GES: AN ESTIMATION SYSTEM IN DEVELOPMENT AT STATISTICS CANADA

Hyunshik Lee, Mike Hidiroglou and Victor Estevao, Statistics Canada
Hyunshik Lee, 11-Q R.H. Coats Bldg., Statistics Canada, Ottawa K1A 0T6

KEY WORDS: Generalized regression estimator; model-assisted approach; design consistency; domains; g-weights.

1. Introduction

We are currently developing a Generalized Estimation System (GES) as a part of the General Survey Function Development at Statistics Canada. This initiative has also led to the earlier development of the Generalized Edit and Imputation System (GEIS) and the more recent Generalized Sampling System (GSAM).

The rationale for the development of general systems is described in several papers such as Outrata and Chinnappa (1989). The GES project is an effort to provide an estimation system that can be used by most of the surveys conducted at Statistics Canada. Many of these surveys have common features. But, until recently, almost all surveys used their customized estimation systems. While this approach has provided the flexibility to meet specific requirements of each survey, many resources have been spent in the development and maintenance of these systems. The system maintenance costs have been significant because of the acquisition and upgrading of different software and hardware products. Also, there has been a constant need to train new system developers due to staff rotation on the projects. Also, because these systems have evolved over several years independently of one another, they tend to reflect different system architectures and methodologies. The development of generalized systems such as GES is a concerted effort to reduce these costs and to standardize development strategies and methodologies.

An important benefit of this development is that it has provided us with a focal point for discussion of existing methodologies. When we set out to develop the GES we were well aware of the complex nature of estimation theory. Several estimation software packages have been developed elsewhere using different approaches to the methodology framework. These include LINWEIGHT (Bethlehem and Keller, 1987), PC-CARP (Schnell et al., 1988), SUDAAN (Shah et al., 1989), ISSA (Rojas and Aliaga, 1993) and others. We decided to adopt a framework based on the theory of the generalized regression estimator (Särndal, Swensson and Wretman, 1992). This has allowed us to classify and use a large family of estimator functions through the specification of a general regression model. The theory permits us to use auxiliary information to improve on the efficiency of the estimators while achieving consistency with the known auxiliary totals. The use of auxiliary data is particularly important because of its availability in many surveys. We characterize a generalized regression estimator through the concepts of model level, model groups, model auxiliary variables and model variance. This provides us with a structure that includes many traditional estimators such as the combined and separate Horvitz-Thompson and ratio estimators, post-stratified estimators as well as more complex estimators such as raking ratio.

Currently, GES produces point estimates and associated estimates of reliability (standard error and cv) for domain size or domain totals, ratios and means of variables of interest. A domain can be defined as any subpopulation of the survey population. Variance estimation is currently based on the model-assisted approach using a Taylor linear approximation for non-linear parameters such as a ratio. This approach is design consistent. The methodology behind the system is described in more detail in Section 2: GES Methodology.

The GES is currently being developed as a microcomputer application. We are using the SAS System to develop the source code, create the GES selection menus and carry out data base management. The menus provide a simple user interface to the GES. The most recent version (GES2.2) runs under SAS 6.08 for Windows and needs the products SAS/BASE, FSP and IML. More information concerning the system structure and hardware and software requirements is given in Section 3: GES Structure and Environment.

The sample design is another important element in the framework of the GES. The current version fits the estimation requirements for stratified one-stage element or cluster design under simple random sampling without replacement. In the future, we plan to accommodate other designs. These include: stratified one-stage element and cluster designs with selection proportional to size, with or without replacement, as well as stratified multi-stage designs. We are examining several possible extensions of the GES. In the future, additional options for variance

estimation will be incorporated. Extensions to the methodology framework regarding calibration estimators (Deville and Särndal, 1992), and additional functions such as outlier detection and treatment, will be included. A discussion of future plans is given in Section 4: Future Development. Some concluding remarks are given in Section 5.

2. GES Methodology

2.1 Population, Sample and Model Groups

We introduce notation to discuss the methodology. Let $U = \{1, ..., k, ..., N\}$ denote the index set for the N units of a finite population. We denote by s a probability sample of units drawn from U by a given sampling design. The inclusion probabilities induced by this design are denoted $\pi_k = P(k \in s)$ and $\pi_{k\ell} = P(k \& \ell \in s)$. We assume that the π_k and the $\pi_{k\ell}$ are known and positive. Set $a_k = 1/\pi_k$, called the sampling weight of the k-th unit.

Let y_k denote the value of a variable of interest, y, for population unit k. The population total of y is denoted $Y = \Sigma_U y_k$. (If A is a set of units, we write Σ_A to denote $\Sigma_{k \in A}$). Estimates of totals are required for the entire population and a variety of domains of interest; domain estimation is considered in Section 2.4.

The Horvitz-Thompson estimators can often be improved with auxiliary information. The auxiliary information considered here are known totals for one or more auxiliary variables. The counting variable that counts the number of units is also treated as an auxiliary variable in this paper if necessary totals are known. These totals may be known for the entire population or for specified subpopulations. The objective is to use this information as efficiently as possible in the estimation.

We use the term *model group* to designate a subpopulation for which auxiliary variable totals are known and to which the model statement is applied. The model groups represent the most detailed level at which auxiliary information is used. Our general notation for a model group is U_p, where $U_p \subset U$. Let x_{pk} be the value for the k-th unit of an auxiliary column vector x_p associated with U_p. More specifically, we call U_p a model group if: (a) the auxiliary value x_{pk} can be observed for every unit $k \in s_p = s \cap U_p$, and (b) the group auxiliary total $X_p = \Sigma_{U_p} x_{pk}$ is known.

A set of model groups, $\{U_p;\ p = 1, ..., P\}$, divides the whole population into mutually exclusive and exhaustive subpopulations. It is possible to have $P = 1$. In this case, the entire population is the only model group. We assume that for every unit $k \in s$ the model group identity and the measurement (y_k, x'_{pk}) are available.

Many commonly used estimators can be justified using a linear regression of y on x_p that can be fitted within each group. Ideally, x_p is a good predictor of the variable of interest y within the model group. However, the structure of x_p can be as simple as $x_{pk} = 1$ for all $k \in U_p$, implying that $X_p = \Sigma_{U_p} x_{pk} = N_p$. The model groups correspond to post-strata (see the following section). The knowledge of the group counts N_p can considerably improve the precision of the estimates. The vector x_p, for which the model group total X_p is known, may be composed of different variables in the different groups, therefore the index p on x_p.

2.2 Regression Approach: The Generalized Regression (GREG) Estimator

For the p-th group, U_p, consider the regression model stating that

$$y_k = x'_{pk} \beta_p + \epsilon_k \quad \text{for} \quad k \in U_p \qquad (1)$$

where $E_\xi(\epsilon_k) = 0$, $\mathrm{Var}_\xi(\epsilon_k) = c_k \sigma^2$, and $\mathrm{Cov}_\xi(\epsilon_k, \epsilon_\ell) = 0$ for all $k \neq \ell$, where the subscript ξ denotes moments with respect to the model. The known constants c_k are determined by the variance structure of the underlying regression model. Here β_p is estimated from the sample s by $\hat{B}_p$, defined as the solution of

$$\sum_{s_p} \frac{a_k x_{pk} x'_{pk}}{c_k} \hat{B}_p = \sum_{s_p} \frac{a_k x_{pk} y_k}{c_k}$$

This represents the system of normal equations when the data $\{(y_k, x'_{pk}): k \in s_p\}$ are used to fit the model (1). The weights a_k in this system of equations serve the purpose of making $\hat{B}_p$ a design consistent estimator of the population regression coefficient vector B_p. The population regression vector is assumed to be the best fit (in the sense of generalized least squares) when all units in U_p are

observed. The regression fit also produces the residuals $e_k = y_k - x'_{pk}\hat{B}_p$ for $k \in s_p = s \cap U_p$. The model group total $Y_p = \Sigma_{U_p} y_k$ is estimated by $\hat{Y}_{p\pi} + (X_p - \hat{X}_{p\pi})'\hat{B}_p$, where $\hat{X}_{p\pi} = \Sigma_{s_p} a_k x_{pk}$ is the Horvitz-Thompson estimator of the known auxiliary group total X_p. (In this paper, estimators identified by a "hat" and the subscript π signifies the Horvitz-Thompson estimator). The total weight given to the k-th unit is the product of the two weights, a_k (design derived) and g_k (auxiliary data derived). That is, the sum of the Horvitz-Thompson estimator $\hat{Y}_{p\pi} = \Sigma_{s_p} a_k y_k$ and a regression adjustment $(X_p - \hat{X}_{p\pi})'\hat{B}_p$. To obtain the estimator of the entire population total, sum over groups, that is

$$\hat{Y}_{GREG} = \sum_{p=1}^{P} \{\hat{Y}_{p\pi} + (X_p - \hat{X}_{p\pi})'\hat{B}_p\} \qquad (2)$$

Note that the above estimator may be written as a weighted linear sum as

$$\hat{Y}_{GREG} = \Sigma_s w_k y_k = \Sigma_{p=1}^{P} \Sigma_{s_p} a_k g_k y_k \qquad (3)$$

where $w_k = a_k g_k$ and

$$g_k = 1 + (X_p - \hat{X}_{p\pi})' \times (\Sigma_{s_p} a_k x_{pk} x'_{pk}/c_k)^{-1} x_{pk}/c_k \qquad (4)$$

The g_k will be referred to as the g-weight.

The regression residuals e_k are needed for computing the estimate of the variance of $\hat{V}(\hat{Y}_{GREG})$ or $\hat{V}$ for short. This variance estimator, is given by

$$\hat{V} = \sum_{k \in s} \sum_{l \in s} \left(\frac{\Delta_{kl}}{\pi_{kl}}\right)\left(\frac{g_k e_k}{\pi_k}\right)\left(\frac{g_l e_l}{\pi_l}\right) \qquad (5)$$

where $\Delta_{kl} = \pi_{kl} - \pi_k \pi_l$, $\pi_{kk} = \pi_k$.

The theoretical justification for g-weighting the residuals in the variance estimator (5) is given in Särndal, Swensson and Wretman (1989). Although (5) defines $\hat{V}$ as a double sum, it is reduced to a single sum in many practical cases. For example, consider a STSRSWOR design with stratum sampling fraction $f_h = n_h / N_h$. Then $s = \cup_{h=1}^{H} s_h$, where s_h is a SRSWOR sample drawn from the h-th stratum, $h = 1, ..., H$. In this case, (5) becomes

$$\hat{V} = \sum_{h=1}^{H} N_h^2 \frac{1-f_h}{n_h} \sum_{s_h} \frac{(g_k e_k - \overline{ge}_h)^2}{n_h - 1}$$

where $\overline{ge}_h = \Sigma_{s_h} g_k e_k / n_h$. As another example, for SRSWOR, we have

$$\hat{V} = N^2 \frac{1-f}{n} \sum_s \frac{(g_k e_k)^2}{n-1} \qquad (6)$$

This is the case when the c_k in the model's variance structure, $\mathrm{Var}_\xi(\epsilon_k) = c_k \sigma^2$, satisfy $c_k = \lambda' x_k$ for all k and for some constant column vector λ. For example, for the homoscedastic variance structure, $\mathrm{Var}_\xi(\epsilon_k) = \sigma^2$ for all k, we have $\Sigma_s g_k e_k = 0$ and thus, (6) if the regression model contains an intercept term. A standard measure of precision used in survey organizations is the coefficient of variation, abbreviated as cv. For the GREG estimator (2), the cv is calculated as $\mathrm{cv} = \sqrt{\hat{V}} / \hat{Y}_{GREG}$.

Post-stratification is a special case of the GREG estimator. It is commonly used in large-scale surveys, mainly to increase the efficiency of the estimators on a conditional basis. However, comparisons based on the unconditional distribution suggest that the post-stratified estimator has a very slight advantage over simple $N\bar{y}_s$.

The traditional post-stratified estimator is derived from a model that is the special case of (1) such that $x_{pk} = 1$ for all $k \in U_p$. That is, the model is

$$y_k = \beta_p + \epsilon_k \text{ for } k \in U_p \qquad (7)$$

where $E_\xi(\epsilon_k) = 0$, $\mathrm{Var}_\xi(\epsilon_k) = \sigma_p^2$ and $\mathrm{Cov}_\xi(\epsilon_k, \epsilon_l) = 0$ for $k \neq l$. The model groups are called post-strata in this case. The required auxiliary information is the post-strata counts $N_p = \Sigma_{U_p} x_{pk}$ for $p = 1, ..., P$. The estimator (2) takes the form

$$\hat{Y}_{POST1} = \sum_{p=1}^{P} N_p \tilde{y}_{s_p}$$

where $\tilde{y}_{s_p} = \hat{Y}_{p\pi} / \hat{N}_p$ with $\hat{N}_p = \Sigma_{s_p} a_k$.

The variance estimator for $\hat{Y}_{POST1}$ is obtained from expression (5) by setting for $p = 1, ..., P$, $e_k = y_k - \tilde{y}_{s_p}$ for $k \in s_p$. The g-weights are $g_k = N_p / \hat{N}_p$ for all $k \in s_p$.

2.3 Calibration Approach

An alternative procedure to the regression approach, for accounting for auxiliary data is to find new weights w_k that are as close as possible to the original weights a_k. These new weights are subject to

the constraint $\Sigma_{s_p} w_k x_{pk} = X_p$ for $p = 1, ..., P$. We require the weights w_k to reproduce X_p group by group, in such a way that the weighted x-total over the sample gives the known group total X_p.

An advantage of this approach over the regression approach is that: (a) it permits to find new weights that are non-negative and bounded by a lower and an upper limit and (b) a wider class of estimators can be obtained.

A distance measure must be specified to quantify the distance between the new weights w_k and old weights a_k. Several possible distance measures are considered in Deville and Särndal (1992). Two commonly used distance functions are:

(a) The Generalized Least Squares (GLS) distance function

$$F(w_k/a_k) = (w_k/a_k - 1)^2/2,$$

(b) The Raking Ratio (RR) distance function

$$F\left(\frac{w_k}{a_k}\right) = \frac{w_k}{a_k}\log\left(\frac{w_k}{a_k}\right) - \frac{w_k}{a_k} + 1$$

The use of the GLS distance function leads to the generalized regression estimator $\hat{Y}_{GREG}$. Thus, the calibration approach is more general than the regression approach.

Computer software exists for this purpose. For example, the program CALMAR (Deville, Särndal and Sautory, 1993), solves the calibration equations by Newton's method and calculates the new weights. Other programs serving a similar purpose are M-WEIGHT by Huang and Fuller (1978) and BASCULA (Göttgens et al., 1991). The g-weights resulting from the output of these programs can easily be incorporated into GES.

Calibration theory can be applied to known marginals of a frequency table in any number of dimensions. A family of distance functions leads to *generalized raking ratio* estimators. When the RR distance function is used, we obtain the *raking ratio* or *iterative proportional fitting* estimators (Deming and Stephan, 1940; Brackstone and Rao, 1979).

2.4 Estimating Domain Totals

Domains are subpopulations for which point estimates of totals, means or other parameters are required, with the corresponding precision measures. Domains are not to be confused with model groups or with strata. These are also subpopulations but serve different purposes. Denote by $s_{(d)} = s \cap U_{(d)}$ the part of the sample s that falls in a domain $U_{(d)}$. Except in rare and controlled situations, such as when $U_{(d)}$ is identical to a stratum, the size of $s_{(d)}$ will be random.

The y-values observed within the domain are $\{y_k : k \in s_{(d)}\}$. Often, this information can be supplemented with auxiliary information to produce estimates with better precision. Here we consider estimation of the following kind. Suppose that x_k is an auxiliary vector whose total is known for specified model groups of the population U. We use the data $\{(y_k, x_k') : k \in s_{(d)}\}$ to estimate the domain total $Y_{(d)} = \Sigma_{U_{(d)}} y_k$. A standard device in domain estimation is to introduce a *domain variable*, denoted $y_{(d)}$, whose value for the k-th unit is

$$y_{(d)k} = \begin{cases} y_k & \text{if } k \in U_{(d)} \\ 0 & \text{if } k \notin U_{(d)} \end{cases} \qquad (8)$$

The domain total $Y_{(d)}$ can then be written as the total over the entire population U of the domain variable $y_{(d)}$. That is, $Y_{(d)} = \Sigma_U y_{(d)k}$.

To obtain the GREG estimator of $Y_{(d)}$, the following design-based procedure given in Estevao, Hidiroglou and Särndal (1992) can be used. A supply of g-weights g_k is first computed according to (4) for each model group, $p = 1, ..., P$. The weights $a_k g_k$ are then applied to observed the y_k-values in the domain. We obtain the estimator

$$\hat{Y}_{(d)GREG} = \sum_{p=1}^{P} \sum_{s_p} a_k g_k y_{(d)k} \qquad (9)$$

Note that the g-weights are functions of auxiliary totals at model group; this may be at a level coarser than the domain level.

We turn now to variance estimation. The variance estimator corresponding to the point estimator (9), $\hat{V}(\hat{Y}_{(d)GREG})$, is denoted $\hat{V}_{(d)}$ for short. It is calculated as

$$\hat{V}_{(d)} = \sum\sum_{s} \left(\frac{\Delta_{k\ell}}{\pi_{k\ell}}\right)\left(\frac{g_{ks} e_{(d)k}}{\pi_k}\right)\left(\frac{g_{\ell s} e_{(d)\ell}}{\pi_\ell}\right) \qquad (10)$$

where $e_{(d)k} = y_{(d)k} - x_{pk}' \hat{B}_{(d)p}$ for $k \in s_p$. Here $\hat{B}_{(d)p}$ is obtained from the normal equations

$$\sum_{s_p} \frac{a_k x_{pk} x_{pk}'}{c_k} \hat{B}_{(d)p} = \sum_{s_p} \frac{a_k x_{pk} y_{(d)k}}{c_k} \qquad (11)$$

assuming a regression model between $y_{(d)k}$ and x_p similar to (1).

Three different types of residuals enter into the

computation of (10). The first two types occur for sample units k belonging to intersecting model groups; the third type occurs for sample units k belonging to nonintersecting model groups. More specifically, for $k \in s_p = s \cap U_p$, we have

$$e_{(d)k} = \begin{cases} y_k - x'_{pk}\hat{B}_{(d)p} & \text{if } k \in U_{(d)}, U_{(d)} \cap U_p \neq \phi; \\ -x'_{pk}\hat{B}_{(d)p} & \text{if } k \notin U_{(d)}, U_{(d)} \cap U_p \neq \phi; \\ 0 & \text{if } U_{(d)} \cap U_p = \phi \end{cases}$$

The calculation of $\hat{V}_{(d)}$ is simplified since $e_{(d)k}$ is zero for all k in nonintersecting model groups.

In the special case where s is drawn by SRSWOR, then (10) becomes

$$\hat{V}_{(d)} = N^2 \frac{1-f}{n} \sum_s \frac{(g_k e_{(d)k})^2}{n-1}$$

when $\Sigma_s g_k e_{(d)k} = 0$, as is the case when the c_k in the model's variance structure satisfies $c_k = \lambda' x_k$ for all k.

The design-based coefficient of variation is computed in a manner completely analogous to (8), namely, $cv_{(d)} = \sqrt{\hat{V}_{(d)}} / \hat{Y}_{(d)GREG}$.

Several remarks are in order.

(a) Computational principle. The computations for a domain mimic the computations carried out for the entire population. To get the point estimator and the variance estimator for the domain $U_{(d)}$, just repeat the calculations made for the entire population, replacing y_k by $y_{(d)k}$ for $k \in s$. This implies that (3) turns into (9) for point estimation. For variance estimation, replacing y_k by $y_{(d)k}$ for $k \in s$, automatically implies replacing e_k by $e_{(d)k}$ and (7) will turn into (10). In other words, the computation of the domain estimator (9) and the corresponding variance estimator (10) is handled formally by replacing the y-variable by the domain variable $y_{(d)}$. Computational simplicity is thereby gained.

(b) Nature of the normal equations. The normal equations (11) correspond formally to the fit of the regression of the domain-specific dependent variable $y_{(d)}$ on the predictor x_p, using the sample observations from the p-th group. This fit may be mediocre because $y_{(d)}$ is not a natural dependent variable: it equals the y-variable inside the domain but is always equal to zero outside. But here we are not primarily interested in the goodness of the fit at the domain level. Instead the primary objective is to work with g-weights that (i) yield additive domain estimates (see Remark (d) below), and (ii) remain unchanged from one domain to another, which has computational advantages and allows calculation of other domain estimates than those officially reported by the organization. For alternative domain estimators, see Särndal, Swensson and Wretman (1992, pp. 408-413).

(c) Design consistency. The reason why the approach adopted here yields *close* estimates for domains hinges on the property of design consistency. It is known that $\hat{Y}_{GREG}$ given by (2) is a design consistent estimator of the entire population total Y. This implies loosely speaking that no matter what the configuration of finite population values $(y_1, ..., y_N)$, $\hat{Y}_{GREG}$ will be near Y with a high probability when the sample size is large. This property holds in particular for the domain-specific vector $(y_{(d)1}, ..., y_{(d)N})$. So, $\hat{Y}_{(d)GREG}$ given by (9) is a design consistent estimator of the domain total $Y_{(d)}$. Similarly for variance estimation, $\hat{V}$ given by (7) is a design consistent variance estimator. It follows that if the formula $\hat{V}$ is calculated on the domain-specific vector $(y_{(d)1}, ..., y_{(d)N})$, which gives the result $\hat{V}_{(d)}$ in (10), then we have a design consistent variance estimator for $\hat{Y}_{(d)GREG}$.

(d) Additivity property. Suppose that we seek to estimate the total for each of D domains $U_{(d)}$, $d = 1, ..., D$, forming a mutually exclusive and exhaustive partition of U. Then $\hat{Y}_{GREG} = \Sigma_{d=1}^{D} \hat{Y}_{(d)GREG}$ where $\hat{Y}_{GREG}$ and $\hat{Y}_{(d)GREG}$ are given by (3) and (9), respectively. This says that the sum of the domain estimates is equal to the estimate made for the entire population. This additivity property is built into the estimates because it is often required by users of official statistics. It follows easily that since $\Sigma_{d=1}^{D} y_{(d)k} = y_k$, for all $k \in U$.

3. GES Structure and Environment

The structure of the GES reflects the methodological components of generalized regression estimation. There are three main functions in the GES: (1) calculate sample design weights, (2) calculate g-weights and (3) calculate domain estimates. These must be carried out in that order. The user selects options and provides inputs

to each function. They can be changed within a given function and rerun the function and subsequent ones. This allows the user to experiment with different estimators.

To use the GES, a survey application must be first defined. Many of these can be defined but each survey application must be associated with a sample design. For each application, the user defines one or more time periods of survey data. This is useful for periodic surveys. But, at any given time, the estimation can be carried out with the available survey data in the specific period.

The GES has several user friendly features for browsing output files, modifying input files and selecting files and variables. Online Help is available for assistance in the use of the system.

The GES is being developed using the SAS System, version 6.08 for Windows 3.1. The SAS/AF product is used to create the GES menu windows. All source code is written in the SAS programming language using SAS data steps and procedures. The menu windows provide a simple user interface with menu bars, selection lists and point and click features.

The GES is a microcomputer application. To run GES, a 386(SX/DX) or 486(SX/DX) microprocessor is needed. A 486DX processor is recommended for optimum performance. In addition, the user must have Windows 3.1 (with DOS 5.0) running under Standard or Enhanced Mode and also SAS version 6.08 of the following products: BASE, AF, FSP and IML.

4. Future Development

We have desired that the GES be in a modular form. Each module produces an important component of the whole estimation process. Three most important ones are: (i) design weights calculation, (ii) g-weights calculation, and (iii) calculation of domain estimates (see Section 3). However, the current version is not as flexible as we intended. We are currently working to make the GES more flexible and modular so that the user can bring in his/her own design weights and/or g-weights into GES more easily.

Currently GES can accept one design: stratified simple random sampling of clusters (the same design for elements is acceptable by treating elements as clusters). Even with this simple design, GES can meet the estimation needs of many business surveys. However, most of social surveys use more complex designs such as stratified multi-stage probability proportional to size (pps) sampling design. We are planning to include this design in the future.

The GREG estimator described in Section 2 is general enough to include most traditional estimators. However, we are planning to incorporate a more general calibration estimator that contains the GREG estimator as a special case.

The variance estimator implemented in GES uses Taylor linear approximation for non-linear statistics. We intend to include the jackknife variance estimator in the GES as an option. The jackknife variance estimator is a general tool for variance estimation that can be used under very general condition.

Some surveys require synthetic type of estimator for small domains. Domain estimates obtained by (9) in Section 2 can be unreliable or even undefined if there is no sampled unit belonging to the domain of interest. In the latter case, however, sometimes auxiliary information is available for the units in the domain. For this and for general small domain (or area) estimation purpose, we are considering to include a synthetic type of estimator.

Almost all surveys use imputation for missing data. It is well known that the variance is underestimated when imputed data are treated as real observations. Several methods have been proposed to remedy this problem. These include multiple imputation of Rubin (1987), model assisted approach of Särndal (1990) and jackknife method of Rao (1992). Some empirical studies have been conducted to investigate the properties of these methods (see Lee, Rancourt and Särndal, 1991; Kovar and Chen, 1992). The Rao's jackknife method can be used with slight modification when the jackknife variance estimator is available. Since the GES follows the model assisted approach in variance estimation, the Särndal approach can be implemented in the GES. Thus, our future plan contains the implementation of these two methods.

The outlier problem is a familiar one in sample surveys. Outliers are influential observations to the particular estimator employed. They can be detected and treated to reduce their influence in estimation or we can use a robust estimator. Robust estimation option is another item we are considering. Outlier methodologies in sample surveys have been reviewed in Lee (1993).

Finally, we are also considering the inclusion of estimators of the population distribution function and quantiles. They are particularly needed when the income distribution is investigated.

5. Concluding Remarks

In this paper we have presented the principles behind the development of the Generalized

Estimation System (GES) at Statistics Canada. An important aspect of the GES is the use of auxiliary information to improve efficiency.

The currently programmed specifications can handle stratified, single-stage sample designs such as stratified simple random sampling without replacement, stratified cluster sampling, and stratified probability-proportional-to-size (PPS) sampling. The system has been programmed to produce Hajek, ratio, simple regression, post-stratified and generalized regression estimators for stratified simple random samples of elements or (single stage) clusters. It is being extended to stratified probability proportional to size sampling, with and without replacement, for elements and clusters and will eventually handle multi-stage designs.

6. References

Bethlehem, J.G., and Keller, W.J. (1987). Linear Weighting of Sample Survey Data. *Journal of Official Statistics*, 3, 141-153.

Brackstone, G.J., and Rao, J.N.K. (1979). An Investigation of Raking Ratio Estimation. *Sankhyā*, Series C, 41, 97-114.

Deming,W.E. and Stephan, F.F. (1940) . On a Least Squares Adjustment of a Sampled Frequency Table When the Expected Marginal Totals Are Known. *Annals of Mathematical Statistics*, 11, 427-444.

Deville, J.-C., and Särndal, C.E. (1992). Calibration Estimators in Survey Sampling. *Journal of the American Statistical Association*, 87, 376-382.

Deville, J.-C., Särndal, C.E., and Sautory, O. (1993). Generalized Raking Procedures in Survey Sampling. *Journal of the American Statistical Association*, 87, 376-382.

Estevao, V., Hidiroglou, M.A., and Särndal, C.E. (1992). Requirements on a Generalized Estimation System at Statistics Canada. Presented at the Workshop on Uses of Auxiliary Information, Statistics Sweden, October 5-7, 1992.

Göttgens, R., Vellen, B., Odekerken, M., and Hofman, L. (1991). Bascula, Version 1.0. A Weighting Package under MS-DOS, User Manual. CBS-Report, Netherlands Central Bureau of Statistics, Voorburg, The Netherlands.

Huang, E. and Fuller, W.A. (1978). Nonnegative Regression Estimation for Sample Survey Data. *Proceedings of the Social Statistics Section, American Statistical Association*, 330-305.

Kovar, J., and Chen, E. (1992). Variance under Imputation: An Empirical Investigation. Presented at the 1992 Annual Meeting of the Statistical Society of Canada. Edmonton, Alberta, May 31 - June 2.

Lee, H. (1993). Outliers in Survey Sampling. A chapter in the monograph for the International Conference on Establishment Surveys, Buffalo, June 27-30. (To appear)

Lee, H., Rancourt, E., and Särndal, C.E. (1991). Experiments with Variance Estimation from Survey Data with Imputed Values. *Proceedings of the Section on Survey Research Methods*, American Statistical Association, pp. 690-695.

Outrata, E, and Chinnappa, N. (1989). General Survey Functions Design at Statistics Canada. Unpublished technical paper, Statistics Canada.

Rao, J.N.K. (1992). Jackknife Variance Estimation under Imputation for Missing Data. Unpublished paper, Statistics Canada.

Rojas, G., and Aliaga, A. (1993). Sampling Errors in the Integrated System for Survey Analysis (ISSA). *Proceedings of the Section on Survey Research Methods*, American Statistical Association (to appear).

Rubin D.B. (1987). *Multiple Imputation for Nonresponse in Surveys*. New York: John Wiley.

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

Schnell, D., Kennedy, W.J., Sullivan, G., Park, J.P., and Fuller, W.A. (1989). Personal Computer Variance Software for Complex Surveys. *Survey Methodology*, 14, 59-69.

Särndal, C.E. (1990). Methods for Estimating the Precision of Survey Estimates When Imputation Has Been Used. *Proceedings of Statistics Canada Symposium '90: Measurement and Improvement of Data Quality*, pp. 369-380.

Särndal, C.E., Swensson, B. and Wretman, J.H. (1989). The Weighted Residual Technique for Estimating the Variance of the General Regression Estimator of the Finite Population Total. *Biometrika*, 76, 527-537.

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