International Statistical Review (2001), **69**, 3, 461–485, Printed in The Netherlands

Estimation Methods and Related Systems at Statistics Canada

M.P Singh, M.A. Hidiroglou, J.G. Gambino and M.S. Kovačević

Methodology Branch, Statistics Canada, Ottawa, K1A 0T6, Canada

Summary

This paper provides an overview of research in estimation techniques, their application, and the development of generalized estimation systems at Statistics Canada. In Canada, the demand for more detailed and better quality cross-sectional data related to various socio-economic issues has increased significantly in recent years. Also, there has been increasing interest in longitudinal data to better understand and interpret the relationships among variables, necessitating the implementation of a number of large scale panel surveys by Statistics Canada. The paper briefly discusses estimation for longitudinal data and a weighting approach developed for cross-sectional data from these surveys. For cross-sectional household and business surveys, as well as the census of population, appropriate calibration estimators developed for each situation are briefly discussed. In addition, regression composite estimation, a method developed to improve the quality of cross-sectional estimates from rotating panel surveys such as the Canadian Labour Force Survey, is presented. With regard to more detailed cross-sectional estimates at sub-provincial levels, different approaches to small area estimation developed for various programs are also presented. We summarize the various modules developed for the Generalized Estimation System. Important new developments within the system include two-phase estimation as well as the estimation of variance for a number of imputation procedures. We briefly review the status of current estimation research on selected topics as well as the direction of future research.

Key words: Complex surveys; Estimation; Variance estimation; Estimation systems.

1 Introduction

As the country's national statistical agency, Statistics Canada conducts a wide range of surveys to shed light on the critical social and economic issues facing Canada and its provinces. In addition, the Census of Population provides benchmark information on the Canadian population and its demographic, social and economic conditions at small geographic levels and for sub-populations that cannot be generated through surveys. The surveys conducted by Statistics Canada vary greatly in their periodicity, content and complexity. They range from simple ad hoc cross-sectional surveys to a number of complex periodic (monthly to annual) and longitudinal household and business surveys. There are a number of challenging issues related to sample design, data collection, data processing and data analysis in our major cross-sectional and longitudinal surveys. However, in this paper we have restricted our discussion to research, development and application of estimation techniques to major Statistics Canada surveys.

The primary objective of most cross-sectional surveys is to produce unbiased (or nearly unbiased) estimates of levels such as totals, means and ratios and also estimates of change from repeated surveys, with associated measures of precision. In providing estimates from these surveys, the weighting and estimation methods reflect the sample design followed in each case. Further, to

improve the efficiency of these estimates, information on suitable auxiliary variables is incorporated in the estimation process and the original sampling weights are adjusted to obtain calibrated weights whose totals match benchmark constraints. This is usually achieved through raking or regression methods. Numerous estimation procedures for cross-sectional surveys have been consolidated in a generalized estimation system (GES). This development, described in section 2, unifies a wide variety of estimation procedures using auxiliary data under one umbrella, using regression (or generalized regression-GREG). Recent developments in calibration have also been implemented in GES, using algorithms that are based on linear programming techniques. Variance estimation for the GREG estimator uses the Taylor or jackknife procedure.

In section 3, we briefly describe the use of regression estimators in two of our major monthly surveys, namely the Labour Force Survey (LFS) and the Survey of Employment, Payrolls and Hours (SEPH), as well as in the Census of Population. Variations of regression that take account of the design and level at which the estimates are produced are briefly discussed, along with the corresponding variance estimates. Further, a new regression composite estimator which improves the quality of LFS estimates by exploiting the rotating panels of the LFS design has been developed. This variation on traditional composite estimation ensures the internal consistency of estimates while achieving significant efficiency gains for key variables.

Section 4 deals with the challenges presented by longitudinal surveys. While cross-sectional data are suitable for monitoring socio-economic patterns and trends, they do not provide information on social processes per se. Rather, it is longitudinal surveys, where data are collected from the same respondents over a period of time, that provide the opportunity to better understand and interpret the underlying causal relationships among variables, such as whether it is low income that leads to poor health or failing health that leads to a decline in income. The resulting in-depth analyses of various phenomena will provide new information for policy changes that can affect Canadians. To respond to these information needs, Statistics Canada has launched several household panel surveys on topics such as labour, income, health and education. Also a new business panel has been launched recently where data on both employees and employers are collected longitudinally. Important design features of these longitudinal surveys, along with weighting and estimation issues, are given in section 4.

Although the primary objective of these surveys remains the production of longitudinal data series, there is growing demand for deriving cross-sectional estimates from them, enhancing their cost-effectiveness. This has implied that their sample design takes this factor into account, and that estimation procedures satisfying cross-sectional as well as longitudinal requirements be developed. Similarly, the requirement for more detailed cross-sectional estimates at sub-provincial levels has meant that different approaches to dealing with small area estimation had to be developed. These approaches, including design modifications, accumulating data over time, combining data from different sources and using model-dependent estimators are covered in section 5.

Recent work on variance estimation is described in section 6. This includes a brief overview of developments in variance estimation in the presence of imputation. Section 7 discusses developments on several estimation topics. Finally, in section 8, we briefly mention the general direction of future research and development on estimation-related issues. In addition to the references cited in this document, research and development work done at Statistics Canada is documented in a series of internal methodology working papers not cited here.

2 Regression Estimation and Generalized Systems

The need to automate increasingly complex estimation and variance estimation procedures was recognized in the mid-eighties. The rationale for the development of generalized systems for automating estimation is described in Outrata & Chinnappa (1989). Generalized estimation systems have the property that they can be applied to a wide variety of survey designs and estimators. Gen-

eralized estimation software has several advantages over customized software, including (i) reduced maintenance costs and training due to staff rotation; (ii) a unified single systems architecture and methodology; (iii) flexibility for the methodologist to try out different estimation procedures for a given survey; and (iv) the embedding of new systems and methodological advances.

Several estimation packages have been developed elsewhere using different approaches for the methodology framework. These include LINWEIGHT (Bethlehem & Keller, 1987), PC-CARP (Schnell *et al.*, 1988), SUDAAN (Shah *et al.*, 1989), CLAN (Andersson & Nordberg, 1994), WESVAR (from Westat) and others. These packages have several features in common with respect to the sampling designs that they accommodate and the parameters that they estimate. For instance, a common feature is that they all handle stratified clustered probability-proportional-to-size (PPS) sampling designs with and without replacement. Common estimated parameters include population totals, means and ratios. The differences between these packages are with respect to (i) the availability of analytic features such as regression, quantiles, logistic regression, and two-way table analysis and (ii) variance estimation procedures (Taylor, jackknife, or replication). The estimation procedures for a variety of sampling designs used at Statistics Canada are increasingly incorporating auxiliary data. Therefore, the framework adopted for building a generalized estimation system (GES) is based on the use of auxiliary information, and of the generalized regression estimator (GREG).

2.1 Sampling Designs and the GES

Specifications for a general estimation system were initially written in 1990 and 1991. A detailed description of the methodology can be found in Estevao, Hidiroglou & Särndal (1995). GES is built around the following elements: the sampling plan, the population parameters to be estimated, the use of auxiliary information, and domains of interest. The sampling designs include (i) single-stage designs such as stratified simple random sampling with and without replacement (SRSWR and SRSWOR), (ii) stratified cluster sampling and stratified PPS sampling, (iii) stratified multistage designs with the components computed one stage at a time, and (iv) stratified two-phase sampling with the sampling units drawn using SRS within each stratum at each phase. GES computes estimates of totals, means, and ratios with their associated measures of reliability given that auxiliary information has been incorporated in the estimation process. This auxiliary information can cut across design strata, or be included within them. This allows the computation of most of the commonly used estimators in survey sampling, including separate and combined ratio or regression estimators (or intermediate combinations), poststratified estimators (separate, combined, or mixed), and others such as the raking ratio estimator. Estimates and their associated measures of reliability are computed for user-specified domains of interest.

PPS with and without replacement sampling have been incorporated for one-stage stratified sample designs, with the estimated variance being computed only for with-replacement sampling. The reason for this is that the estimated variance for PPS without replacement schemes requires the computation of joint selection probabilities. The computation of such joint probabilities is usually not trivial and differs among the many selection mechanisms that exist for drawing PPS samples (see Brewer & Hanif, 1983). Approximations exist for eliminating the need to compute these joint selection probabilities. These approximations alter only slightly the estimated variance for PPS with replacement schemes by incorporating correction factors for each sampled unit.

2.2 Regression Estimation

The straight expansion estimator of a population total Y is

$$\hat{Y} = \sum_{k \in s} w_k y_k$$

where y_k is the value of the characteristic of interest for the kth unit in the sample s. Here, w_k is the design weight adjusted for unit nonresponse. Let $\mathbf{X}$ denote a vector of p known population totals, sometimes referred to as control totals. For example, in a household survey, these could be the number of people in various age-sex groups. Let $\boldsymbol{x}_k = (x_{1k}, x_{2k}, \ldots, x_{pk})$ be the corresponding set of variables for the kth unit in the sample. In the age-sex example, each element of this vector indicates whether or not the kth individual is in the corresponding age-sex group. Thus $\mathbf{X}$ is the sum of the $\boldsymbol{x}_k$ over the whole population. The generalized regression estimator of Y is

$$\hat{Y}_{\text{GREG}} = \hat{Y} + \left(\mathbf{X} - \hat{\mathbf{X}}\right)' \hat{\boldsymbol{\beta}}$$

where

$$\hat{\boldsymbol{\beta}} = \left(\sum_{k \in s} w_k \boldsymbol{x}_k \boldsymbol{x}_k'\right)^{-1} \sum_{k \in s} w_k \boldsymbol{x}_k y_k.$$

This estimator has the property that $\hat{X}_{\text{GREG}} = X$, i.e., it reproduces all the control totals exactly.

Let

$$\mathbf{X} = [\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n]'$$

be the n by p matrix of auxiliary variables for the sample. It can be shown that if the n-vector of ones is in the column space of $\mathbf{X}$ then the GREG estimator simplifies to

$$\hat{Y}_{\text{GREG}} = \mathbf{X}' \hat{\boldsymbol{\beta}}.$$

In practice, this is often the case since one set of auxiliary variables, such as age-sex groups, are mutually exclusive and exhaustive. As long as each age-sex group contains at least one person in the sample, then the above condition holds. The condition also holds if the estimation area is partitioned into geographical regions which are all used as auxiliary variables: as long as there is some sample in each region, then the unit vector will be in the column space of $\mathbf{X}$.

The GREG estimator can be written as

$$\hat{Y}_{\text{GREG}} = \sum_{k \in s} \tilde{w}_k y_k,$$

where

$$\tilde{w}_k = w_k g_k = w_k \left[1 + \left(\mathbf{X} - \hat{\mathbf{X}}\right)' \left(\sum_{j \in s} w_j \boldsymbol{x}_j \boldsymbol{x}_j'\right)^{-1} \boldsymbol{x}_k\right].$$

Here, we have implicitly defined the g-factor g_k. In the case where the auxiliary variables include a mutually exclusive and exhaustive set of categories, this reduces to

$$\tilde{w}_k = w_k g_k = w_k \left[\mathbf{X}' \left(\sum_{j \in s} w_j \boldsymbol{x}_j \boldsymbol{x}_j'\right)^{-1} \boldsymbol{x}_k\right].$$

The weights $\tilde{w}_k$ are called regression, calibration or final weights.

A more general form of the GREG estimator uses the g-factor

$$g_k = 1 + \left(\mathbf{X} - \hat{\mathbf{X}}\right)' \left(\sum_{s} w_k \frac{\boldsymbol{x}_k \boldsymbol{x}_k'}{\sigma_k^2}\right)^{-1} \frac{\boldsymbol{x}_k}{\sigma_k^2}.$$

Some surveys conducted by Statistics Canada use these more general estimators—see section 2.3.

A disadvantage of the GREG estimator is that the resulting final weights $\tilde{w}_k$ may be negative,

smaller than one or very large. A number of authors including Huang & Fuller (1978), Deville & Särndal (1992), Singh & Mohl (1996), Rao & Singh (1997) and Théberge (2000), have developed procedures that ensure that the calibration weights are bounded.

The variance of $\hat{Y}_{\text{GREG}}$ can be estimated by

$$v_{\text{TAY}}\left(\hat{Y}_{\text{GREG}}\right) = \sum_{(k,\ell)\in s}\sum (1 - f_{k\ell})\, z_k\, z_\ell$$

where the finite population correction factor $f_{kl} = w_{kl}/w_k\, w_l$, $z_k = \widetilde{w}_k\, e_k$, $w_{kl} = 1/\pi_{kl}$, where π_{kl} is the joint probability of including units k and l into the sample, and $e_k = y_k - x_k'\hat{\beta}$. It should be noted that GES does not necessarily compute the variances in the double sum form. For example, in the case of *stratified SRSWOR* the computational form is

$$v_{\text{TAY}}\,(Y_{\text{GREG}}) = \sum_{h=1}^{H} \frac{n_h\,(1 - f_h)}{n_h - 1} \sum_{s_h} (z_k - \bar{z}_h)^2$$

where n_h is the number of sampled units in stratum h, f_h is the associated finite population factor, $\bar{z}_h$ is the mean of the z_k variables in stratum h, and H is the number of strata. For PPS sampling with replacement, the correction factor f_h disappears.

The jackknife can also be used to estimate variances for surveys with multistage designs, where each stratum contains a sample of several first stage units (FSUs). To estimate the jackknife variance of $\hat{Y}_{\text{GREG}}$, we begin by deleting, from stratum h, an FSU j and adjusting the weights of the sample in the remaining FSUs in stratum h to compensate for the deleted sample. This produces an estimate $\hat{Y}_{\text{GREG}(hj)}$ of Y. This is repeated for all FSUs and all strata in a province. Typically, there are several hundred FSUs in a province. The jackknife variance estimate of Y is

$$v_J\left(\hat{Y}_{\text{GREG}}\right) = \sum_{h=1}^{H} \frac{n_h - 1}{n_h} \sum_{j=1}^{n_h} \left(\hat{Y}_{\text{GREG}(hj)} - \hat{Y}_{\text{GREG}}\right)^2$$

where n_h is the number of FSUs in stratum h and H is the total number of strata. For a recent survey of replication-based variance estimators, such as the jackknife, used in complex surveys, see Rust & Rao (1996).

To estimate the variance of a ratio Y/Z of two totals, such as the unemployment rate, we simply replace $\hat{Y}$ and $\hat{Y}_{(hj)}$ in the above variance formula by $\hat{Y}/\hat{Z}$ and $\hat{Y}_{(hj)}/\hat{Z}_{(hj)}$, respectively.

2.3 *Regression Estimation for Two-Phase Sampling*

Two-phase sampling is increasingly being used at Statistics Canada due to the wealth of timely administrative data of reasonably good quality that is becoming available. This is especially the case in business surveys where this procedure has been used for several annual and sub-annual surveys. Examples include the use of two-phase sampling annually (Choudhry *et al.*, 1989; Armstrong & St-Jean, 1994), and sub-annually (Binder *et al.*, 2000; Hidiroglou *et al.*, 1995).

Point estimation and variance estimation procedures for this growing set of varied two-phase designs have been developed for eventual integration in GES. Hidiroglou & Särndal (1998) developed a general framework for estimation in two-phase designs with the use of auxiliary information. A short description of this general setup is as follows. A first-phase probability sample $s_1 (s_1 \subseteq U)$ is drawn from the population U, according to a sampling design with selection probabilities $\pi_{1k} = P(k \in s_1)$. Given s_1, the second-phase sample $s_2 (s_2 \subseteq s_1 \subseteq U)$ is selected from s_1, according to a sampling design with selection probabilities $\pi_{2k} = P(k \in s_2 | s_1)$. We assume that $\pi_{1k} > 0$ for all $k \in U$ and $\pi_{2k} > 0$ for all $k \in s_1$. The first-phase sampling weight of unit k is denoted as $w_{1k} = 1/\pi_{1k}$, and the second-phase sampling weight as $w_{2k} = 1/\pi_{2k}$. The overall sampling weight for a selected unit is

$w_k^* = w_{1k}\, w_{2k}$.

Given that the design weights incorporating the first- and second-phase design weights are $w_k^* = w_{1k}\, w_{2k}$ we seek a set of calibrated weights $\tilde{w}_k^*$ that lie as close to them as possible. These weights are computed through two successive stages of calibration. First-phase calibration factors g_{1k} are computed as stated in section 2.1. Let these first-phase calibrated weights be $\tilde{w}_{1k} = w_{1k}\, g_{1k}$ $(k \in s_1)$. Given positive factors $\{\sigma_{2k}^2 : k \in s_2\}$ the overall calibrated weights $\tilde{w}_k^*$ are obtained by minimizing a distance function subject to the additional constraint that the resulting $g_{2k} = \tilde{w}_k^*/(w_{1k}\, w_{2k})$ factors are bounded above and below. The resulting estimator of total is $\hat{Y}_{\mathrm{CAL}} = \sum_{s_2} \tilde{w}_k^*\, y_k$. The first order and second order inclusion probabilities are used in the variance formula. Two sets of regression residuals, one for each phase, are also required. The estimator of the variance is given by

$$v\left(\hat{Y}_{\mathrm{CAL}}\right) = \sum_{k,} \sum_{\ell \in s_2} w_{2k\ell}\,(1 - f_{1k\ell})\, z_{1k}\, z_{1\ell} + \sum_{k,} \sum_{\ell \in s_2} w_{1k}\, w_{1\ell}\,(1 - f_{2k\ell})\, z_{2k}\, z_{2\ell}$$

where correction factors $f_{1k\ell}$, $f_{2k\ell}$, $w_{1k\ell}$, $w_{2k\ell}$, are defined as before. Also, $z_{1k} = \tilde{w}_{1k}\, e_{1k}$, $z_{2k} = \tilde{w}_{2k}\, e_{2k}$, where $\tilde{w}_{2k} = \tilde{w}_k^*/\tilde{w}_{1k}$, and the residuals are estimated from the implied regression models fit at each phase. Again, the above variance is not necessarily computed using double sums for designs that do not require it. Särndal *et al.* (1992) show how the numerical computations can be simplified for a two-phase design that involves an arbitrary sample design at the first phase, and a second-phase sample from an arbitrary re-stratification of the first-phase sample. More recent developments in this area are described in section 6.

3 Cross-sectional Surveys

3.1 Estimation in Household Surveys

Most household surveys conducted by Statistics Canada are related to the Labour Force Survey (LFS): they are either supplements to the LFS, or use former LFS-sampled households, or use the LFS frame to select non-LFS households. As a result, these surveys tend to use the same or similar estimation methods as the ones used by the LFS. In this section, therefore, we will focus on the LFS.

The Canadian Labour Force Survey is a monthly survey of 53,000 households. The survey has a complex multistage design consisting of six rotation groups. Each month, the households in one rotation group are replaced, and each household stays in the sample for six consecutive months. All members of a selected household are in the sample, but children less than fifteen years of age do not receive a labour force questionnaire. The LFS publishes monthly, annual and three-month average estimates for labour force characteristics by industry, occupation, demographic group and various levels of geography. It also publishes data on wages, union membership and hours of work. For a detailed description of the methodology of the LFS, see Singh *et al.* (1990) and Gambino *et al.* (1998).

The LFS uses a generalized regression estimator to produce estimates. It is based on the estimator described in section 2.2. This approach exploits the availability of demographic estimates for various age-sex groups, subprovincial Economic Regions and Census Metropolitan Areas to improve the quality of the straight expansion estimator. This estimator was adopted by the LFS in 1988 following a comparison between the GREG estimator and the raking ratio estimator used before 1988. This work is described by Lemaître & Dufour (1987), who also showed how the regression approach can deal effectively with a long-standing problem, namely, the desire to have a unique final weight for all members of the same household. Their approach, which was adopted by the LFS, amounts to replacing the indicators x_k for person k in the regression matrix X with the average vector for the household. For example, in a household of five people, if there are two male infants, then everyone in the household has a value of 2/5 for the "indicator" for the male 0–4 age-sex group. See Lemaître & Dufour (1987) for details.

For variance estimation, the Labour Force Survey uses the jackknife, as described in section 2.2.

3.1.1 Composite estimation in the Labour Force Survey

Until January 2000, the Labour Force Survey did not use the fact that five-sixths of the LFS sample is common between consecutive months to improve published estimates. It is well known that in a rotating sample design the common sample can be used to produce a better estimate of change compared to simply taking the difference between the usual estimates for two consecutive months. This improved estimate of change can then be used to improve the estimate of level. For example, the traditional K-composite estimator is a linear combination of the usual estimate of level, say a regression estimator, and another estimate of level obtained by taking last month's estimate of level and updating it using an estimate of change based on the common sample, i.e.,

$$\hat{Y}_t^c = K \times \hat{Y}_t + (1 - K) \times \left[\hat{Y}_{t-1}^c + \text{change}_{\text{common}} \right]$$

where the superscript c denotes a composite estimate, t denotes the current month and $\hat{Y}$ denotes an estimator of the variable of interest. Although traditional composite estimators lead to improved estimates, they suffer from a number of drawbacks such as consistency of estimates (in the sense of parts adding up to totals). Therefore, this kind of composite estimation has not been implemented in the LFS.

We present a brief description of an estimator which we will refer to as the regression composite estimator. This estimator deals simultaneously with all characteristics that are to be "composited" and takes care of the consistency issue. The method has the operational advantage that it fits well into the estimation framework used by the LFS—the characteristics of interest enter into the estimation procedure as control totals. It also has two essential properties: each sampled household will have a single weight (i.e., the weight does not depend on the characteristic of interest) and parts will add up to the corresponding total (e.g., the sum of *employed* and *unemployed* will still equal the size of the labour force, which is not the case in the traditional approach where each variable is treated separately).

The regression composite estimator implemented in the LFS extends the regression estimation method used by the survey by adding several labour force characteristics, based on data from the previous month, to the set of demographic characteristics used as auxiliary variables in the past. Thus, to the demographic controls for the current month mentioned in section 3.1, we add controls for the previous month such as *employed, unemployed* and *not in labour force* at the provincial level and for broad age-sex groups, and employment in several industries such as agriculture and construction.

Let y denote one of the above labour force variables and let Y denote its population total. The new estimator uses the estimate of Y from the previous month as an auxiliary variable. This is achieved by first modifying last month's individual weights to reflect the current month's population, resulting in an adjusted estimate $\hat{Y}_{t-1}^*$ for last month's total. Then the weights for the current month are adjusted so that $\alpha \left[\hat{Y}_t - (\text{estimate of change based on the } \textit{common} \text{ sample}) \right] + (1 - \alpha)$ $\left[\text{estimate of last month's total based on the common sample} \right]$ equals $\hat{Y}_{t-1}^*$.

Singh *et al.* (1997) treated the two terms in the square brackets separately (i.e., as separate regressors). The use of a linear combination of the two terms was suggested by W. Fuller (1998). The choice $\alpha = 1$ results in an estimator that performs well for change, and $\alpha = 0$ for level. Thus the choice of α depends on the relative importance one gives to estimates of change versus estimates of level: α is chosen close to one if change is much more important than level. The LFS uses the value $\alpha = 2/3$.

Unlike the demographic totals, the new control totals are random variables, and this must be taken into account when estimating variances. This is accomplished by jackknifing the new control

totals as well. Note that, like the demographic controls, the new controls are incorporated in the estimator *simultaneously*. This differs from the traditional K-composite estimator which treats each characteristic y separately.

For the employment characteristics that are controlled in the regression process, there can be substantial improvement in efficiency as measured by their variance. For example, our studies show that for employment estimates in certain industries whose regression estimates are volatile, the gain in efficiency can exceed 40 percent. For province-level employment and unemployment estimates, the efficiency gains are more modest, typically in the five to ten percent range. For estimates of month-to-month change, the gains can be much more pronounced, especially for α close to one. For example, when $\alpha = 1$, the variance of the estimate of month-to-month change in employment in Ontario is cut in half. For change in employment in some industries, the variance is reduced even more. One important consequence of the latter result is that certain time series which could not be seasonally adjusted effectively in the past are adjustable when regression composite estimation is used, i.e., it increases the signal-to-noise ratio sufficiently to allow the seasonal adjustment procedure to detect the seasonal pattern. Based on these encouraging results, the LFS implemented regression composite estimation in January 2000.

3.2 Estimation in the Canadian Census

The Canadian Census of Population is conducted every five years. Out of every five households in the population, four get a short questionnaire containing basic demographic questions. The remaining households get a long questionnaire which is used to compile detailed information on the Canadian population. Various procedures to weight the detailed information are possible. Since the households that get the long questionnaire are selected systematically, one approach is simply to apply a weight of five to each household. This simple approach, however, can lead to substantial discrepancies between estimates based on the long questionnaire and counts based on the whole population for demographic characteristics, particularly for very small areas. In this section we describe the weighting procedure adopted for the 1991 and 1996 censuses, which improves consistency between weighted sample counts and population counts.

For each census, Canada is divided into Enumeration Areas (EAs) containing approximately 250 households each. For estimation purposes, the EAs are combined into weighting areas (WAs). On average, each WA consists of 7 EAs. For the 1986 census, raking ratio estimation (Brackstone & Rao, 1976, 1979) was used to make estimates of key WA-level population counts, such as the number of males in the WA and the number of people in various age groups in the WA, agree with the full population in each WA. However, as described by Bankier *et al.* (1992, 1997a), there were often substantial discrepancies between estimates and population values at the EA level. To deal with this problem, a two-step GREG procedure was developed for the 1991 Census. The two-step procedure uses GREG at both the EA level and the WA level, using the population totals at the EA and WA levels as control totals, to produce a final weight for each person and household in the 1-in-5 sample. The initial sampling weight is multiplied by an EA-level based g-factor and then by a WA-level g-factor, with each g-factor coming from an application of GREG at the corresponding level. The GREG procedure can be described using the notation in section 2.2.

Since EAs can have a small sample, the final weights obtained by blindly applying GREG can be extreme, i.e., either very large or negative. To deal with this problem, Bankier *et al.* (1992) developed a procedure to reduce the number of controls to ensure that the final weights fall in the interval [1,25]. Briefly, they

- drop all constraints involving fewer than 60 households
- look for constraints that are exactly linearly dependent and drop the one that applies to the fewest households

- look for constraints that are nearly linearly dependent and remove constraints to eliminate the dependence
- if some final weights are still outside the interval [1,25], identify the constraints whose removal would eliminate the problem.

Bankier *et al.* (1992) give the details of how each of these steps is performed. They also present a detailed comparison of this method with the raking ratio method used in the 1986 census and a one-step GREG method. The new method showed better performance, especially at the EA level. It was applied successfully, in a completely automated fashion, to all 5730 WAs in the 1991 census.

3.3 Estimation in Business Surveys

Business surveys vary in their periodicity (annual, subannual), the frame that they use for sampling, the sampling unit they use, the target population, and in their sampling and estimation procedures. We will discuss the methodology for the Survey of Employment, Payrolls, and Hours (SEPH) as it represents the newer methodology to be used in our sub-annual business surveys. SEPH is a monthly survey that collects data on employment, payrolls, working hours, overtime pay and hours, summarized earnings and categories of employment. The primary objectives of the survey are to provide monthly estimates of the total number of paid employees, payrolls, average weekly earnings, average weekly hours and other related variables at the three digit Standard Industrial Classification (SIC) level for Canada and the provinces. The survey covers all industries except agriculture, fishing and trapping, private household services, religious organizations and military services.

The current design for SEPH consists of two independent samples drawn monthly from two separate frames representing the same population: an administrative sample (or payrolls sample) and an establishment sample. There is not a one-to-one correspondence between payroll accounts and establishment, but the set of all payroll accounts and the set of all establishments cover the same population. The payroll sample consists of some 200,000 payroll deduction (PD) accounts sampled systematically (using the last digit as control). Payrolls and number of employees are provided by this sample. The establishment sample, consisting of some 10,000 establishments, is drawn from the Business Register, and the full range of SEPH variables is collected. The sample is stratified by major industry (such as retail trade, wholesale trade, etc.), by geography (groups of provinces) and by size (take-all, take-some). Rotation of the take-some establishments (the smaller establishments) occurs monthly, with selected establishments staying in sample for at most 12 months, and rotated out establishments staying out of the sample for at least 12 months.

The estimation process, described in mathematical notation below, uses auxiliary data from the administrative sample, where known counts are used to adjust the weights. This results in poststratified estimation, allowing the usual stratified SRS variance to be used. The establishment sample is used to obtain a regression model for the estimation of totals for variables not collected in the PD sample. Using total hours as an example, a linear regression is estimated across groups of strata (model groups) using the payroll sample data. Total employees and total payrolls for the month are the independent variables, while total hours and summarized earnings are the dependent variables. Using total employees and total payrolls reported on the administrative sample as the auxiliary variables, parameter values from the regression are used to predict total hours for each unit in the model group. Ratios of total hours by category of employee to total hours are also estimated. Finally, the ratios are used to prorate the total into categories of employee (part time/full time).

For a given domain U_d, the PD administrative totals (employment and payrolls) are estimated from the administrative sample s_1 as $\hat{\mathbf{X}}_1(d) = \sum_{s_1} \tilde{w}_{1k}\, x_{1,k}(d)$, where $\tilde{w}_{1k} = \left\lfloor N_p/\hat{N}_p \right\rfloor w_{1k}$, N_p, is the number of PD accounts for a given partition of the administrative universe, $\hat{N}_p$ is estimated from the administrative sample s_1 and w_{1k} is the original sampling weight for unit $k \in s_1$. The estimated

variance for $\hat{\mathbf{X}}_1$ is the standard post-stratified variance.

Regression coefficients are obtained separately in each of Q model groups. A model group is a group of strata in which there is a strong relationship between the auxiliary variables and the dependent variables. Obtaining a different regression for each model group produces better results than using one global regression since the relationship between variables varies widely from group to group. The model groups are determined ahead of time. Thus we obtain Q regression coefficients b_q, one per model group. The vector b_q corresponds to $\hat{\beta}$ from section two. The predicted variables are produced by multiplying x'_{1k} by b_q. This yields the following estimator for the predicted variables (say y) for domain U_q.

$$\hat{Y}(d) = \sum_{q=1}^{Q} \sum \tilde{w}_{1k}\, x'_{1k}(d) b_q$$

where b_q is obtained by regressing y_k on the x variables (x_{2k}, say) available from the establishment sample. That is, $b_q = \left(\sum_{s_{2q}} w_{2k}\, x_{2,k}\, x'_{2,k}/\hat{\sigma}^2_{2k}\right)^{-1} \sum_{s_{2q}} w_{2k}\, x_{2k}\, y_k/\hat{\sigma}^2_{2k}$ where s_{2q} is the subset of the establishment sample s_2 where the fit between y_k and x_{2k} was obtained, x_{2k} are the data that correspond to the x_{1k} data from the administrative source, and $\hat{\sigma}^2_{2k}$ is a variance factor that results in homogeneous residuals. The estimated variance is obtained by recognizing that $\hat{Y}(d)$ is a product estimator. The estimated variance is made up of two components, one due to the post-stratified estimator for the administrative sample, and the other due to the predicted regression fit from the establishment sample. More details of the methodology are available in Hidiroglou (1995) and Rancourt & Hidiroglou (1998).

4 Longitudinal Surveys

Recognizing a growing need to understand the determinants of changes in the Canadian population and the necessity to use this knowledge in policy development, Statistics Canada has launched several major panel surveys in recent years. Many dynamic aspects of the Canadian population are covered, such as labour, income, health and education. Also, to respond to an increasing demand for longitudinal information about businesses, a business panel survey has recently been introduced. A comprehensive account of the methodological issues in longitudinal surveys is given in Kasprzyk *et al.* (1989) and Binder (1998). The journal *Survey Methodology* (1998) contains a special section on selected papers presented at the international symposium sponsored in 1997 by the International Association of Survey Statisticians. In this section, we give an overview of the design, weighting and estimation issues specific to longitudinal surveys and present a summary of current research in the modelling of longitudinal survey data.

4.1 Some Design Features of Canadian Longitudinal Surveys

The basic design characteristics of ongoing longitudinal surveys are summarized in Table 1 in a comparative manner using a list of "design decisions for a panel survey" by Kalton & Citro (1993). A brief review of the objectives of these surveys is given below.

The goal of the Survey of Labour and Income Dynamics (SLID) is to collect data from Canadian families and individuals to support studies of employment-unemployment dynamics, life-cycle labour market transitions, job quality, quality of working life, family income mobility, dynamics of low income and change in family circumstances.

The main objective of the National Population Health Survey (NPHS) is to provide comprehensive information on the health status of Canadians over time, and to measure the effects of socio-economic and environmental factors, and the relationship between utilization of the health care system and

Table 1

Some Design Characteristics of Canadian Longitudinal Surveys

	SLID	NPHS	NLSCY	WES
Longitudinal Reference Population	1993, 1996, 1999, ... Non-institutionalized residents of 10 provinces of age 16+	1993 All residents of age 12+	1994–95 (First panel) Children of age 0–11	Canadian firms and employees as of 1998
Longitudinal Respondents	All individuals in originally selected households	A longitudinal individual selected from the originally selected household	Up to two children in originally selected households	A representative for the employer, and a selected employee
Additional Cross-Sectional Respondents	Cohabitants of the longitudinal individuals	Cohabitants of the longitudinal individuals	The most knowledgeable person about the longitudinal individual (usually a parent), teachers and the school principal[5]	NA
Length of the panel	6 years	Up to 20 years	At most 25 years	At least 4 years for workplaces and 2 years for employees
Length of the reference period	Varies (1 month to 1 year)	Varies (1 month to 2 years)	Varies (from a current observation to up to 2 years)	1 year
Number of waves	6	Up to 10	At most 12	At least 4 for the workplace, and 2 for employees
Overlapping of panels	Yes, two panels	No	Yes	No
Panel sample size	15,000 households · 31,000 individuals	20,000 individuals	18,000 children (first panel)	7500 establishments and about 40,000 employees
Sample design	LFS[2] (2 rotations)	LFS and ESS[3] (in Quebec)	LFS (9 rotations)	1-stage stratified for employers and 2-stage stratified for employees
Tracking and tracing	All longitudinal individuals while stay in Canada and USA	All longitudinal individuals	Longitudinal individual	Longitudinal workplace and employee
Data collection method	CAPI, Tax Admin Data	CAPI, CATI	CAPI, mail-mail (self administered), face-to-face	CAPI and CATI
Data collection period	January (Labour) and May (Income)	June, August, November or March	November, February, or May	April, May, June, July or August
Special Additional Samples for Cross-Sectional Purposes	Top-up samples from the relevant cross-sectional population, (about 7500 households).	Top-up sample from the relevant cross-sectional population. Extra "Buy-in" samples samples for provinces	Extra "Buy-in" samples for provinces. An additional sample of births (0–23 months old) starting with the 2 wave[4]	A sample of births before each odd wave[4]

[1] Residents in institutions are covered by a separate institutional component
[2] LFS - Canadian Labour Force Survey
[3] ESS - Enquête Sociale et de Santé in Quebec
[4] This addition is actually a new panel that will be followed longitudinally
[5] The role of these essentially cross-sectional respondents is to enrich information about the longitudinal individuals

individual well-being.

The purpose of the National Longitudinal Survey of Children and Youth (NLSCY) is to monitor the development and well-being of Canada's children as they grow from infancy to adulthood, that is, to measure various biological, social and economic characteristics and risk factors among children and youth, and to aid in developing effective policies and strategies to support young people.

The objective of the Workplace and Employee Survey (WES) is to follow an integrated sample of employers and employees in order to determine and study the relationships among different strategies and approaches to management and human resource practices on the employer side, and the resulting outcomes concerning job stability, use of technology, training and earnings on the employee side.

4.2 *Longitudinal Weighting and Estimation*

All three longitudinal social surveys use the frame and design features of the Canadian Labour Force Survey, whose design was described briefly in section 3.1. The exception is the Quebec portion of the NPHS which uses a sample of dwellings selected for the *Enquête sociale et de santé*, a survey managed by Santé Québec.

Data collected in longitudinal surveys present special analytical problems due to the dynamic nature of the units of interest over time. The surveys start with samples of households, and then either all the members with certain characteristics (e.g., SLID and NLSCY) or a sample of members (e.g., NPHS) of those households are followed for the life of the panel. Data are collected not only for the original sample of individuals but for all the persons who are living with the sample members at the time. These persons are usually called cohabitants. While the concept of a longitudinal person is easy to define, a longitudinal household presents a more difficult notion, so usually the household characteristics are considered as characteristics of the persons.

Longitudinal surveys are sensitive to the adverse effects of sample attrition. Unit losses over time occur because of death, migration, inability to trace, and refusal. Nonresponse for these reasons accumulates with each wave, weakens the precision of estimates and leads to bias due to sample unrepresentativeness.

Complex procedures have been developed to determine the response homogeneity (weighting) groups in the second wave (and subsequently) based on detailed information available for respondents from the first wave. Two model-based approaches for creating response (weighting) groups have been used: a segmentation algorithm like CHAID has been used in NPHS (Tambay *et al.*, 1997), and logistic regression modelling has been applied in SLID (Michaud & Hunter, 1992). In the first case, the algorithm creates clusters that are maximally different in propensity to respond provided that the members of the same cluster have a very similar propensity to respond. In the case of the logistic regression approach, the probability to respond is modelled using a set of available covariates, and then the response (weighting) groups are formed using significant covariates in such a way that the probability of response for all individuals within a group is the same. The response groups are further used in a reweighting procedure (Stukel *et al.*, 1997), or for imputation in the case of partial nonresponse.

The last essential step in obtaining the final longitudinal weights is post-stratification, or more generally, calibration where the weights are adjusted so that the weighted counts (totals) for selected domains (post-strata) are equal to known population counts (totals) for these domains pertaining to the panel's year of selection (Latouche & Michaud, 1997).

Longitudinal weighting is a multi-step process carried out independently for each panel, and is essentially done for each wave. For example, the first panel of SLID results in six longitudinal files with six sets of longitudinal weights. A set of criteria determines whether a person is eligible for longitudinal weighting. A person may be eligible for longitudinal weighting but the survey data are not collected; examples include movers into institutions.

Longitudinal estimates are calculated using the final longitudinal weights assigned to the persons in the longitudinal sample. New entrants (e.g., 0–23 month old children in the NLSCY) may be included in the analysis for periods beginning after the start of the panel.

The complexity in calculating the longitudinal weights has inspired research on alternative weighting of longitudinal data (Dufour *et al.*, 1998). A method was developed for the decomposition of the difference between the initial and the final weights according to the major stages in a weighting procedure: the initial (probability) weighting, the non-response adjustment, and the post-stratification. The method allows comparison of weighting procedures through the analysis of the impact of different stages on the final weights. A study conducted to compare the two methods for determination of the response (weighting) groups, i.e., the logistic regression and the segmentation method, showed that the segmentation method is better in creating the more efficient response homogeneity groups. This study also showed that some steps in the very complex weighting scheme can be simplified with minimal loss in efficiency. This is especially important for variance estimation via resampling methods which requires the repetition of the complete weighting process for each replicate.

Much of estimation for longitudinal surveys is associated with measuring change. Typical longitudinal quantities that are estimated from the longitudinal samples are gross changes and transition rates from one state to another.

Variance estimation presents a special challenge. Currently, the jackknifing methodology has been adapted and used for variance estimation of both longitudinal and cross-sectional estimators. For the public use microdata files derived from the NPHS, the bootstrap method was suggested as the most suitable, under confidentiality constraints and considering disclosure problems (Mayda *et al.*, 1996).

4.3 Cross-Sectional Weighting and Estimation

Panel surveys are designed primarily for longitudinal purposes, although very often they are expected to produce cross-sectional estimates as well. In addition to the originally selected individuals (longitudinal individuals), both new entrants to the population and cohabitants originally present in the reference population have to be considered to maintain the cross-sectional representativeness of the sample (see Lavallée, 1995). The main difficulties in cross-sectional estimation arise from the dynamic aspects of a panel, such as attrition, movers, cohabitants and new entrants to the population. There is a danger of a decline in the representativeness of a panel because of attrition; also the representativeness of the cohabitants who joined the originally selected longitudinal persons is always questionable. Weighting begins with computation of the basic weights for the three groups of individuals: the longitudinal ones, the cohabitants originally present and the new entrants. For the longitudinal part of the sample the basic cross-sectional weights are determined after the adjustment for nonresponse. The different procedures for determination of the basic weights are developed for different surveys and thoroughly described in the literature: for SLID —Lavallée & Hunter (1992), Lavallée (1995); for NPHS—Tambay & Catlin (1995); for NLSCY—Statistics Canada (1997); for WES—Patak *et al.* (1998). Additional complexities arise from subsampling of individuals from a selected household (e.g., in NPHS).

Combining overlapping panels of a household panel survey (such as SLID) for cross-sectional purposes presents special estimation problems. These problems have been addressed in Merkouris (1997). An approach outlined there involves the initial construction of a combined cross-sectional sample by weight adjustment of units from domains of the different panels that represent common domains of the cross-sectional population. This is followed by the application of a weight share procedure to the combined sample (in order to deal with the dynamic aspects of the panel). In the final step of weight adjustment, the weights of the combined sample are calibrated to totals of the cross-sectional population.

The problem of incomplete cross-sectional coverage remains even after the combining of overlap-

ping panels. To rectify this problem one may select a special sample from the non-covered population (new entrants), if available, or simply select a new sample from the cross-sectional population at a given time. This one-time sample (also called a top-up sample), in combination with a panel (or with overlapping panels) provides complete coverage of the cross-sectional population. The cross-sectional estimates based on these, say, three sources (two panels of different age and a top-up sample) can be produced from either the combined sample (where samples are combined before post-stratification), or by the use of a combined cross-sectional estimator where the coefficients in the linear combination of corresponding totals are chosen to minimize the variance of the combination (Merkouris, 1997).

4.4 Analysis of Longitudinal Survey Data

Considerable effort has been directed toward research on methods for the analysis of longitudinal survey data. An objective of the research in this area has been to adapt existing inferential methods and develop new ones so that the survey design is accounted for.

Low income issues are one of the government's priorities and a number of research projects have been developed around them. The low income bound is usually defined as one half of the median income for that size (or type) of household, or it is defined using a consumption principle as the minimal necessary income to cover the basic needs of a household. The impact of estimating low income bounds by cross-sectional estimates while calculating transition rates using longitudinal weights was assessed in comparison to longitudinal estimators that use longitudinal weights for estimation of both the low income bound and the change. Simulations were carried out under two attrition scenarios: missing at random and attrition concentrated in low income groups. It was found that the longitudinal estimator is less sensitive to misspecification of the nonresponse adjustment model (Bleuer & Kovačević, 1999a).

Estimation of gross flows and transition rates and their variances (Bleuer & Kovačević, 1999b) is only a starting point on which the next stage, the modelling of change, relies. Use of weight calibration in dealing with nonresponse for estimation of gross flows is discussed by Singh *et al.* (1995). In Kovačević (1999), standard log-linear models of symmetry, marginal homogeneity and independence are modified to account for change and applied to data with temporal dependence. Also, some implications of the survey design on parameter estimates and their estimated variances for semi-parametric models were studied using SLID data on work absence (Hapuarachchi, 1997).

5 Small Area Estimation

With the increase in the planning, administration, and monitoring of various social and fiscal programs at local area levels, there have been increasing demands for more and better quality data at these levels. In Canada the quinquennial census of population provides a benchmark and serves as the richest source of data for small areas, various characteristic domains and target groups of policy interest. Administrative records are another increasingly important source of statistical data (Brackstone, 1987a).

As these demands for small area/domain data relate to important social and economic issues, it is normally the situation that some large scale surveys are already in place. In many cases, the existence of information from such surveys, which were designed to provide reliable estimates at higher levels, itself generates demands for lower level data. Thus survey-based small area estimates, due to their timeliness and policy relevance, are in great demand.

In Canada, we made an early start by organizing an International Symposium on Small Area Statistics (see the volume edited by Platek *et al.*, 1987). There are numerous policy and technical issues that need to be addressed in the provision of small area data. Brackstone (1987b) addresses

these issues in the context of Statistics Canada. For more recent reviews reference should be made to Schaible (1996) for programs in the United States, Ghosh & Rao (1994) for an excellent review of estimation methods and M.P. Singh *et al.* (1992, 1994) for an overall strategy that includes the planning, designing and estimation stages of the survey process. For an overview of small area methods used in practice at Statistics Canada, see Gambino & Dick (2000).

For discussion in this section, we classify the estimators in two broad categories, namely, design-based and model-based. Design-based estimators include the direct estimator (Schaible, 1992) which uses information on study variables only from the domain and for the time period of interest and also the modified direct estimators (see Singh, Gambino & Mantel, 1994) which may use information on the study variable from other domains and time periods. Both these estimators may use auxiliary variables from other domains.

For the majority of large scale surveys, we exploit the opportunities at the design stage to obtain significantly more efficient design-based regional estimates at the expense of small increases in the coefficient of variation at the national and provincial levels. Survey design techniques are usually not sufficient to achieve an adequate degree of precision at local area or rare characteristic domain levels, through the use of design-based estimators. A combination of design-based and model-based estimators is used in such cases. Below we present briefly the techniques used at Statistics Canada.

Combining small areas: What constitutes a small area generally differs from survey to survey and even within the same survey from client to client. For example, for health issues, data may be needed for health regions, for education issues it may be school boards or education planning regions and for labour force surveys, data may be required for geo-political regions such as federal electoral districts, employment centres, census divisions and so on.

One common feature of such areas is that they usually vary greatly in size. There are census divisions consisting of only a couple of census enumeration areas (EAs) and also those with several thousand EAs. One of the challenges then is to determine the actual purpose for which the data are to be used, and in consultation with clients, to arrive at a suitable grouping of the small areas for which reliable design-based estimates can be produced. This usually is achievable.

Pooling data over time: For periodic surveys pooling data over successive occasions to increase the reliability of estimates is a common practice. Depending on the rotation pattern used for such surveys, significant gains in reliability can be achieved. There may, however, be conceptual issues to be sorted out for pooled estimates since such estimates refer to an average of the parameter of interest (e.g., unemployment) over a period of time. Together the grouping of smaller areas and pooling over time provide reasonable estimates from periodic surveys in many situations.

It should be noted that in both these cases, as in the case of indirect estimation, the principle of "borrowing strength" is being used, without, however, having an explicit model.

Combined estimators: Following the work on synthetic estimation at the National Center for Health Statistics (1968) and by Gonzalez (1973), the initial studies carried out at Statistics Canada are reported in papers by Ghangurde & Singh (1977, 1978) and Singh & Tessier (1976). These and several other studies reported in the literature clearly indicated that, in most practical situations, the design-based estimators, though unbiased, can be highly unstable, whereas model-based (indirect) synthetic estimators, though highly efficient, can be heavily biased. A natural approach then was to consider a weighted average of the estimators in order to balance the instability of the design-based estimator and the potential bias of the synthetic estimator.

Such a (design-model) combined estimator of a total for domain d may be written as

$$\hat{Y}_d = \lambda_d \hat{Y}_{1d} + (1 - \lambda_d) \hat{Y}_{2d}$$

where $\hat{Y}_{1d}$ and $\hat{Y}_{2d}$ respectively denote the design-based and model-based estimators and λ is a suitably chosen weight ($0 \leq \lambda \leq 1$). For example, $\hat{Y}_1$ can be the simple expansion, ratio or regression estimator and $\hat{Y}_2$ can be similarly defined as a ratio, regression or other type of synthetic (indirect) estimator. Many of the small area estimators proposed in the literature, both design-based and model-based have the above form. The crucial question is the determination of the weights λ_d and the mean square error (MSE) of this combined estimator. In the following, we present two design-based approaches developed at Statistics Canada for determining the weights. Drew *et al.* (1982) proposed a sample size dependent estimator which uses the weight

$$\lambda_d = \begin{cases} 1 & \text{if } \hat{N}_d \geq \delta N_d \\ \hat{N}_d / \delta N_d & \text{otherwise} \end{cases}$$

where $\hat{N}_d$ is an unbiased estimate of the known domain size N_d and δ is subjectively chosen to control the contribution of the synthetic estimator (or the magnitude of the bias due to the use of the synthetic component). This estimator with $\delta = 2/3$, a generalized regression synthetic estimator as $\hat{Y}_{2d}$ and a generalized regression estimator as $\hat{Y}_{1d}$ is currently used in the Canadian Labour Force Survey to produce domain estimates. Särndal & Hidiroglou (1989) proposed an alternative estimator where weights are defined as

$$\lambda'_d = \begin{cases} 1 & \text{if } \hat{N}_d \geq N_d \\ \left(\hat{N}_d / N_d\right)^{h-1} & \text{otherwise,} \end{cases}$$

where h is subjectively chosen. They suggest $h = 2$ as a general purpose value. Note that λ_d and λ'_d are identical if one chooses $\delta = 1$ and $h = 2$.

It may be noted that Schaible (1979) proposed an averaging of weights based on optimization of the MSE of the combined estimator for several variables, whereas Purcell & Kish (1979) use a common weight and then minimize the average MSE to get optimum weights leading to James–Stein type weights as a special case. Singh & Mian (1995) present a generalization of the sample size dependent estimator. For an appraisal of the developments in design and model based estimation, reference is made to Ghosh & Rao (1994). Developments at Statistics Canada using model-based approaches are reported in papers by Choudhry & Rao (1989), Singh & Mantel (1991), A. Singh *et al.* (1994), Royce (1992), Dick (1995), Pfeffermann & Bleuer (1993), Singh & Wu (1998), and Singh *et al.* (1998).

6 Variance Estimation

Variance estimation techniques have been developed, modified and applied in a number of studies conducted at Statistics Canada. Generally, development has been concentrated in the following areas: linearization methods and their application in variance estimation for two-phase sampling designs, for generalized regression estimators, and for non-linear non-smooth statistics; linearization of the jackknife variance estimator; variance estimation in longitudinal studies; empirical comparison of different resampling methods; and variance estimation under confidentiality constraints.

Binder & Patak (1994) provided the estimating equations (EE) approach for derivation of Taylor series approximations to the variances of a wide class of estimators from complex surveys. Next, Binder (1996) presented several examples of application of the EE approach: the variance of the usual GREG estimator and of the Wilcoxon rank sum test statistic. This approach was also extended to the case of two-phase samples.

Binder & Kovačević (1995) and Kovačević & Binder (1997) derived the Taylor linearized variance estimators, using the EE approach, for a number of complex income inequality and polarization mea-

sures including the Gini coefficient, the Lorenz curve ordinate, the quantile share, the low income proportion, the exponential measure, the polarization index, and the polarization curve ordinate.

In Kovačević & Yung (1997) different variance estimation methods for complex income distribution statistics were compared empirically. Variance estimation methods included in the study are the jackknife, the bootstrap, the grouped balanced half-sample method, the repeatedly grouped balanced half-sample method, and the Taylor linearization method based on estimating equations. Based on the comparison of relative bias, relative stability, and coverage properties of the resulting confidence intervals, it was shown that the Taylor linearized variance estimators perform the best, with the bootstrap method coming second.

Variance estimation for the generalized regression estimator in a two-phase context was derived by Särndal *et al.* (1992). Hidiroglou & Särndal (1998) extended this theory by providing a unified and systematized theory for two-phase sampling with auxiliary information. However, these papers did not provide simple computational expressions that are possible when stratified sampling is implicated at both phases. In Binder *et al.* (2000) such an expression is provided for the case when the first phase sample has been restratified using information gathered from the first phase sample. Simple expressions for variance estimation are provided in this paper for the double expansion estimator and for the reweighted expansion estimator suggested by Kott (1995). The computational simplifications have been incorporated into Statistics Canada's GES. Variance estimation for the same two estimators within the jackknifing methodology was investigated by Kott & Stukel (1997). Under a common two-phase design they found that there was a remarkable difference in performance of the jackknife estimator for these two estimators in favor of the latter one. The linearized jackknife variance estimator, proposed in Yung & Rao (1996), has the good statistical properties of the usual jackknife but is less computationally intensive. The specific form of it is developed for the GREG estimator of a total and for the ratio of two GREG estimators. It is shown empirically that for these estimators the usual jackknife, the linearized jackknife, and the Taylor linearized variance estimator perform similarly.

In an empirical study, Stukel *et al.* (1996) investigated a number of calibration estimators with an emphasis on the properties of their variance estimators: the jackknife and the Taylor. The conclusion of the study was that the jackknife variance estimator had consistently smaller bias than the Taylor estimator, although both variance estimators showed very small bias even under severe restriction of the final weights.

Many Statistics Canada surveys release Public Use Microdata Files (PUMF) that enable analysts to perform their own analyses, but due to confidentiality constraints design information must be suppressed. This lack of design information prevents users of the PUMFs to calculate proper design-based variance estimates. The purpose of the PUMF variance estimation project is to find ways to allow the users of the PUMFs to obtain better estimates of CVs than the existing method while still ensuring the confidentiality of the survey respondents. Two methods are currently being investigated (Yung, 1997): the use of Generalized Variance Functions (GVF) to 'predict' the CV of a point estimate based on the point estimate itself, and a method to include a set of replicate weights representing bootstrap samples which would allow the users to calculate bootstrap variance estimators (Mayda *et al.*, 1996).

Current variance estimation procedures usually do not take imputation into account, i.e., they treat imputed data the same as actual responses. This results in an underestimation of the variance, which may be more or less important depending on the imputation procedure (mean, ratio, hot deck, nearest neighbour, regression), the response rate, and the domains affected. Several papers have addressed this problem over the years at Statistics Canada: Särndal *et al.* (1992), Rao & Shao (1992), Rancourt *et al.* (1994), Rao & Sitter (1995) and Rao (1996) to mention a few. Incorporation of variance correction procedures in the GES is emphasized in Lee, Rancourt & Särndal (1994), Gagnon *et al.* (1996), and Gagnon *et al.* (1997). A preliminary version of such a system, called SIMPVAR, exists. This

system is currently not integrated with GES. The inputs required for GES include sampling design, type of estimator, auxiliary variables, stratum information, population parameters to be estimated, domain definitions, design weights, g-factors, and point estimates to be produced. SIMPVAR requires additional inputs to define the imputation process: imputation method, imputation groups, auxiliary data used for imputation, respondent flags, and donor identifiers for donor imputation methods.

The framework for the procedures in SIMPVAR is as follows. A sample s is drawn from a population U, yielding design weight w_k, and an estimator of the population total Y_d, for a given domain $U_d \subseteq U$, would be computed as $\hat{Y}_d = \sum_{s_d} \widetilde{w}_k y_k$ if all units in the sample had responded. Here, the calibration weight $\widetilde{w}_k = w_k g_k$ includes any use of auxiliary data via calibration (g_k). If we denote the response set for a given sample s as s_r, and the nonresponse set as s_o, then the estimator of total for the given domain U_d is given by $\hat{Y}_d^{\oplus} = \sum_{s_d} \widetilde{w}_k y_k^{\oplus}$, where $y_k^{\oplus}$ takes the value y_k if $k \in s_r$, and y_k^{IMP} otherwise. Noting that the difference between $\hat{Y}_d^{\oplus}$ and Y_d can be expressed as $\hat{Y}_d^{\oplus} - Y_d = \left(\hat{Y}_d - Y_d\right) + \left(\hat{Y}_d^{\oplus} - \hat{Y}_d\right)$, the estimated variance for $\hat{Y}_d^{\oplus}$ is simply $v_{\text{TOT}}\left(\hat{Y}_d^{\oplus}\right) = v_{\text{SAMP}}\left(\hat{Y}_d^{\oplus}\right) + v_{\text{IMP}}\left(\hat{Y}_d^{\oplus}\right) + 2\text{cov}\left(\hat{Y}_d^{\oplus}, \hat{Y}_d\right)$. The first term $v_{\text{SAMP}}\left(\hat{Y}_d^{\oplus}\right)$ is computed using the sample design and the imputed data set; that is

$$v_{\text{SAMP}}\left(\hat{Y}_d^{\oplus}\right) = \sum_{k \neq \ell \in s} \sum (1 - f_{k\ell})\, z_{k|x}^{\oplus}(d) z_{\ell|x}^{\oplus}(d),$$

where $z_{k|x}^{\oplus}(d) = \widetilde{w}_k e_{k|x}^{\oplus}(d)$ and the residuals $e_{k|x}^{\oplus}(d)$ are obtained from the fit of $y_k^{\oplus}(d)$ on the auxiliary data vectors $\boldsymbol{x}_k$, $k = 1, \ldots, n$. Denoting the imputed values by γ_k, where x_k may be the $\boldsymbol{x}_k$ values themselves, the resulting estimated imputation variance is given by

$$v_{\text{IMP}}\left(\hat{Y}_d^{\oplus}\right) = \sum_{k \neq \ell \in s} \sum (1 - f_{k\ell})\, z_{k|\gamma}^{\oplus}(d) z_{\ell|\gamma}^{\oplus}(d),$$

where $z_{k|\gamma}^{\oplus}(d) = \widetilde{w}_k e_{k|\gamma}^{\oplus}(d)$ and the residuals $e_{k|\gamma}^{\oplus}(d)$ are obtained from the fit of $y_k^{\oplus}(d)$ on the auxiliary data vectors x_k when $k \in s_r$, and are zero otherwise. The covariance $\text{cov}\left(\hat{Y}_d^{\oplus}, \hat{Y}_d\right)$ is estimated similarly.

7 Special Topics

In this section we briefly describe developments related to a few selected estimation topics that have occurred in recent years, namely, multiple frame estimation, distribution function and quantile estimation, two-phase sampling, and the treatment of outliers. Some other areas of research related to estimation not discussed here include record linkage (Fellegi & Sunter, 1969; Armstrong & Mayda, 1994), statistical matching of survey data files (Singh *et al.*, 1993; Kovačević & Liu, 1994; Liu & Kovačević, 1996, 1997), confidentiality and disclosure issues (Robertson & Schiopu-Kratina, 1997; Boudreau, 1997), and edit and imputation methods (Fellegi & Holt, 1976; Kovar & Whitridge, 1995; Granquist & Kovar, 1997; Bankier *et al.*, 1997b). Also not discussed are a number of research initiatives in the time series area.

7.1 Multiple Frame Estimation

In a multiple-frame survey, the overall sample consists of the union of samples selected from separate overlapping sampling frames. The union of these frames results in a frame that represents the target population of interest. Common examples of dual-frame methodology include the case where one frame can be sampled cheaply but does not represent the whole population, whereas

the other frame has complete coverage but is expensive to sample. The use of multiple-frame methodology in this context often arises in agricultural surveys where the incomplete frame is a list frame that is out of date, and the more complete, but more expensive frame is an area frame. Various estimators of the population total are possible, including those proposed by Hartley (1962), Lund (1968), and Fuller & Burmeister (1972). Those estimators optimize the linear combination of the estimators for the overlap portion of those frames. Bankier (1986) examined a variant of the traditional multiple frame problem in the context of sampling administrative tax files. The problem is as follows. An initial stratified simple random sample is selected from a tax data frame that is incomplete, out-of-date, but that can be stratified by industry, geography and size. A subsequent simple random sample stratified by size only is selected from a second administrative tax file that is complete with respect to coverage. Bankier's approach differs from other procedures in that he computes a weight that is the inverse of the probability of sample units for the overlapped unit, yielding a Horvitz–Thompson estimator of the total.

Stukel *et al.* (1997) adapted the Skinner & Rao (1996) "pseudo"-maximum likelihood estimator of a population total in the context of a dual frame. The use of the Skinner–Rao procedure was in the context of pooling of two frames. The procedure was chosen because it yields the same weights adjustment for all variables, which was particularly advantageous for the survey of interest (NPHS), a multipurpose survey with scores of variables of interest.

7.2 CDF and Quantile Estimation

Recently, research has been conducted to examine the application of the generalized regression (GREG) approach to cumulative distribution function (CDF) and quantile estimation (Kovačević, 1997). It is shown that a simple GREG estimator is asymptotically both design and model unbiased. The problem of possibly negative g-factors is solved by imposing range restrictions and using an iterative calibration procedure (Deville *et al.*, 1993). By correcting the simple calibration estimator for the estimated model-bias, an estimator with improved model properties is obtained. This estimator is more efficient than the modified difference estimator suggested by Rao *et al.* (1990) due to an additional utilization of auxiliary variables through the use of 'g-factors'. However, it is very cumbersome to compute either of these estimators when the population is large. When these estimators are applied for quantile estimation the computational burden increases by the order of the sample size. In that perspective, the simple calibration estimator comes out as a practical and efficient alternative despite slightly worse asymptotic properties. A small empirical study (Kovačević, 1997) confirmed that the inversion of the simple calibration estimator of the CDF resulted in acceptably stable and accurate estimates of quantiles. Overall, the research showed that the simple calibration estimator of the CDF, which is computationally easy to obtain and which can be easily implemented within the Generalized Estimation System (GES) of Statistics Canada performs comparably well for both CDF and quantile estimation.

7.3 Treatment of Outliers

Outliers are a common occurrence in sample surveys of highly skewed populations. The impact of an outlier on the estimate of total can be quite significant. Procedures have been developed to detect and treat such units mostly in the context of simple random sampling. The detection of outliers has been based on computing ordered residuals standardized by the median and the interquartile distance as measures of location and scale respectively. Approaches that have been developed or studied for treating outliers include changing the values of the outliers by Winsorization, (Fuller, 1991), adjusting the sampling weights (Hidiroglou & Srinath, 1981, and Ghangurde, 1989), and using robust estimation procedures (Gwet & Rivest, 1992; Lee, 1991, and Chambers & Kokic, 1993). The

robust estimation techniques estimate a given total by adding to the sum of the sampled observations the sum of the "robustified" non-sampled observations. Lee (1995) has provided an excellent review of these procedures. Lee & Patak (1998) proposed and studied the properties of a robustified GREG estimator that is design consistent.

8 Future Directions

Data analysis and data integration are among the primary foci of methodology research and development at Statistics Canada, in light of the need to enhance the quality of our analytical outputs and to respond to the continuing environment of restrained resources. Efforts in data analysis include continuing the development of suitable methods and systems, including their incorporation in GES, as well as increasing the awareness and use of such products among analysts (through our Data Analysis Resource Centre, a unit in the Methodology Branch) for analyzing data from our major cross-sectional surveys.

There is a need to develop software for analytic statistics in complex surveys as discussed in Skinner *et al.* (1989). These include suitable chi-squared tests for contingency tables (Fellegi, 1980; Rao & Scott, 1984). The software packages PC-CARP, SUDAAN, WESVAR and STATA provide a limited range of statistics such as regression, logistic regression, and quantile estimation. The range can be increased dramatically by automating the derivation of the required variance expressions. Automation begins with the associated estimating equations, followed by use of the sandwich estimator given by Binder (1983).

With regard to the analysis of data from longitudinal surveys, the parameters involved in the understanding and interpretation of various patterns of social and economic change are much more complex than the cross-sectional parameters, as they refer to observations over time. Examples include modelling of length of spells (e.g., unemployment, poverty, diseases), inference about transitions from one state to another (school-to-work, gross changes, health status), development of new measures and indices (health indicators, income cut-offs), inference about regression coefficients and odds ratios, etc. It should further be stressed that estimation issues cannot be isolated from sample design issues in longitudinal surveys. Accordingly, research is required to determine the appropriate duration of a panel, the need for overlapping panels and split-panel designs, the number of waves, and other design features, such as sample allocation using suitable cost-variance models.

Data integration efforts include not only the integration of surveys at the design and data collection stages to optimize cost efficiencies and minimize respondent burden but also dealing with issues such as harmonization of concepts and definitions and internal consistencies (e.g., consistency of employment data from LFS, SEPH and the Unified Enterprise Survey, of income data from SLID and tax records, and of expenditure data from the Survey of Household Spending and other sources). Statistics Canada's 1999 methodology symposium on "Combining Data from Different Sources" is indicative of the importance we put on this topic.

Research on this topic includes continued emphasis on development of suitable methods and systems for record linkage, statistical matching, multiple frame estimation, benchmarking and refined calibration techniques. Also, research is needed on combining information for small area estimation. Sample size dependent estimators provide reasonably good estimates for regional/medium-sized domains but not for local/smaller domains. Studies are required on the determination of stable and operationally feasible weights for (design-model) combined estimators and also on the design bias of such estimators.

Confidentiality issues cut across all these topics and suitable methods for dealing with them need to be developed, particularly in the context of longitudinal data and public use microdata files. There is also a need to continue the research on variance estimation for data from such files, as described in section 6. We need to develop methods that can be applied to complex statistics from regressions,

logistic regressions and contingency tables.

Finally, in addition to the continuing investigation of estimation techniques for sampling error, there is a need to re-emphasize the importance of the measurement of total error, especially in light of the adoption of computer-assisted interviewing in our major cross-sectional and longitudinal surveys. This entails building on earlier work (e.g., Fellegi, 1974) to develop techniques for estimating factors such as interviewer effects that affect the quality of our products.

Acknowledgement

We would like to thank Editor Vijay Nair and two referees. Based on their comments, the structure and readability of the paper were improved considerably.

References

Andersson, C. & Nordberg, L. (1994). A method for variance estimation of non-linear functions of totals in surveys. Theory and software implementation. *Journal of Official Statistics*, **10**, 395–405.

Armstrong, J.B. & Mayda, J.E. (1994). Model-based estimation of record linkage error rates. *Survey Methodology*, **19**, 137–148.

Armstrong, J. & St-Jean, H. (1994). Generalized regression estimation for a two-phase sample of tax records. *Survey Methodology*, **20**, 91–105.

Bankier, M.D. (1986). Power allocations: determining sample sizes for subnational areas. *The American Statistician*, **42**, 174–177.

Bankier, M., Houle, A.M. & Luc, M. (1997a). Calibration estimation in the 1991 and 1996 Canadian censuses. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, pp. 66–75.

Bankier, M., Houle, A.M., Luc, M. & Newcombe, P. (1997b). 1996 Canadian Census demographic variables imputation. *Proceedings of the Section on Survey Research Methods Section, American Statistical Association*, pp. 389–394.

Bankier, M.D., Rathwell, S. & Majkowski, M. (1992). Two step generalized least squares estimation in the 1991 Canadian Census. *Proceedings of the Workshop on Uses of Auxiliary Information in Surveys*, Statistics Sweden, Örebro.

Bethlehem, J.G. & Keller, W.K. (1987). Linear weighting of sample survey data. *Journal of Official Statistics*, **3**, 141–153.

Binder, D.A. (1983). On the variance of asymptotically normal estimators from complex surveys. *International Statistical Review*, **51**, 279–292.

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

Binder, D.A. (1998). Longitudinal surveys: Why are these surveys different from all other surveys. *Survey Methodology*, **24**, 101–108.

Binder, D.A., Babyak, C., Brodeur, M., Hidiroglou, M. & Jocelyn, W. (2000). Variance Estimation for Two-Phase Stratified Sampling. *The Canadian Journal of Statistics*, **28**, 751–764.

Binder, D.A. & Kovačević, M.S. (1995). Estimating some measures of income inequality from survey data: An application of the estimating equations approach. *Survey Methodology*, **21**, 137–146.

Binder, D.A. & Patak, Z. (1994). Use of Estimating Functions For Estimation From Complex Surveys. *Journal of American Statistical Association*, **89**, 1035–1043.

Bleuer, R.S. & Kovačević, M.S. (1999a). Some Issues In the Estimation of Income Dynamics. *Survey Methodology*, **25**, 87–98.

Bleuer, R.S. & Kovačević, M.S. (1999b). Variance estimation in longitudinal studies of income dynamics. *Proceedings: Symposium 98, Longitudinal Data Analysis*, pp. 123–128. Statistics Canada, Ottawa, Canada.

Boudreau, J.-R. (1997). Assessing the risk of disclosure by modelling the number of subpopulations of the same size. *Proceedings of the Survey Methods Section. Statistical Society of Canada*, pp. 15–24.

Brackstone, G.J. (1987a). Small area data: Policy issues and technical challenges. In *Small Area Statistics*, Eds. R. Platek, J.N.K. Rao, C.-E. Särndal and M.P. Singh, pp. 3–20. New York: Wiley and Sons.

Brackstone, G.J. (1987b). Statistical Uses of Administrative Data: Issues and Challenges. *Proceedings: Symposium 87, Statistical uses of Administrative Data*, pp. 6–16. Statistics Canada.

Brackstone, G.J. & Rao, J.N.K. (1976). Raking ratio estimation. *Survey Methodology*, **2**, 63–69.

Brackstone, G.J. & Rao, J.N.K. (1979). An investigation of raking ratio estimations. *Sankhyā, Series C*, Part 2, **41**, 97–114.

Brewer, K.R.W. & Hanif, M. (1983). *Sampling with Unequal Probabilities*. New York: Springer-Verlag.

Chambers, R. & Kokic, P. (1993). Outlier robust sample survey inference. *Proceedings of the 49th Session of the International Statistical Institute*, Firenze, Vol. 2, 55–72.

Choudhry, G.H., Lavallée, P. & Hidiroglou, M. (1989). Two-Phase sample design for tax data. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, pp. 646–651.

Choudhry, G.H. & Rao, J.N.K. (1989). Small area estimation using models that combine time series and cross-sectional data. *Proceedings: Symposium 89, Analysis of Data in Time*, pp. 67–74. Statistics Canada.

Deville, J.-C. & Särndal, C.E. (1992). Calibration estimators in survey sampling. *Journal of the American Statistical Associ-*

ation, **87**, 376–382.

Deville, J.-C., Särndal, C.-E. & Sautory, O. (1993). Generalized raking procedures in survey sampling. *Journal of the American Statistical Association*, **88**, 1013–1020.

Dick, P. (1995). Modelling net undercoverage in the 1991 Census. *Survey Methodology*, **21**, 45–54.

Drew, J.D., Singh, M.P. & Choudhry, G.H. (1982). Evaluation of small area estimation techniques for the Canadian Labour Force Survey. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, pp. 545–550.

Dufour, J., Gagnon, F., Morin, Y., Renaud, M. & 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*, pp. 552–557.

Estevao, V., Hidiroglou, M.A. & Särndal, C.-E. (1995). Methodological principles for a generalized estimation system at Statistics Canada. *Journal of Official Statistics*, **11**, 181–204.

Fellegi, I.P. & Holt, D. (1976). A systematic approach to automatic edit and imputation. *Journal of the American Statistical Association*, **71**, 17–35.

Fellegi, I.P. & Sunter, A.B. (1969). A theory of record linkage. *Journal of the American Statistical Association*, **64**, 1183–1210.

Fellegi, I.P. (1974). An improved method of estimating the correlated response variance. *Journal of the American Statistical Association*, **69**, 496–501.

Fellegi, I.P. (1980). Approximate tests of independence and goodness of fit based on stratified multistage samples. *Journal of the American Statistical Association*, **75**, 261–268.

Fuller, W.A. (1991). Simple estimators of the mean of skewed population. *Statistica Sinica*, **1**, 137–138.

Fuller, W.A. & Burmeister, L. (1972). Estimators for samples selected from two overlapping frames. *Proceedings of Social Statistics Section, American Statistical Association*, pp. 245–249.

Fuller, W.A. (1998). Suggestion made during the meeting of the Advisory Committee on Statistical Methods of Statistics Canada, held October 19 and 20, 1998, in Ottawa, Canada.

Gagnon, F., Lee, H., Provost, M., Rancourt, E. & Särndal, C.-E. (1997). Estimation of variance in the presence of imputation. In *Proceedings: Symposium 97, New Directions in Surveys and Censuses*. Statistics Canada, Ottawa.

Gagnon, F., Lee, H., Rancourt, E. & Särndal, C.-E. (1996). Estimating the variance of the generalized regression estimator in the presence of imputation for the generalized estimation system. *Proceedings of the Survey Methods Section, Statistical Society of Canada*, pp. 151–156.

Gambino, J. & Dick, P. (2000). Small area estimation practice at Statistics Canada. *Statistics in Transition*, **4**, 597–610.

Gambino, J.G., Singh, M.P., Dufour, J., Kennedy, B. & Lindeyer, J. (1998). *Methodology of the Canadian Labour Force Survey*. Statistics Canada. Catalogue No. 71-526.

Ghangurde, P.D. (1989). Outliers in sample surveys. *Proceedings of the Survey Research Methods Section, American Statistical Association*, pp. 736–739.

Ghangurde, P.D. & Singh, M.P. (1977). Synthetic estimates in periodic household surveys. *Survey Methodology*, **3**, 152–181.

Ghangurde, P.D. & Singh, M.P. (1978). Evaluation of efficiency of synthetic estimates. *Proceedings of the Social Statistics Section, American Statistical Association*, pp. 53–61.

Ghosh, M. & Rao J.N.K. (1994). Small-area estimates—an appraisal. *Statistical Science*, **8**(1), 55–93.

Gonzalez, M.E. (1973). Use and evaluation of synthetic estimators. *Proceedings of Social Statistics Section, American Statistical Association*, pp. 33–36.

Granquist, I. & Kovar, J.G. (1997). Editing of survey data: how much is enough? In *Survey Measurement and Process Quality*, Eds. L. Lyberg, P. Biemer, M. Collins, E. deLeeuw, C. Dippo, N. Schwarz and D. Trewin, chapter 18, pp. 415–436. New York: John Wiley & Sons.

Gwet, J.-P. & Rivest, L.-P. (1992). Outlier resistant alternatives to the ratio estimator. *Journal of the American Statistical Association*, **87**, 736–739.

Hapuarachchi, K.P. (1997). Proportional hazards model for survey data: An empirical investigation. *Proceedings of the Survey Methods Section, Statistical Society of Canada Annual Meeting*, pp. 67–72.

Hartley, H. (1962). Multiple frame surveys. *Proceedings of the Social Statistics Section, American Statistical Association*, pp. 203–206.

Hidiroglou, M.A. (1995). Sampling and estimation for stage one of the Canadian Survey of Employment, Payrolls and Hours survey redesign. *Proceedings of the Survey Methods Section, Statistical Society of Canada*, pp. 123–128.

Hidiroglou, M.A., Latouche, M., Armstrong, B. & Gossen, M. (1995). Improving survey information using administrative records: the case of the Canadian employment survey. *Proceedings of the Annual Research Conference Proceedings*, pp. 171–197. Washington: Bureau of the Census.

Hidiroglou, M.A. & Särndal C.-E. (1998). Use of auxiliary information for two-phase sampling. *Survey Methodology*, **24**, 11–20.

Hidiroglou, M.A. & Srinath, K.P. (1981). Some estimators of population total category large units. *Journal of the American Statistical Association*, **78**, 690–695.

Huang, E. & Fuller, W.A. (1978). Nonnegative regression estimators. *Proceedings of the Social Statistics Section, American Statistical Association*, pp. 300–305.

Kalton, G. & Citro, C.F. (1993). Panel surveys: adding the fourth dimension. *Survey Methodology*, **19**, 205–215.

Kasprzyk, D., Duncan, G., Kalton, G. & Singh, M.P. (Eds.) (1989). *Panel Surveys*. New York: Wiley.

Kott, P.S. (1995). Can the jackknife be used with a two-phase sample? *Proceedings of the Survey Methods Section, Statistical Society of Canada*, pp. 107–110.

Kott, P.S. & Stukel, D.M. (1997). Can the jackknife be used with a two-phase sample? *Survey Methodology*, **23**, 81–90.

Kovačević, M.S. (1997). Calibration estimation of cumulative distribution function and quantiles from survey data. *Proceedings of the Survey Methods Section, Statistical Society of Canada Annual Meeting*, pp. 139–144.

Kovačević, M.S. (1999). Tools for inference on dynamics of low income status. *Proceedings: Symposium 98, Longitudinal Data Analysis*, pp. 101–107. Statistics Canada, Ottawa.

Kovačević, M.S. & Binder, D.A. (1997). Variance estimation for measures of income inequality and polarization—the estimating equations approach. *Journal of Official Statistics*, **13**, 41–58.

Kovačević, M.S. & Liu, T.P. (1994). Statistical matching of survey datafiles: a simulation study. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, pp. 479–484.

Kovačević, M.S. & Yung, W. (1997). Variance estimation for measures of income inequality and polarization—an empirical study. *Survey Methodology*, **23**, 41–52.

Kovar, J.G. & Whitridge, P.J. (1995). Imputation of establishment survey data. In *Business Survey Methods*, Eds. B.G. Cox, D.A. Binder, N. Chinnappa, A. Christianson, M.J. Colledge and P.S. Kott, pp. 403–423. New York: Wiley.

Latouche, M. & Michaud, S. (1997). Concerns pertaining to weighting of longitudinal surveys. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, pp. 111–119.

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

Lavallée, P. & Hunter, L. (1992). Weighting for the survey of labour and income dynamics. *Proceedings: Symposium 92, Design and Analysis of Longitudinal Surveys*, pp. 65–78. Statistics Canada, Ottawa.

Lee, H. (1991). Model-based Estimators that are robust to outliers. *Proceedings of the Annual Research Conference*, pp. 178–202. Washington, DC: Bureau of the Census.

Lee, H. (1995). Outliers in Business Surveys. In *Business Survey Methods*, Eds. B.G. Cox, D.A. Binder, B.N. Chinnappa, A. Christianson, M.J. Colledge and P.S. Kott, pp. 503–526.

Lee, H. & Patak, Z. (1998). Outlier Robust Generalized Regression Estimator. *Proceedings of the Survey Methods Section, Statistical Society of Canada*, pp. 241–247.

Lee, H., Rancourt, E. & Särndal, C.-E. (1994). Experiments with variance estimation from survey data with imputed values. *Journal of Official Statistics*, **10**, 231–243.

Lemaître, G. & Dufour, J. (1987). An integrated method for weighting persons and families. *Survey Methodology*, **13**, 199–207.

Liu, T.P. & Kovačević, M.S. (1996). Categorically constrained matching. *Proceedings of the Survey Methods Section, Statistical Society of Canada*, pp. 123–134.

Liu, T.P. & Kovačević, M.S. (1997). An empirical study on categorically constrained matching. *Proceedings of the Survey Methods Section, Statistical Society of Canada*, pp. 167–178.

Lund, R.E. (1968). Estimators in multiple frame surveys. *Proceedings of the Social Statistics Section, American Statistical Association*, pp. 282–288.

Mayda, J.E., Mohl. C. & Tambay, J.-L. (1996). Variance estimation and confidentiality: They are related! *Proceedings of the Survey Methods Section, Statistical Society of Canada*, pp. 135–142.

Merkouris, T. (1997). On cross-sectional estimation for repeated panel household surveys. *Proceedings of the Survey Methods Section, Statistical Society of Canada*, pp. 93–98.

Michaud, S. & Hunter, L. (1992). Strategy for minimizing the impact of nonresponse for the survey of labour and income dynamics. *Proceedings: Symposium 92, Design and Analysis of Longitudinal Surveys*, pp. 89–98. Statistics Canada, Ottawa, Canada.

National Center for Health Statistics (1968). *Synthetic State Estimates of Disability*. U.S. Government Printing Office, Washington, D.C.

Outrata, E. & Chinnappa, B.N. (1989). General survey functions at Statistics Canada. *Bulletin of the International Statistical Institute*, **53**(2), 219–238.

Patak, Z., Hidiroglou, M. & Lavallée, P. (1998). The Methodology of the Workplace and Employee Survey. *Proceedings of the Survey Research Methods Section, American Statistical Association*, pp. 83–91.

Pfeffermann, D. & Bleuer, S.R. (1993). Robust joint modelling of labour force series of small areas. *Survey Methodology*, **19**, 149–164.

Platek, R., Rao, J.N.K., Särndal, C.-E. & Singh, M.P. (1987). *Small Area Statistics*. New York: Wiley.

Purcell, N.J. & Kish, L. (1979). Estimation for small domains. *Biometrics*, **35**, 365–384.

Rancourt, E. & Hidiroglou, M.A. (1998). Use of administrative records in the Canadian Survey of Employment, Payrolls and Hours. *Proceedings of the Survey Methods Section, Statistical Society of Canada*, pp. 39–48.

Rancourt, E., Särndal, C.E. & Lee, H. (1994). Estimation of variance in presence of nearest neighbour imputation. *Proceedings of the Survey Research Methods Section, American Statistical Association*, pp. 888–893.

Rao, J.N.K. (1996). On variance estimation with imputed survey data. *Journal American Statistical Association*, **91**, 499–506.

Rao, J.N.K., Kovar, J.G. & Mantel, H.J. (1990). On estimating distribution function and quantile from survey data using auxiliary information. *Biometrika*, **77**, 365–375.

Rao, J.N.K. & Shao, J. (1992). Jackknife variance estimation with survey data under hotdeck imputation. *Biometrika*, **79**, 811–822.

Rao, J.N.K. & Scott, A.S. (1984). On chi-squared tests for multiway contingency tables with cell proportions estimated from survey plate. *Annals of Statistics*, **12**, 46–60.

Rao, J.N.K. & Singh, A.C. (1997). A ridge-shrinkage method for range-restricted weight calibration in survey sampling. *Proceedings of the Survey Research Methods Section, American Statistical Association*, pp. 57–65.

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

Robertson, D. & Schiopu-Kratina, J. (1997). The mathematical basis for Statistics Canada cell suppression software: CONFID. *Proceedings of the Survey Methods Section, Statistical Society of Canada Annual Meeting*, pp. 7–14.

Royce, D. (1992). A comparison of some estimators of a set of population totals. *Survey Methodology*, **18**, 109–125.

Rust, K.F. & Rao, J.N.K. (1996). Variance estimation for complex surveys using replication techniques. *Statistical Methods in Medical Research*, **5**, 283–310.

Särndal, C.-E. & Hidiroglou, M.A. (1989). Small domain estimation: A conditional analysis. *Journal of the American Statistical Association*, **84**, 266–275.

Särndal, C.-E., Swensson, B. & Wretman, J.H. (1992). *Model Assisted Survey Sampling*. Springer-Verlag.

Schaible, W.L. (1979). A composite estimator for small area statistics. *Synthetic Estimates for Small Area*, NIDA Research Monograph Series 24, U.S. Department of Health, Education and Welfare, Library of Congress Catalogue No. 79–600067, pp. 36–53.

Schaible, W.L. (1992). Use of small area estimators in U.S. federal programs. *Small Area Statistics and Survey Designs*, Volume 1, 95–114. Eds. G. Kalton, J. Kordos and R. Platek. Warsaw: Central Statistical Office.

Schaible, W.L., Editor (1996). *Indirect estimators in U.S. federal programs*. New York: Springer-Verlag.

Schnell, D., Kennedy, W.J., Sullivan, G., Park, J.P. & Fuller, W.A. (1988). Personal Computer variance software for complex surveys. *Survey Methodology*, **14**, 59–69.

Shah, B.V., Lavange, L.M., Barnwell, B.G., Killinger, J.E. & Wheeless, S.C. (1989). *SUDAAN: Procedures for Descriptive Statistics Users' Guide*. Research Triangle Park: Research Triangle Institute Report.

Singh, A.C., Kennedy, B., Wu, S. & Brisebois, F. (1997). Composite estimation for the Canadian Labour Force Survey. *Proceedings of the Survey Research Methods Section, American Statistical Association*, pp. 300–305.

Singh, A.C. & Mantel, H. (1991). State space composite estimation for small areas. *Proceedings: Symposium 91, Spatial Issues in Statistics*, pp. 17–25. Statistics Canada.

Singh, A.C., Mantel, H.J., Kinack, M. & Rowe, G. (1993). Statistical matching: use of auxiliary information as an alternative to the conditional independence assumption. *Survey Methodology*, **19**, 59–79.

Singh, A.C., Mantel, H.J. & Thomas, B.W. (1994). Time series EBLUPs for small areas using survey data. *Survey Methodology*, **20**, 33–43.

Singh, A.C. & Mian, I. (1995). Generalized sample size dependent estimators for small areas. *Proceedings, U.S. Census Bureau, Annual Research Conference*, pp. 687–701.

Singh, A.C. & Mohl, C. (1996). Understanding calibration estimation in survey sampling. *Survey Methodology*, **22**, 107–115.

Singh, A.C., Stukel, D.M. & Pfeffermann, D. (1998). Bayesian versus frequentist measures of error in small area estimation. *Journal of the Royal Statistical Society, Series B*, **60**(2), 377–396.

Singh, A.C. & Wu, S. (1998). Hierarchical covariance modelling for nonlinear regression with random parameters. *Proceedings of the Survey Research Methods Section, American Statistical Association*, pp. 64–73.

Singh, A.C., Wu, S. & Boyer, R. (1995). Longitudinal Survey Nonresponse Adjustment by Weight Calibration. *Proceedings of the Survey Research Methods Section, American Statistical Association*, pp. 396–401.

Singh, M.P., Drew, J.D., Gambino, J.G. & Mayda, F. (1990). *Methodology of the Canadian Labour Force Survey*. Statistics Canada, Catalogue No. 71-526.

Singh, M.P., Gambino, J.G. & Mantel, H. (1992). Issues and options in the provision of small area statistics. *Small Area Statistics and Survey Designs*, **1**, 37–75, Eds. G. Kalton, J. Kordos and R. Platek. Warsaw: Central Statistical Office.

Singh, M.P., Gambino, J. & Mantel, H. (1994). Issues and strategies for small area data (with discussion). *Survey Methodology*, **20**, 3–22.

Singh, M.P. & Tessier, R. (1976). Some estimators for domain totals. *Journal of the American Statistical Association*, **71**, 322–325.

Skinner, C.J. & Rao, J.N.K. (1996). Estimation in dual frame surveys with complex design. *Journal of the American Statistical Association*, **91**(433), 349–356.

Skinner, C.J, Holt, D. & Smith, T.M.F (1989). *Analysis of Complex Surveys*. Chichester: Wiley.

Statistics Canada (1997). National Longitudinal Survey of Children & Youth. Overview of Survey Instruments for 1996–97 Data Collection Cycle 2. Catalogue 89F0078XIE.

Stukel, D., Hidiroglou, M. & Särndal, C.-E. (1996). Variance estimation for calibration estimators: a comparison of jackknifing versus Taylor linearization. *Survey Methodology*, **22**, 117–126.

Stukel, D.M., Mohl, C.A. & 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 Annual Meeting*, pp. 111–116.

Tambay, J-L. & Catlin, G. (1995). Sample design of the National Population Health Survey. *Health Reports*, **7**, 29–38.

Tambay, J-L., Schiopu-Kratina, I., Mayda, J., Stukel, D. & Nadon, S. (1997). Treatment of Nonresponse in Cycle Two of the National Population Health Survey. *Survey Methodology*, **24**, 147–156.

Théberge, A. (2000). Calibration and restricted weights. *Survey Methodology*, **26**, 99–107.

Yung, W. (1997). Variance estimation for public use files under confidentiality constraints. *Proceedings of the Survey Research Methods Section, American Statistical Association*, pp. 434–439.

Yung, W. & Rao, J.N.K. (1996). Jackknife linearization variance estimators under stratified multi-stage sampling. *Survey Methodology*, **22**, 23–32.

Résumé

Cet article donne une vue d'ensemble de la recherche portant sur les techniques d'estimation, leur application, et le développement de systèmes généralisés pour l'estimation à Statistique Canada. Au Canada, la demande pour des données transversales plus détaillées et de meilleure qualité touchant de nombreuses questions socio-économiques a augmenté considérablement ces dernières années. Aussi, on dénote l'intérêt croissant pour ces données longitudinales afin de mieux comprendre et interpréter les relations entre les variables, et nécessitant la mise en œuvre de plusieurs grandes enquêtes par

panel à Statistique Canada. Cet article discute brièvement de l'estimation pour des données longitudinales ainsi que d'une approche de pondération transversale pour les données provenant de ces enquêtes. On discute brièvement des estimateurs par calage appropriés pour des enquêtes transversales ménages ou d'entreprises, ainsi que pour le recensement de la population. De plus, on présente l'estimateur de régression composite, une méthode développée afin d'améliorer la qualité des estimations transversales pour des enquêtes avec rotation de panels telles que l'enquête canadienne sur la population active. On présente aussi plusieurs approches pour obtenir des estimations transversales plus détaillées au niveau infra-provincial, c'est-à-dire pour des petites régions. Nous résumons plusieurs modules développés pour le Système d'Estimation Généralisé. De nouveaux développements importants pour ce système tels que l'estimation à deux phases et l'estimation de la variance pour l'imputation sont présentés. Nous examinons brièvement le statut actuel de la recherche sur l'estimation portant sur un ensemble de sujets précis, ainsi que la direction de la recherche future.

[*Received April 1999, accepted January 2001*]