\zeta = .3965$, $\theta = .2827$.

The following table shows relative efficiencies of the proposed Strategies 1, 2 and the one proposed by Prasad and Graham (1994) with respect to Chotai (1974) which are respectively denoted by $E_1 = V_c/M_1$, $E_2 = V_c/M_2$ and $E_3 = V_c/V_{PG}$.

Table 1
Efficiencies of the Strategies

f	Population 1			Population 2			Population 3		
	E_1	E_2	E_3	E_1	E_2	E_3	E_1	E_2	E_3
.05	1.0463	1.1033	1.0181	1.0196	1.0850	.8262	1.0053	1.0864	1.0030
.10	1.0479	1.0895	1.0187	1.0202	1.0711	.8212	1.0055	1.0711	1.0031
.15	1.0496	1.0776	1.0194	1.0209	1.0579	.8172	1.0057	1.0577	.0033
.20	1.0514	1.0683	1.0200	1.0216	1.0519	.8123	1.0058	1.0469	1.0034
.25	1.0533	1.0622	1.0208	1.0224	1.0490	.8071	1.0061	1.0396	1.0035
.30	1.0554	1.0604	1.0216	1.0232	1.0530	.8017	1.0063	1.0368	1.0036

From the above table, we note that in all the three populations, Strategy 2 fares better than the others. It is also worth noting that both the proposed strategies fare better than those of Chotai (1974) and Prasad and Graham (1994). For the population 1, $\zeta = .3811$ which is quite favourable for Prasad and Graham's (1994) strategy, hence for the proposed Strategy 1. Both Prasad and Graham's strategy and Strategy 1, performed better than Chotai's (1974) strategy. For the population 2, $\zeta = 1.1478$ which is high and unfavourable for Prasad and Graham's (1994) strategy, but $\delta = .7666$ is quite favourable to Strategy 1. Hence, for the population 2, Prasad and Graham's strategy becomes less efficient than that of Chotai (1974), but the proposed Strategy 1 remains better. For the population 3, $\zeta = .3965$ which is quite favourable for Prasad and Graham (1994) but at the same time $\delta^* = .7729$ and this (δ^*) favours Chotai (1974). In fact Chotai's (1974) strategy is marginally inferior to Prasad and Graham's (1994) strategy but the proposed Strategy 2 remains better than both. It should be noted that the examples shown here are quite unusual in the

sense that they present low correlation between y_2 and z (in example 1, $\delta = .3638$ and in example 3, $\delta = .1057$) and there is a negative correlation between y_2 and y_1 ($\delta^* = -.6939$) in example 2. The correlations δ and δ^* are expected to be high and positive. Hence, further investigation is needed to compare the performances of the present strategies with suitable data.

Table 2
Sensitivity of Efficiency $E^* = V_{PG}/M_{\tilde{\beta}}$

$\lvert v \rvert$	.05	.10	f .15	.20	.25	.30
			Population 1			
0	1.028	1.029	1.030	1.031	1.032	1.033
.2	1.027	1.027	1.028	1.029	1.031	1.032
.4	1.023	1.024	1.027	1.026	1.027	1.028
.6	1.017	1.108	1.019	1.019	1.020	1.021
.8	1.010	1.010	1.010	1.011	1.011	1.011
1.0	1.000	1.000	1.000	1.000	1.000	1.000
1.2	.989	.988	.988	.988	.988	.987
1.4	.976	.976	.975	.974	.973	.972
			Population 2			
0	1.234	1.241	1.249	1.257	1.266	1.278
.2	1.219	1.227	1.233	1.241	1.249	1.258
.4	1.180	1.186	1.191	1.197	1.204	1.211
.6	1.125	1.128	1.133	1.137	1.141	1.146
.8	1.063	1.065	1.067	1.068	1.070	1.073
1.0	1.000	1.000	1.000	1.000	1.000	1.000
1.2	.939	.938	.936	.935	.933	.931
1.4	.883	.880	.877	.875	.871	.869
			Population 3			
0	1.002	1.002	1.004	1.003	1.003	1.003
.2	1.002	1.002	1.002	1.002	1.003	1.002
.4	1.002	1.002	1.002	1.002	1.002	1.002
.6	1.001	1.002	1.002	1.002	1.002	1.001
.8	1.001	1.001	1.001	1.001	1.001	1.001
1.0	1.000	1.000	1.000	1.000	1.000	1.000
1.2	.999	.999	.999	.999	.999	.999
1.4	.998	.997	.998	.998	.998	.998

To study the effect of departure of the optimum value of $\beta = \beta_0$ when some guess value of β is used in Strategy 1, one may consider sensitivity of efficiency of $\hat{Y}_2$ for the Strategy 1 for different choices of β, following Prasad and Srivenkataramana (1980). The minimum variance of $\hat{Y}_2$ for the Strategy 1 when some guess value of $\beta_0 = \tilde{\beta}$ is used, produces

$$V_{\min}(\hat{Y}_2 \mid \tilde{\beta}) = k(1 - f + \sqrt{\zeta^{**}})\,\sigma_2^2/2 = M_{\tilde{\beta}} \qquad (9)$$

where $\zeta^{**} = [1 - (1 - v^2)\,\delta^2]\,\zeta$ and $v = 1 - \tilde{\beta}/\beta_0$.

From (9), we note that the proposed Strategy 1 with the guess value $\tilde{\beta}$ fares better or worse than Prasad and Graham's (1994) strategy according to $|v| < 1$ or $|v| > 1$. Similarly, the proposed Strategy 1 with $\beta = \tilde{\beta}$ performs better or worse than Chotai's (1974) strategy according to $v^2 >$ or $< (1 - 1/\delta^2)(1 - 1/\zeta)$. Table 2 proceeds sensitivity E^* of the estimator $\hat{Y}_2$ compared to Prasad and Graham's (1994) strategy where $E^* = V_{PG}/M_{\tilde{\beta}}$. From the Table 2, the loss with $v > 1$ is likely to be more than the gain with $v < 1$ for population 1 and population 3 but the situation is reverse for population 2.

CONCLUSION

In sampling over two occasions, one should utilize data collected on the first occasion to get an efficient estimator for the population total on the second occasion. Chotai (1974) used data collected on the first occasion at the stage of estimation, while Prasad and Graham did so at the stage of selection (and hence estimation) of the matched sample. In this article, two strategies have been proposed. The first one utilizes data collected at the first occasion for the selection of the matched sample similar to Prasad and Graham and formation of a regression estimator as determined by Chotai (1974). These make Strategy 1 more efficient than that of Prasad and Graham. The proposed Strategy 2 utilized first occasion values as a stratification variable, measure of size for the selection of the matched sample for the second occasion, and formation of a regression type estimator involving auxiliary variable (z), available on the first occasion. Intuitively one should expect the proposed Strategy 2 to perform better than the others mentioned here, but no theoretical result was established due to the complexity of the expression for the minimum variance of the proposed estimator. However, superiority of the Strategy 2 was established through numerical data.

ACKNOWLEDGEMENTS

The author is grateful to the referee, Associate Editor and the Editor for their valuable comments that substantially improved the earlier version of this paper. This work was supported by the FRD, South Africa.

REFERENCES

ARNAB, R. (1991). On sampling over two occasions using varying probabilities. *Journal of the Indian Society of Agricultural Statistics*, 43(3), 282-290.

AVADHANI, M.S., and SUKHATME, B.V. (1970). A comparison of two sampling procedures with an application to successive sampling. *Applied Statistics*, 19, 251-259.

CHOTAI, J. (1974). A note on the Rao-Hartley-Cochran method for PPS sampling over two occasions. *Sankhyā*, Series C, 36, 173-180.

PRASAD, N.G.N., and SRIVENKATARAMANA, T. (1980). A modification to the Horvitz-Thompson estimator under the Midzuno sampling scheme. *Biometrika*, 67, 709-711.

PRASAD, N.G.N., and GRAHAM, J.E. (1994). PPS sampling over two occasions. *Survey Methodology*, 20, 59-64.

RAJ, D. (1965). On sampling over two occasions with probabilities proportional to size. *Annals of Mathematical Statistics*, 36, 327-330.

RAO, J.N.K., HARTLEY, H.O., and COCHRAN, W.G. (1962). On a simple procedure of unequal probability sampling without replacement. *Journal of the Royal Statistical Society*, 24, 482-491

SINGH, D., and CHAUDHURI, F.S. (1986): *Theory and Analysis of Sample Survey Designs*. India: Wiley Eastern Limited, 166.

SINGH, M.P. (1967). The relative efficiency of some two-phase sampling schemes. *Annals of Mathematical Statistics*, 38, 937-940.

SUKHATME, P.V., and SUKHATME, B.V. (1970). *Sampling Theory of Surveys with Applications*. Ames, Iowa: Iowa State University Press, 185.

UNITED NATIONS (1992). *Statistical Year Book*, (1988/89). New York: United Nations, 356.

Survey Methodology, December 1998
Vol. 24, No.2, pp. 193-201
Statistics Canada

Confidence Intervals for Proportions With Small Expected Number of Positive Counts Estimated From Survey Data

EDWARD L. KORN and BARRY I. GRAUBARD[1]

ABSTRACT

In the nonsurvey setting, "exact" confidence intervals for proportions calculated using the binomial distribution are frequently used instead of intervals based on approximate normality when the number of positive counts is small. With complex survey data, the binomial intervals are not applicable, so intervals based on the assumed approximate normality of the sample-weighted proportion are used, even if the number of positive counts is small. We propose a simple modification of the binomial intervals to be used in this situation. Limited simulations are presented that show the coverage probability of the proposed intervals is superior to that of the normality-based intervals, logit-transform intervals, and intervals based on a Poisson approximation. Applications are given involving the prevalence of Human Immunodeficiency Virus (HIV) based on data from the third National Health and Nutrition Examination Survey, and the proportion of users of cocaine based on data from the Hispanic Health and Nutrition Examination Survey.

KEY WORDS: Binomial confidence interval; Exact confidence interval; Logit transformation; Poisson confidence interval.

1. INTRODUCTION

With complex survey data, the typical construction of a $1 - \alpha$ level confidence interval for a proportion of positive counts for a 0-1 variable is

$$\hat{p} \pm t_d\,(1 - \alpha/2)\,[\hat{\text{var}}(\hat{p})]^{1/2} \tag{1.1}$$

where $\hat{p}$ is the sample-weighted estimator of the proportion, $\hat{\text{var}}(\hat{p})$ is the variance estimator of $\hat{p}$, and $t_d(1 - \alpha/2)$ is the $1 - \alpha/2$ quantile of a t distribution with d degrees of freedom. The estimator $\hat{\text{var}}(\hat{p})$ is computed using linearization or a replication method to reflect the sample design, including the fact that $\hat{p}$ is a sample-weighted estimator. By complex survey data, we mean data obtained from a multistage design with stratified selection of clusters at the first stage. For such a sample design, d is usually taken to be equal to the number of sampled clusters minus the number of strata (Korn and Graubard 1990). The confidence interval (1.1), which we shall refer to as the "linear interval", is based on the assumption that $\hat{p}$ is approximately normally distributed. Under various reasonable asymptotics, this is known to be true (Krewski and Rao 1981). The use of the t quantile rather than a normal-distribution quantile in (1.1) is based on empirical evidence (Frankel 1971, ch. 7), and it can also be formally justified using strong assumptions (Korn and Graubard 1990).

When the expected number of positive counts is small, the approximate normality of $\hat{p}$ breaks down (Cochran 1977, p. 58). For a simple random sample (or in the nonsurvey setting), one can avoid the normality assumption by using the Clopper and Pearson (1934) confidence interval based on the binomial distribution; see Vollset (1993) for a complete discussion of confidence intervals for proportions in the nonsurvey setting. When x positive responses are seen in a simple random sample of size n, the Clopper-Pearson $1 - \alpha$ level confidence interval $(p_L(x, n),\ p_U(x, n))$ can be expressed as (Johnson, Kotz and Kemp 1993, p. 130):

$$p_L(x, n) = \frac{v_1 F_{v_1, v_2}(\alpha/2)}{v_2 + v_1 F_{v_1, v_2}(\alpha/2)}$$

$$p_U(x, n) = \frac{v_3 F_{v_3, v_4}(1 - \alpha/2)}{v_4 + v_3 F_{v_3, v_4}(1 - \alpha/2)} \tag{1.2}$$

where $v_1 = 2x, v_2 = 2\,(n - x + 1), v_3 = 2(x + 1), v_4 = 2\,(n - x)$ and $F_{d_1, d_2}(\beta)$ is the β quantile of an F distribution with d_1 and d_2 degrees of freedom. For one-sided confidence bounds, α is used instead of $\alpha/2$ in the above expressions. For a simple random sample, these intervals are known to have coverage probability greater than or equal to their nominal level, regardless of the expected number of positive counts. They are sometimes referred to as "exact" confidence intervals; we shall refer to them as the "binomial intervals".

In this paper we suggest a simple modification to the binomial intervals to make them applicable for a proportion estimated from complex survey data. We are especially interested in the situation when the expected number of positive counts is small. Many survey analysts would not

[1] Edward L. Korn, Biometric Research Branch, EPN-739, National Cancer Institute, Bethesda, MD 20892, U.S.A.; Barry I. Graubard, Biostatistics Branch, National Cancer Institute, Bethesda, MD 20892, U.S.A.

present estimated proportions in this situation, since they are unreliable. For example, applying the relative-standard-error criterion for presenting proportions in the 1996 National Household Survey on Drug Abuse (SAMHSA 1998), the estimated proportion of women using cocaine in Table 7 would not be presented. We believe such proportions can provide valuable information, but that their lack of precision needs to be explicitly stated by presenting confidence intervals. In section 2, we define our proposed confidence intervals and define intervals based on a logit transformation and the Poisson distribution that have been suggested in the literature. Simulation results are presented in section 3 that compare the intervals. We find that the proposed intervals behave well in terms of coverage probability of the true proportion and in terms of their average width. Two applications are given in section 4 involving large surveys, but where the number of positive counts is expected to be small. We end with a discussion of some related work that constructs confidence intervals that are guaranteed to attain their nominal coverage probability regardless of the population configuration of counts.

2. PROPOSED AND OTHER CONFIDENCE LIMITS

For a $1-\alpha$ level confidence interval based on a sample of size n, first define the effective sample size by

$$n^* = \frac{\hat{p}(1-\hat{p})}{\hat{\text{var}}(\hat{p})} \tag{2.1}$$

and the degrees-of-freedom adjusted effective sample size by

$$n_{df}^* = \frac{\hat{p}(1-\hat{p})}{\hat{\text{var}}(\hat{p})}\left(\frac{t_{n-1}(1-\alpha/2)}{t_d(1-\alpha/2)}\right)^2 \tag{2.2}$$

Both n^* and n_{df}^* are set equal to n when $\hat{p}=0$. The proposed limits substitute n_{df}^* for n, and $\hat{p}n_{df}^*$ for x in (1.2), *viz.* $p_L(\hat{p}n_{df}^*, n_{df}^*)$ and $p_U(\hat{p}n_{df}^*, n_{df}^*)$. (When n is large, the $1-\alpha/2$ quantile of a normal distribution can be used in place of $t_{n-1}(1-\alpha/2)$ in (2.2).) For estimating a confidence interval for a proportion on a subdomain of the population, the sample size n is taken to be equal to the sample size restricted to the subdomain.

A heuristic justification for this procedure is as follows. The effective sample size (2.1) is n divided by an estimator of the design effect of the survey. This seems to be a reasonable way to incorporate the additional variability of $\hat{p}$ due to the complex sampling. For confidence interval construction, the variability of the variance estimator is also important. The second fraction in (2.2) takes into account the fact that $\hat{\text{var}}(\hat{p})$ will typically be more variable than a variance estimator that would be used for simple random sampling. If d is large, then this factor is close to one and unneeded. For small d and large n and $\hat{p}n_{df}^*$, we would like the proposed interval to be close to the interval (1.1), which is appropriate in this situation. Using the fact that $F_{u,w}(\beta) \approx 1 + z(\beta)\sqrt{2(1/u + 1/w)}$ for large u and w (Johnson and Kotz 1970, p. 81), this is true, *i.e.*, $\hat{p} - p_L(\hat{p}n_{df}^*, n_{df}^*) \approx p_U(\hat{p}n_{df}^*, n_{df}^*) - \hat{p} \approx t_d(1-\alpha/2)[\hat{\text{var}}(\hat{p})]^{1/2}$.

A procedure closely related to the proposed procedure was developed by Breeze (1990) for use in the U.K. General Household Survey. This procedure is based on the simple-random-sampling $1-\alpha$ confidence interval $(po_L(x), po_U(x))$ for a Poisson random variable x, which can be expressed as (Johnson *et al.* 1993, p. 171):

$$po_L(x) = 0.5\,\chi^2_{\nu_1}(\alpha/2) \text{ and } po_U(x) = 0.5\,\chi^2_{\nu_2}(1-\alpha/2)$$

where $\nu_1 = 2x$, $\nu_2 = 2(x+1)$, and $\chi^2_\nu(\beta)$ is the β quantile of a χ^2 distribution with ν degrees of freedom. With complex survey data, the confidence interval is taken to be $(po_L(\hat{p}n^*)/n^*, po_U(\hat{p}n^*)/n^*)$.

A third procedure for confidence interval construction is based on a logit transform. For a $1-\alpha$ level confidence interval, the interval is

$$\left(\frac{1}{1+\exp(-LLOGIT)}, \frac{1}{1+\exp(-ULOGIT)}\right)$$

where

$$LLOGIT = \log\frac{\hat{p}}{1-\hat{p}} - t_d(1-\alpha/2)\frac{[\hat{\text{var}}(\hat{p})]^{1/2}}{\hat{p}(1-\hat{p})} \tag{2.3}$$

and

$$ULOGIT = \log\frac{\hat{p}}{1-\hat{p}} + t_d(1-\alpha/2)\frac{[\hat{\text{var}}(\hat{p})]^{1/2}}{\hat{p}(1-\hat{p})} \tag{2.4}$$

These intervals, with a normal-distribution quantile instead of a t distribution quantile, were suggested for use with the 1996 National Household Survey on Drug Abuse (SAMHSA 1998). When $\hat{p}=0$, in the nonsurvey setting one might add a small constant to the observed number of events and nonevents, *e.g.*, 1/2, to be able to calculate the logit-transform confidence interval (Agresti 1990, pp. 249-250). In the present setting, when $\hat{p}=0$, we set the confidence interval equal to the binomial interval $(p_L(0,n), p_U(0,n))$.

In applications where it is known before sampling that the (true) design effect will be greater than 1, various modifications of the above procedures are possible. For our proposal, we recommend in this situation truncating the degrees-of-freedom adjusted effective sample size at n. That is, if n_{df}^* is greater than n, we set its value to n, and define the lower and upper confidence limits to be $p_L(\hat{p}n, n)$ and $p_U(\hat{p}n, n)$. For the Breeze intervals, one could set n^* to be n if $n^* > n$. For the linear or logit intervals, one can use the simple-random-sampling variance estimator $\hat{p}(1-\hat{p})/n$ in place of $\hat{\text{var}}(\hat{p})$ in (1.1), (2.3) and (2.4) if $n^* > n$; see SAMHSA (1998) for additional truncation suggestions. The justification of these truncation

procedures is that the design effect may be estimated to be less than one because of instability of the variance estimator $\hat{\text{var}}(\hat{p})$. This type of instability may be especially large because $\hat{p}$ is small (SAMHSA 1998). The effect of these truncation procedures is to make the confidence intervals wider and more conservative. In theory, one could also adjust the estimated effective sample sizes when it is known before the sampling that the (true) design effect is less than one. However, to be conservative, we do not recommend doing this.

Our focus in this paper is on confidence intervals for the "superpopulation" probability that the outcome $Y = 1$ rather that the finite-population proportion. That is, the target parameter is $p = \sum_{u=1}^{N} p_u / N$ rather than $P = \sum_{u=1}^{N} Y_u / N$, where Y_u has a Bernoulli distribution with parameter p_u, and N is the population size. The simulated coverage probabilities given in the next section therefore refer to coverage of p. With this target parameter in mind, we do not use finite-population correction factors when estimating $\hat{\text{var}}(\hat{p})$ for use in (2.2); additional adjustments to the design-based variance $\hat{\text{var}}(\hat{p})$ for superpopulation inference are not pursued here (Korn and Graubard 1998). A referee suggests the possibility of a model-based approach to estimating a confidence interval for p. However, in our limited experience, such approaches yield estimators similar to weighted estimators and offer no advantages for inference (Pfeffermann and LaVange 1989; Graubard and Korn 1996).

If one were interested in a confidence interval for P, we would recommend using the proposed intervals but with $\hat{\text{var}}(\hat{p})$ in (2.2) containing the finite-population correction factors. A confidence interval for $\sum_{u=1}^{N} Y_u$ could be obtained by multiplying the ends of the confidence interval for P by N, if known, or by an estimator $\hat{N}$ of N, if not known. (In theory, one could account for the variability of $\hat{N}$, but this additional variability will be small.) An alternative approach for estimating a confidence interval for P would be to modify the usual limits (Guenther 1983) appropriate for a simple random sample (based on the hypergeometric distribution) similarly to the way the proposed intervals modify the binomial intervals.

3. SIMULATIONS

The main simulation results are presented in Tables 1-5. Table 1 presents the results of simulations in which datasets of 32 clusters, each with sample size 100, were simulated. Within cluster i, the number of positive events was simulated with a binomial distribution with probability parameter p_i. In Table 1, we refer to the $\{p_i, i = 1, ..., 32\}$ as the cluster probabilities. For the top third of the table, the cluster probabilities are taken to be the constant $p = .1, .02,$

Table 1
Simulated Lack of Coverage (Percent) of Upper and Lower One-sided 95% Confidence Bounds for Sample Design of 32 Clusters and 100 Observations Per Cluster; Sample Weights are 1 Or 10 with Probability 1/2 (Noninformative Sample Weights)

Distribution of cluster proportions[a]	Overall proportion	Expected number positive	Method of calculating confidence bounds							
			Linear		Logit		Breeze		Proposed	
			Lower	Upper	Lower	Upper	Lower	Upper	Lower	Upper
(1)										
.1	.1	320	4.6	5.5	5.3	4.6	4.5	4.1	4.8	4.4
.02	.02	64	3.4	7.1	5.2	4.6	4.5	4.7	4.2	4.4
.01	.01	32	2.9	8.0	5.4	4.5	4.4	4.5	4.0	4.1
.0025	.0025	8	1.6	9.5	5.5	1.8	3.6	2.2	3.3	1.8
(1/2, 1/2)										
.05, .15	.1	320	4.3	5.8	5.5	4.3	4.3	3.8	4.7	4.1
.01, .03	.02	64	3.1	7.5	5.2	4.8	4.3	4.8	4.0	4.5
.005, .015	.01	32	2.7	8.6	5.2	4.7	4.1	4.9	3.7	4.4
.00125, .00375	.0025	8	1.5	9.9	5.4	2.0	3.4	2.3	3.1	2.0
(3/4, 1/4)										
.05, .25	.1	320	3.1	7.8	4.7	5.6	3.4	5.0	3.6	5.3
.01, .05	.02	64	2.7	8.6	5.1	5.3	4.0	5.4	3.7	5.0
.005, .025	.01	32	2.2	9.8	5.0	5.3	3.7	5.5	3.3	5.0
.00125, .00625	.0025	8	1.3	10.7	5.3	2.2	3.3	2.5	3.0	2.2

(a) Fractions in parentheses are the probabilities that the cluster proportions have the stated value.

Table 2
Simulated Lack of Coverage (Percent) of Upper and Lower One-sided 95% Confidence Bounds for Sample Design of 32 Clusters and 100 Observations Per Cluster; Informative Sample Weights are 1 or 10 (See Text)

Distribution of cluster proportions[a]	Overall weighted proportion	Expected number positive	Method of calculating confidence bounds							
			Linear		Logit		Breeze		Proposed	
			Lower	Upper	Lower	Upper	Lower	Upper	Lower	Upper
(1)										
.1	.1	191.0	4.3	5.9	5.1	4.9	4.2	4.4	4.6	4.6
.02	.02	36.9	3.3	7.3	5.3	4.3	4.4	4.4	4.1	4.1
.01	.01	18.4	2.8	8.7	5.5	4.0	4.3	4.3	3.9	3.7
.0025	.0025	4.6	1.3	18.7	6.1	4.8	3.2	4.8	2.8	4.8
(1/2, 1/2)										
.05, .15	.1	191.0	5.0	5.0	6.4	3.7	5.1	3.2	5.4	3.4
.01, .03	.02	36.9	3.0	7.9	5.4	4.5	4.3	4.6	4.0	4.3
.005, .015	.01	18.4	2.5	9.2	5.4	4.2	4.1	4.4	3.7	3.9
.00125, .00375	.0025	4.6	1.3	19.0	6.1	4.9	3.2	4.9	2.8	4.9
(3/4, 1/4)										
.05, .25	.1	191.0	4.7	5.7	7.1	4.1	5.1	3.6	5.5	3.8
.01, .05	.02	36.9	2.6	8.9	5.2	5.2	4.0	5.3	3.7	4.9
.005, .025	.01	18.4	2.1	10.1	5.3	4.8	3.8	5.1	3.4	4.5
.00125, .00625	.0025	4.6	1.2	19.8	5.9	5.3	3.2	5.3	2.8	5.3

(a) Fractions in parentheses are the probabilities that the cluster weighted proportions have the stated value.

Table 3
Simulated Lack of Coverage (Percent) of Upper and Lower One-sided 95% Confidence Bounds for Sample Design of 32 Clusters and 100 Observations Per Cluster; Unweighted Analyses

Distribution of cluster proportions[a]	Overall proportion	Expected number positive	Method of calculating confidence bounds							
			Linear		Logit		Breeze		Proposed	
			Lower	Upper	Lower	Upper	Lower	Upper	Lower	Upper
(1)										
.1	.1	320	5.0	4.9	5.7	4.2	4.9	3.8	5.2	4.1
.02	.02	64	3.8	6.3	5.2	4.5	4.7	4.8	4.4	4.4
.01	.01	32	3.5	6.8	5.6	4.4	4.7	4.4	4.3	4.0
.0025	.0025	8	2.5	8.8	5.6	3.8	4.1	3.9	3.9	3.9
(1/2, 1/2)										
.05, .15	.1	320	4.5	5.6	5.6	4.2	4.5	3.7	4.8	4.0
.01, .03	.02	64	3.4	7.0	5.1	4.8	4.5	4.9	4.1	4.6
.005, .015	.01	32	3.0	7.6	5.2	4.8	4.4	4.8	3.9	4.4
.00125, .00375	.0025	8	2.2	9.2	5.4	4.3	3.8	4.3	3.5	4.3
(3/4, 1/4)										
.05, .25	.1	320	3.3	7.7	4.8	5.6	3.5	5.1	3.7	5.3
.01, .05	.02	64	2.9	8.1	5.1	5.2	4.1	5.3	3.8	4.9
.005, .025	.01	32	2.5	9.2	4.9	5.6	3.9	5.6	3.5	5.2
.00125, .00625	.0025	8	2.0	10.4	5.3	5.1	3.8	5.1	3.3	5.1

(a) Fractions in parentheses are the probabilities that the cluster proportions have the stated value.

Table 4
Simulated Lack of Coverage (Percent) of Upper and Lower One-sided 95% Confidence Bounds for Sample Design of 32 Clusters and 10 Observations Per Cluster; Sample Weights are 1 or 10 with Probability 1/2
(Noninformative Sample Weights)

Distribution of cluster proportions[a]	Overall proportion	Expected number positive	Method of calculating confidence bounds							
			Linear		Logit		Breeze		Proposed	
			Lower	Upper	Lower	Upper	Lower	Upper	Lower	Upper
(1)										
.2	.2	64	4.0	6.6	5.2	4.7	3.1	4.7	4.2	4.3
.1	.1	32	3.2	7.8	5.3	4.4	3.6	3.8	3.9	4.0
.025	.025	8	1.7	10.2	5.5	2.1	3.4	2.1	3.2	2.4
(1/2, 1/2)										
.1, .3	.2	64	3.6	7.0	5.0	4.9	2.8	3.4	3.9	4.4
.05, .15	.1	32	3.0	8.1	5.1	4.6	3.4	4.0	3.7	4.2
.0125, .0375	.025	8	1.6	10.6	5.4	2.1	3.3	2.1	3.1	2.5
(3/4, 1/4)										
.1, .5	.2	64	3.1	7.8	4.6	5.3	2.4	3.9	3.3	4.8
.05, .25	.1	32	2.5	9.2	4.8	5.2	3.0	4.6	3.3	4.8
.0125, .0625	.025	8	1.5	11.5	5.3	2.4	3.2	3.5	3.0	2.8

(a) Fractions in parentheses are the probabilities that the cluster proportions have the stated value.

Table 5
Simulated Lack of Coverage (Percent) of Upper and Lower One-sided 95% Confidence Bounds for Sample Design of 32 Clusters and 10 or 100 Observations Per Cluster with Probability 1/2; Sample Weights are 1 or 10 with Probability 1/2
(Noninformative Sample Weights)

Distribution of cluster proportions[a]	Overall proportion	Expected number positive	Method of calculating confidence bounds							
			Linear		Logit		Breeze		Proposed	
			Lower	Upper	Lower	Upper	Lower	Upper	Lower	Upper
(1)										
.1818	.1818	320	5.1	6.0	5.7	5.2	4.2	4.1	5.2	5.0
.0364	.0364	64	4.1	7.6	5.7	5.2	5.0	5.2	4.8	4.9
.0182	.0182	32	3.4	8.5	5.7	5.0	4.7	5.1	4.4	4.7
.0045	.0045	8	2.0	12.7	5.9	3.4	4.0	4.3	3.6	3.8
(1/2, 1/2)										
.0909, .2727	.1818	320	5.0	6.4	6.1	4.8	4.2	3.6	5.2	4.4
.0182, .0545	.0364	64	3.9	8.1	6.0	5.1	4.9	5.0	4.7	4.8
.0091, .0273	.0182	32	3.1	9.3	5.8	5.2	4.5	5.3	4.2	4.9
.0023, .0068	.0045	8	1.8	13.2	5.9	3.6	3.9	4.5	3.5	4.0
(3/4, 1/4)										
.0909, .4545	.1818	320	3.1	9.9	4.6	7.6	2.5	6.3	3.3	7.1
.0182, .0909	.0364	64	2.8	10.9	5.3	7.3	3.9	7.3	3.7	7.0
.0091, .0455	.0182	32	2.4	11.5	5.4	6.8	3.9	6.9	3.6	6.5
.0023, .0114	.0045	8	1.6	14.5	5.7	4.0	3.7	5.0	3.3	4.4

(a) Fractions in parentheses are the probabilities that the cluster weighted proportions have the stated value.

.01, or .0025, corresponding to an expected number of positive events equal to 320, 64, 32, or 8 out of the sample size of 3200. For the middle third of the table, the cluster probabilities are taken to be $p/2$ with probability 1/2 or $3p/2$ with probability 1/2, with p as in the first third of the table. Varying the p_i across the clusters induces an intracluster correlation among the observations. For the middle third of the table, these correlations (ignoring the sample weights) are .00278, .0051, .0025 and .0006 corresponding to the expected number of positive events being 320, 64. 32, or 8, respectively. For the bottom third of the table, the cluster probabilities are taken to be $p/2$ with probability 3/4 or $5p/2$ with probability 1/4, corresponding to intraclass correlations of .0833, .0153, .0076 and .0019. For all simulations in Table 1, sample weights of 1 or 10 are randomly assigned with probability 1/2 to each observation (noninformative weights).

The results presented in Table 1 are appropriate for one-sided 95% upper and lower confidence limits; ideally the lack-of-coverage percentages in the table should be less than or equal to the nominal value of 5.0. The results are also relevant for two-sided 90% confidence intervals, for which ideally both the upper and lower values in the table should both be ≤ 5.0 (Jennings 1987). For each line of the table, 100,000 datasets were simulated using the random number generator in SAS (1990, p. 631) to estimate the probabilities of noncoverage of the confidence limits.

For the linear confidence bounds, the upper confidence limit falls below the true value more than the 5% nominal level. Somewhat surprisingly, this is true even with as large as an expected 320 positive counts, especially with positive intracluster correlation (middle and bottom third of the table). For the logit-transform confidence bounds, the noncoverage appears slightly higher than the nominal level, especially for the lower limits. Both the Breeze and proposed confidence bounds appear generally conservative. Simple-random-sampling binomial limits are not appropriate for the cases simulated in Table 1 because of the sample weights and the intracluster correlation (in the bottom two-thirds of the table). This can be demonstrated by noting that the lack of coverage for both the upper and lower binomial bounds are greater than 8% for all the cases considered in the table (results not shown).

As it is slightly complicated to discuss confidence interval "lengths" for one-sided bounds, we restrict discussion to the lengths of the two-sided 90% confidence intervals. Over all the simulations presented in Table 1, the Breeze and proposed intervals are 3.3% and 4.9% wider on average than the logit-transform intervals.

Table 2 presents simulation results for the same setup as Table 1 except that the sample weights were taken to be informative. This was done by setting the sample weight to be 10 with probability 2/3 if the event was positive and with probability 1/3 if the event was not positive, otherwise the weight was set to 1. The probability that an event was positive in each cluster was adjusted downwards so that the overall weighted proportions were the same as in Table 1. The results in Table 2 look similar to those in Table 1 except the linear and logit intervals tend to have worse coverage probabilities.

Table 3 presents simulation results for the same setup as Table 1 except the analysis is unweighted. The results are very similar to the Table 1 results. Since the top third of Table 3 corresponds to no intracluster correlation, one could also use the simple-random-sampling binomial limits there. Averaging over the four situations in this third of the table, the proposed limits are 2.5% wider that the binomial limits (results not shown). As the true design effect is 1.0 in the top third of Table 3, these simulations can be used to examine the effect of truncation of n_{df}^* in the proposed procedure. (Truncation is uncommon in the simulations in Table 1, since the true design effects there are all >1.) Simulation using the proposed procedure with truncation lead to wider more conservative intervals than for the proposed intervals in the top third of Table 3. Averaging over the four situations considered, the proposed limits with truncation are 4.0% wider than the proposed limits (results not shown for truncated limits).

Table 4 presents simulation results for the same setup as Table 1 except 10 rather than 100 observations are simulated within each cluster. The results are very similar to Table 1 when one compares simulations with the same expected number of positive events. The one exception is the increased conservativeness of the Breeze intervals as compared to the proposed method. This is because the overall proportions are higher in Table 4 than Table 1 for a given expected number of positive events (since the sample size is smaller in Table 4). The Poisson intervals of Breeze do not work well with proportions that are not small. For example, we performed a simulation corresponding to the top third of Table 1 except that the overall proportion $p = .5$ with 1600 expected number of positive events. The simulated lower and upper lack-of-coverage percentages for the Breeze bounds were 1.2% and 1.3%, compared to 4.6% and 4.7% for the proposed method. The Breeze intervals were on average 37% wider than the proposed intervals.

The Breeze intervals also do not work well when the number of clusters is very small, since they do not account for degrees of freedom of the variance estimation. For example, we performed a simulation corresponding to the top third of Table 1 except that data from only 8 clusters were simulated (with 100 observations per cluster), and $p_i \equiv .1$ so that the expected number of positive events was 80. The simulated lower and upper lack-of-coverage percentages for the Breeze bounds were 6.1% and 5.4%, compared to 4.7% and 4.0% for the proposed method.

Table 5 presents simulation results for the same setup as Table 1 except the cluster sizes were taken to be 10 or 100 with probability 1/2. The lack-of-coverage probabilities are larger than the nominal 5% in the bottom third of the table for all the methods. The logit intervals also do not behave as well as in Table 1 for the top two-thirds of the table.

An additional set of simulations was done in which two clusters (each of sample size 50) were simulated from each 32 strata. The expected numbers of positive event were taken as in Table 1, the weights were randomly set to 1 or 10, and the probability of a positive event was taken to be different in the different strata to simulate an intracluster correlation. The results (not shown) were very similar to the results given in Table 1.

4. APPLICATIONS

In this section we consider two applications in which the numbers of positive counts are small. In the first application, involving estimating HIV positivity in an unselected population, the numbers of positive counts are small because the rates of HIV infection are small. In the second application, involving estimating whether individuals have ever used cocaine, the rates are not small but the numbers of positive counts are small because we restrict the analyses to relatively small subdomains. For both applications, SUDAAN (Shah, Barnwell and Bieler 1995) was used to calculate the (design-based) standard errors of the proportions, and the function "FINV" in SAS (1990, p. 547) was used to calculate the quantiles of the F distribution in (1.2).

4.1 Seroprevalence of HIV Estimated From the Third National Health and Nutrition Examination Survey (NHANES III)

NHANES III was a survey conducted in 1988-1994 of the civilian noninstitutionalized population ages 2 months or older of the United States (National Center for Health Statistics 1994). An HIV test was performed on participating individuals 18 years of age or older. McQuillan, Khare, Karon, Schable and Vlahov (1997) studied the seroprevalence of HIV for individuals under the age of 60 years and various subgroups, some of which are displayed in Table 6. Of the 11,202 individuals tested, 59 were infected. The estimated prevalence in Table 6, 0.32%, is far from the unweighted proportion, 0.53% = 59/11202, because the estimated prevalence is a weighted proportion utilizing the sample weights. Because the testing for HIV was anonymous, for these analyses the sample weights were derived from the original NHANES III sample weights of all individuals in the same stand (survey location), race/ethnicity group, sex, and age group (18-39 *vs.* 40-59) of the tested individual (M. Khare, personal communication). The pseudo-design for variance estimation was the sampling of 2 pseudo-PSU's from each of 23 strata (M. Khare, personal communication), which is not the pseudo-design typically used for NHANES III variance estimation.

The linear 90% confidence intervals for prevalence for the various groups in Table 6 are shifted to the left and shorter than the other intervals, which are similar to each other. The proposed intervals are very slightly wider than the Breeze or logit intervals. The effective sample sizes calculated in (2.1) are markedly smaller than the sample sizes because of the design effects of the survey; the confidence intervals based on the truncated procedures will therefore be identical to the ones given in Table 6. The differences between n^* and n^*_{df} are relatively minor. For this relatively rare outcome, the simulations given in section 3 suggest that the Breeze and proposed confidence intervals may maintain their nominal 90% coverage probabilities better than the other intervals.

Table 6
Seroprevalence of HIV Among Adults Aged 18-59 Years Based on the Third National Health and Nutrition Examination Survey

	Total	Sex		Race/ethnicity		
		Male	Female	White	Black	Mex. - Amer.
Sample size	11202	5142	6060	4128	3579	3495
Number infected	59	44	15	9	38	12
Prevalence (%) ± SE	0.320 ± 0.076	0.519 ± 0.130	0.127 ± 0.053	0.203 ± 0.071	1.100 ± 0.247	0.368 ± 0.134
Effective sample size						
n^*	5588	3056	4433	3976	1779	2039
n^*_{df}	5148	2816	4084	3664	1640	1880
Linear 90% con. int.	(0.19, 0.45)	(0.30, 0.74)	(0.04, 0.22)	(0.08, 0.33)	(0.68, 1.52)	(0.14, 0.60)
Logit 90% con. int.	(0.21, 0.48)	(0.34, 0.80)	(0.06, 0.26)	(0.11, 0.37)	(0.75, 1.62)	(0.20, 0.69)
Breeze 90% con. int.	(0.21, 0.48)	(0.32, 0.79)	(0.05, 0.26)	(0.10, 0.37)	(0.73, 1.61)	(0.18, 0.68)
Proposed 90% con. int.	(0.20, 0.48)	(0.32, 0.80)	(0.05, 0.26)	(0.10, 0.37)	(0.71, 1.63)	(0.17, 0.69)

Table 7
"Ever Users" of Cocaine Among Adults Ages 12-44 Years Based on Individuals with 16 or More Years of Education Sampled in Hispanic Health and Nutrition Examination Survey

	Total	Sex	
		Male	Female
Sample size	123	69	54
Ever-users	13	10	3
Proportion (%) ± SE	11.6 ± 2.5	14.3 ± 3.4	7.0 ± 4.8
Effective sample size			
n^*	167.1	105.0	28.2
n^*_{df}	132.8	84.4	22.9
Linear 90% confidence int.	(7.0, 16.2)	(8.0, 20.7)	(-1.9[a], 15.9)
Logit 90% confidence int.	(7.8, 17.1)	(9.1, 21.9)	(1.9, 22.8)
Breeze 90% confidence int.	(7.7, 17.0)	(8.3, 23.2)	(0.9, 24.8)
Proposed 90% confidence int.	(7.4, 17.2)	(8.5, 22.1)	(0.9, 22.7)
Truncated Procedures			
Linear 90% confidence int.	(6.3, 17.0)	(6.5, 22.2)	same as above
Logit 90% confidence int.	(7.2, 18.2)	(8.1, 24.1)	"
Breeze 90% confidence int.	(7.1, 18.1)	(7.7, 24.4)	"
Proposed 90% confidence int.	(7.2, 17.5)	(8.0, 23.2)	"

(a) In practice, this interval would be presented as (0, 15.9) since negative proportions are impossible.

4.2 Use of Cocaine Among College-educated Individuals Sampled in the Hispanic Health and Nutrition Examination Survey (HHANES)

HHANES was a survey conducted in 1982-1983 of three Hispanic groups living in the United States (National Center for Health Statistics 1985). We restrict attention here to the Mexican-American sample. Individuals ages 12-44 years were asked "About how old were you the first time you tried cocaine?". The possible answers were the age of the individual (in years) when he first tried cocaine, a "never used" category, and a "don't know" category. We consider estimating the proportion of "ever-users" among individuals who completed 16 or more years of education (for which there were no "don't know" responses).

There were 13 ever-users among 123 sampled individuals, with the sample-weighted proportion being 11.6% (Table 7). The design-based standard error, 2.5%, is estimated with only 8 degrees of freedom since the sampling design of HHANES can be approximated by the sampling of 2 PSU's from each of 8 strata (Kovar and Johnson 1986). The effective sample sizes are $n^* = 167.1$ and $n^*_{df} = 132.8$, which are both greater than the sample size. This is because the estimated design effect is .736, so that $n^* = 123/.736 = 167.1$. (The second factor in (2.2) is 0.794.) Despite the stratification, we think that the true design effect is greater than 1 for this survey because of the clustering and the sample weighting. (The estimated design effect is estimated poorly because of the limited degrees of freedom.) We therefore think that the truncated procedures are reasonable for this application.

Because of the limited degrees of freedom, and because the outcome is not rare, there are more differences between the logit, Breeze and proposed confidence intervals displayed in Table 7. Based on the simulations given in section 3, we recommend the proposed (truncated) confidence intervals.

Our approach may appear slightly inconsistent for this survey in that we accept poorly-estimated effective sample sizes less than the sample size but truncate those greater. We believe that this is a reasonable conservative approach to use when it is thought that the true design effect is probably greater than 1.

5. DISCUSSION

Although the confidence intervals proposed here had adequate coverage probability for almost all the simulations performed, this is not guaranteed for all possible configurations of the population, *e.g.*, see the bottom third of Table 5. An example with a more serious lack of coverage can also easily be constructed: Suppose that the population consists of clusters of size 100, and that 10% of the clusters have all positive units and the remaining 90% have all zero units. If we sample 10 clusters as a simple random sample, and subsample all the units in the sampled clusters, then 35% (= $(1-.1)^{10}$) of the time we will observe no positive units in the sample size of 1000. In this situation, our proposed intervals reduce to the usual binomial ones, so that, *e.g.*, the upper 95% confidence limit for the population proportion is given by .003 (= $1-.05^{1/1000}$). This implies that

the upper 95% confidence interval is less that the true value of .10 at least 35% of the time, a serious undercoverage.

It is possible in simple sampling situations to construct confidence intervals that are guaranteed to have at least their nominal coverage probability by considering all possible configurations of the population, and using a least-favorable configuration for the coverage probability. For the hypothetical single-stage cluster sample mentioned above, for example, an upper 95% confidence limit could be given by the binomial limit based on 0 positive units out of 10, *i.e.*, .26 ($=1-.05^{1/10}$). Such confidence intervals, which can become computationally intensive to calculate, have been studied by Gross and Frankel (1991), who also suggest some less computationally intensive approximations.

The advantages of our proposed intervals over such approaches are (1) they are easy to calculate, (2) they accommodate any complex sampling design, including nonresponse and postsratification adjustments to the sample weights, (3) they will generally maintain their nominal coverage probability, (4) they will be less conservative than intervals that are guaranteed to maintain their nominal coverage probability for all population configurations, and (5) they have better properties than the linear intervals, logit-transform or Breeze intervals. Conclusions (2) and (5) are based on our simulation results, which of course do not cover all possible situations. More research would be useful in this regard.

ACKNOWLEDGEMENTS

The authors like to thank M. Khare for providing the prevalence estimates and their design-based standard errors given in Table 6, and the Associate Editor and referees for their helpful comments.

REFERENCES

AGRESTI, A. (1990). *Categorical Data Anlaysis*. New York: Wiley.

BREEZE, E. (1990). *General Household Survey: Report on Sampling Error*. London: Her Majesty's Stationery Office (Office of Population Censuses and Surveys).

CLOPPER, C.J., and PEARSON, E.S. (1934). The use of confidence or fiducial limits illustrated in the case of the binomial. *Biometrika* 26, 404-413.

COCHRAN, W.G. (1977). *Sampling Techniques*. Third Edition. New York: Wiley.

FRANKEL, M.R. (1971). *Inference from Survey Samples: An Empirical Investigation*. Ann Arbor, MI: Institute for Social Research, The University of Michigan.

GRAUBARD, B.I., and KORN, E.L. (1996). Modelling the sampling design in the analysis of health surveys. *Statistical Methods in Medical Research*, 5, 263-281.

GROSS, S.T., and FRANKEL, M.R. (1991). Confidence limits for small proportions in complex samples. *Communications in Statistics – Theory and Methods*, 20, 951-975.

GUENTHER, W.C. (1983). Hypergeometric distributions. In *Encyclopedia of Statistical Sciences*, Volume 3, (Eds. S. Kotz and N.L. Johnson). New York: Wiley, 707-712.

JENNINGS, D.E. (1987). How do we judge confidence-interval adequacy? *American Statistician*, 41, 335-337.

JOHNSON, N.L., and KOTZ, S. (1970). *Continuous Univariate Distributions – 2*. New York: Wiley.

JOHNSON, N.L., KOTZ, S., and KEMP, A.W. (1993). *Univariate Discrete Distributions*. Second Edition. New York: Wiley.

KORN, E.L., and GRAUBARD, B.I. (1990). Simultaneous testing or regression coefficients with complex survey data: use of Bonferroni *t*-statistics. *American Statistician*, 44, 270-276.

KORN, E.L., and GRAUBARD, B.I. (1998). Variance estimation for superpopulation parameters. *Statistica Sinica*, 8, 1131-1151.

KOVAR, M.G., and JOHNSON, C. (1986). Design effects from the Mexican American portion of the Hispanic Health and Nutrition Examination Survey: a strategy for analysts. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 396-399.

KREWSKI, D., and RAO, J.N.K. (1981). Inference from stratified samples: properties of the linearization, jackknife and balanced repeated replication methods. *Annals of Statistics*, 9, 1010-1019.

McQUILLAN, G.M., KHARE, M., KARON, J.M., SCHABLE, C.A., and VLAHOV, D. (1997). Update on the seroepidemiology of Human Immunodeficiency Virus in the United States household population: NHANES III, 1988-94. *Journal of Acquired Immune Deficiency Syndromes and Human Retrovirology*, 14, 355-360.

NATIONAL CENTER FOR HEALTH STATISTICS (1985). Plan and operation of the Hispanic Health and Nutrition Examination Survey, 1982-84. *Vital and Health Statistics* 1(19). Hyattsville, MD: National Center for Health Statistics.

NATIONAL CENTER FOR HEALTH STATISTICS (1994). Plan and operation of the Third National Health and Nutrition Examination Survey, 1988-94. *Vital and Health Statistics* 1(32). Hyattsville, MD: National Center for Health Statistics.

PFEFFERMANN, D., and LAVANGE, L. (1989). Regression models for stratified multi-stage cluster samples. In *Analysis of Complex Surveys*, (Eds. C.J. Skinner, D. Holt, and T.M.F. Smith). New York: Wiley, 237-260.

SAMHSA (1998). National Household Survey on Drug Abuse: Main Findings 1996. (DHHS Publication No. (SMA) 98-3200). Rockville, MD: SAMHSA.

SAS (1990). *SAS Language: Reference, Version 6*, First Edition. Cary, NC: SAS Institute Inc.

SHAH, B.V., BARNWELL, B.G., and BIELER, G.S. (1995). *SUDAAN User's Manual, Release 6.40*. Research Triangle Park, NC: Research Triangle Institute.

VOLLSET, S.E. (1993). Confidence intervals for a binomial proportion. *Statistics in Medicine*, 12, 809-824.

ACKNOWLEDGEMENTS

Survey Methodology wishes to thank the following persons who have served as referees during 1998. An asterisk indicates that the person served more than once.

B.M. Balk, *Statistics Netherlands*
* D.R. Bellhouse, *University of Western Ontario*
J. Bethel, *Westat, Inc.*
P. Biemer, *Research Triangle Institute*
* D.A. Binder, *Statistics Canada*
J.-R. Boudreau, *Statistics Canada*
K. Brewer, *Australian National University*
J.M. Brick, *Westat, Inc.*
A. Brinkley, *U.S. Bureau of the Census*
T.W. Broene, *Energy Information Administration, U.S.A.*
N. Buck, *University of Essex*
P. Cantwell, *U.S. Bureau of the Census*
* R. Chambers, *University of Southampton*
* M. Cohen, *National Center for Education Statistics*
B. Cox, *Mathematica Policy Research*
* J. Denis, *Statistics Canada*
* J.-C. Deville, *Institut national de la statistique et des études économiques*
* P. Dick, *Statistics Canada*
A. Dorfman, *U.S. Bureau of Labor Statistics*
* J.D. Drew, *Statistics Canada*
J. Dufour, *Statistics Canada*
* J. Eltinge, *Texas A&M University*
R. Evans, *University of Nebraska - Lincoln*
M. Feder, *University of Southampton*
* W.A. Fuller, *Iowa State University*
* J. Gambino, *Statistics Canada*
B. Graubard, *U.S. National Cancer Institute*
* R.M. Groves, *University of Maryland*
S.J. Haslett, *Massey University, New Zealand*
L. Hattersley, *Office for National Statistics, U.K.*
* M.A. Hidiroglou, *Statistics Canada*
* D. Holt, *Office for National Statistics, U.K.*
K. Humphreys, *University of Glasgow*
J.-S. Hwang, *Academia Sinica*
W. Jocelyn, *Statistics Canada*
E. Johnson, *Educational Testing Service, U.S.A.*
D. Judkins. *Westat, Inc.*
* G. Kalton, *Westat, Inc.*
B. King, *Case Western Reserve University*
* P.S. Kott, *National Agricultural Statistics Service*
M. Kovačević, *Statistics Canada*
M. Kramer, *U.S. Bureau of the Census*
A. Krieger, *University of Pennsylvania*
* R. Lachapelle, *Statistics Canada*
* P. Lahiri, *University of Nebraska - Lincoln*
M. Larson, *Harvard University*
* P. Lavallée, *Statistics Canada*
J. Lawless, *University of Waterloo*
G. Lee, *Australian Bureau of Statistics*
* S. Linacre, *Australian Bureau of Statistics*
* R. Little, *University of Michigan*
* S. Lohr, *Arizona State University*
* H. Mantel, *Statistics Canada*
P.L. do Nascimento Silva, *IBGE, Brasil*
* G. Nathan, *Central Bureau of Statistics, Israel*
D. Paton, *Statistics Canada*
* D. Pfeffermann, *Hebrew University*
N.G.N. Prasad, *University of Alberta*
* B. Quenneville, *Statistics Canada*
T.E. Raghunathan, *University of Michigan*
* J.N.K. Rao, *Carleton University*
J.O. Ramsey, *McGill University*
E. Rancourt, *Statistics Canada*
* L.-P. Rivest, *Université Laval*
G. Roberts, *Statistics Canada*
G. Robinson, *Commonwealth Scientific and Industrial Research Organisation*
* I. Sande, *Bell Communications Research, U.S.A.*
C.-E. Särndal, *Université de Montréal*
* N. Schenker, *University of California - Los Angeles*
* F.J. Scheuren, *George Washington University*
A.J. Scott, *University of Auckland*
* J. Sedransk, *Case Western Reserve University*
* M.P. Singh, *Statistics Canada*
* R. Sitter, *Simon Fraser University*
* C.J. Skinner, *University of Southampton*
K.P. Srinath, *ABT Associates, Inc.*
L. Stokes, *University of Texas - Austin*
* D. Stukel, *Statistics Canada*
A. Théberge, *Statistics Canada*
S. Thivierge, *Statistics Canada*
R. Thomas, *Carleton University*
I. Thomsen, *Statistics Norway*
R. Tortora, *Gallup Organization*
C. Tucker, *U.S. Bureau of Labor Statistics*
* R. Valliant, *U.S. Bureau of Labor Statistics*
* V.K. Verma, *University of Essex*
* P.J. Waite, *U.S. Bureau of the Census*
* J. Waksberg, *Westat, Inc.*
* K.M. Wolter, *National Opinion Research Center*
F. Yu, *Australian Bureau of Statistics*
E. Zanutto, *University of Pennsylvania*
* A. Zaslavsky, *Harvard University*

Acknowledgements are also due to those who assisted during the production of the 1998 issues: J. Beauseigle (Dissemination Division) and L. Perreault (Official Languages and Translation Division). Finally we wish to acknowledge D. Blair, C. Larabie, D. Lemire, C. Marleau and G. Ray of Household Survey Methods Division, for their support with coordination, typing and copy editing.

CONTENTS TABLE DES MATIÈRES

Volume 26, No. 3, September/septembre 1998

Volume 26, No. 4, December/decembre 1998

JOURNAL OF OFFICIAL STATISTICS

An International Review Published by Statistics Sweden

JOS is a scholarly quarterly that specializes in statistical methodology and applications. Survey methodology and other issues pertinent to the production of statistics at national offices and other statistical organizations are emphasized. All manuscripts are rigorously reviewed by independent referees and members of the Editorial Board.

Contents
Volume 14, Number 2, 1998

Volume 14, Number 3, 1998

All inquires about submissions and subscriptions should be directed to the Chief Editor:
Lars Lyberg, R&D Department, Statistics Sweden, Box 24 300, S - 104 51 Stockholm, Sweden.

GUIDELINES FOR MANUSCRIPTS

Before having a manuscript typed for submission, please examine a recent issue (Vol. 19, No. 1 and onward) of *Survey Methodology* as a guide and note particularly the following points:

1. Layout

1.1 Manuscripts should be typed on white bond paper of standard size (8½ × 11 inch), one side only, entirely double spaced with margins of at least 1½ inches on all sides.
1.2 The manuscripts should be divided into numbered sections with suitable verbal titles.
1.3 The name and address of each author should be given as a footnote on the first page of the manuscript.
1.4 Acknowledgements should appear at the end of the text.
1.5 Any appendix should be placed after the acknowledgements but before the list of references.

2. Abstract

The manuscript should begin with an abstract consisting of one paragraph followed by three to six key words. Avoid mathematical expressions in the abstract.

3. Style

3.1 Avoid footnotes, abbreviations, and acronyms.
3.2 Mathematical symbols will be italicized unless specified otherwise except for functional symbols such as "exp(·)" and "log(·)", etc.
3.3 Short formulae should be left in the text but everything in the text should fit in single spacing. Long and important equations should be separated from the text and numbered consecutively with arabic numerals on the right if they are to be referred to later.
3.4 Write fractions in the text using a solidus.
3.5 Distinguish between ambiguous characters, (e.g., w, ω; o, O, 0; l, 1).
3.6 Italics are used for emphasis. Indicate italics by underlining on the manuscript.

4. Figures and Tables

4.1 All figures and tables should be numbered consecutively with arabic numerals, with titles which are as nearly self explanatory as possible, at the bottom for figures and at the top for tables.
4.2 They should be put on separate pages with an indication of their appropriate placement in the text. (Normally they should appear near where they are first referred to).

5. References

5.1 References in the text should be cited with authors' names and the date of publication. If part of a reference is cited, indicate after the reference, e.g., Cochran (1977, p. 164).
5.2 The list of references at the end of the manuscript should be arranged alphabetically and for the same author chronologically. Distinguish publications of the same author in the same year by attaching a, b, c to the year of publication. Journal titles should not be abbreviated. Follow the same format used in recent issues.