CLAN – A SAS-PROGRAM FOR COMPUTATION OF POINT- AND STANDARD ERROR ESTIMATES IN SAMPLE SURVEYS

Claes Andersson, Statistics Sweden
Lennart Nordberg, Statistics Sweden
Claes Andersson, Statistics Sweden, AM/LEDN, S-701 89 Örebro, Sweden, claes.andersson@scb.se

ABSTRACT

CLAN is a program developed at Statistics Sweden and designed to compute point- and standard error estimates in sample surveys. All parameters that can be written as arbitrary rational functions of population (domain) totals can be handled. For example simple ratios, ratios between ratios, differences between different domain totals, etc. CLAN computes an estimate of the parameter and its standard error using a Taylor linearization approach. The Horvitz-Thompson estimator and/or the generalized regression (GREG) estimator can be used for simple totals. CLAN can handle all parameters that can be written as rational functions of totals where the user only has to specify the functional form of the parameters in terms of population or domain totals. The theory and its implementation are described.

Key Words: Linearization, Domain estimation, Complex parameters, Statistical software, Auxiliary information.

1. INTRODUCTION

In this paper we will present the main features of CLAN, which is a program designed to compute point and standard error of estimates in sample surveys. CLAN is written in the SAS macro language and based on the Taylor linearization method for the variance estimation. A method is implemented where the user only has to specify the functional form of the parameters in terms of population or domain totals. Expressions for the necessary derivatives are generated automatically.

All parameters that can be written as rational functions of totals can be handled by this technique. Typical examples in business statistics are production per head within different industries and production of an industry relative to the whole population. Measurement of change often leads to complex estimation problems, the relative change in production by industry from one period to another is often measured as a ratio where the numerator and denominator have different reference times. The sampling units involved in the numerator and denominator are partly the same, partly different, and units that contribute to the total on both occasions may have changed industry in between. Indices of production built up by functions of ratios, etc. are other examples.

Usually, exact expressions for the sampling variances of non-linear estimators are not available, neither are simple unbiased estimators of the variances.

The Taylor linearization is computationally much less intensive than for example the Jackknife but it requires that new expressions are worked out for each different parameter that is considered. Resampling plans, in general only require that the functional form of the parameter is specified.

The Taylor method has been used in survey sampling for a long time, examples of its applicability are given by Tepping (1968), Woodruff (1971) and Woodruff and Causey (1976). The technique has also been used in general computer programs for the estimation of ratios, regression coefficients, etc. and their variances. Examples of such computer programs are SUPER-CARP (Hidiroglou et al 1976), SUDAAN (Research Triangle Institute 1989), PC-CARP (Schnell et al 1988), and GES (Estevao et al 1995).

Section 2 contains an overview of CLAN including the sampling designs and non-response models that can be used, in section 3 details on the available estimators are given. In section 4 the idea behind the technique is shown. Finally in section 5 the pros and cons of the Taylor linearization and the Jackknife is discussed and in section 6 new developments are mentioned.

2. CLAN – A TOOL-KIT FOR VARIANCE ESTIMATION

The SAS program CLAN is not an ordinary program of the type we are used to see today with nice screen layouts, buttons, boxes with available alternatives etc. It is more like a tool-kit for point- and standard error estimation of complex parameters in survey sampling. The intended user is a professional statistician. The user needs some elementary knowledge in computer programming to write a SAS-macro (a program fragment) where the form of the parameters of interest is expressed in terms of population totals. CLAN provides the user with a tool-kit, which makes this task fairly easy.

Being written in the SAS macro language CLAN works in different computer environments, e.g. under Windows 9x, UNIX and on mainframe computers. Only the Base SAS software is necessary.

CLAN permits the user to choose between a large number of estimators, including estimators that use auxiliary information. The major strength of the program lies in the flexibility by which the user may combine estimators with the specification of complex sets of domains.

2.1. What Parameters can be Estimated?

Population and domain totals are the most basic and probably the most common parameters of interest in surveys. Other examples of frequently occurring parameters are population means, ratios or differences between population and/or domain totals. A difference between two ratios under a panel survey design, where the two ratios refer to different time periods is a further example.

In summary, most parameters of interest in surveys may be expressed as functions of population (or domain) totals. Consider the parameter θ of a fixed finite population U of size N. Let θ be a function of J totals $\mathbf{t}=(t_1, ..., t_j, ..., t_J)'$, that is

$$\theta=f(t_1, ..., t_j, ..., t_J)=f(\mathbf{t}) \tag{2.1}$$

where $t_j = \sum_U y_{jk}$ is the total of the variable y_j in population U and y_{jk} is the value of y_j for unit k.

All parameters θ that can be written in the form (2.1), where f is an arbitrary rational function, can be handled (meaning that a point estimate and a corresponding standard error estimate is produced), by CLAN. In fact, the computational technique which is applied (see section 4 ahead) can easily be extended to include other functions such as *log, exp, sqrt etc.* However, since there has not yet been a demand for such functions, only rational functions are currently allowed in CLAN.

For some important survey parameters such as the median and other fractiles the function f is defined implicitly via an estimating function. Unfortunately such implicitly defined functions are not covered by the computational approach used for CLAN. Hence variances for fractile estimators can not be computed by CLAN.

The y-variables mentioned above can be defined in the following way. Consider a set of variables $y_1, ..., y_j, ..., y_J$ and let y_{pk} be the value of variable y_p for unit k in the finite population U. Also consider a partitioning of U into D possibly overlapping domains $U_1, ..., U_d, ..., U_D$. For each one of the $P\times D$ possible combinations of variables and domains we can define a total $t_{pd} = \sum_{k\in U_d} y_{pk} = \sum_{k\in U} y_{pk} c_{dk}$ where c_{dk} is an indicator variable such that

$$c_{dk} = \begin{cases} 1 \text{ if unit } k \in U_d \\ 0 \text{ otherwise.} \end{cases} \tag{2.2}$$

In general only a subset of the $P\times D$ possible totals are of interest. Let $\mathbf{t}=(t_1, ..., t_j, ..., t_J)'$ be the vector of the J totals we are interested in. To every t_j in $\mathbf{t}$ we associate one variable y_p and one domain U_d. Formally $t_j = \sum_{k\in U_{d(j)}} y_{p(j)k}$, where $y_{p(j)}$ and $U_{d(j)}$ mean that y_p and U_d are associated with the total t_j. Notice that several t_j may be associated with the same y_p or U_d.

Since each t_j is associated with only one variable and one domain we will henceforth suppress the indices p and d and write y_{jk} instead of $y_{p(j)k}$ and U_j instead of $U_{d(j)}$ in order to simplify notation.

The vector $\mathbf{t}$ of simple totals is generated by J arbitrary combinations of y-variables and domains. At one extreme $\mathbf{t}$ might be a vector of totals of *one variable* y (i.e. P=1), in J (=D) different, possibly overlapping domains $U_1, U_2, \ldots, U_J$. At the other extreme $\mathbf{t}$ might be a vector of totals in *one domain* (i.e. D=1) for J (=P) different y-variables.

There are two families of estimators available in CLAN for the estimation of the population totals $(t_1, ..., t_j, ..., t_J)'$. The basic one is the Horvitz-Thompson (HT) estimator (see section 3.1 ahead). The second family of estimators of $\mathbf{t}$ implemented in CLAN is the calibration and generalized regression (GREG-) estimation, see section 3.2.

As an estimator of θ, CLAN uses $\hat{\theta} = f(\hat{\mathbf{t}})$. CLAN computes $\hat{\theta}$ and an estimate – based on Taylor linearization – of the standard error $\sqrt{V(\hat{\theta})}$. This estimation can be done, simultaneously, for an arbitrary number of functions $\theta_q=f_q(t_1, ..., t_j, ..., t_J)$, q=1, 2, …, Q.

2.2. Sample Designs

The sampling designs implemented in CLAN include stratified (or non-stratified) simple random sampling (SRS) without replacement of *i*) elements or *ii*) cluster of elements. When the sampling unit is cluster no subsampling of elements is assumed. This means for example that probability-proportional-to-size (PPS) and multistage cluster sampling designs are missing. The reason is that such designs are currently used very little at Statistics Sweden. Nevertheless, under certain conditions it is possible to use CLAN for the computation of *approximate* estimates of the variances under PPS or multistage cluster designs by appropriately weighting.

2.3. Non-response Models

In most surveys non-response occurs. There are two main ways to treat this, by weighting or by imputation, CLAN can be used for weighting but does not take imputation into account.

We can look upon the non-response compensation adjustment weight as an approximation of the inverse of the response probability for the sample units. Särndal *et al.* (1992) suggested that the response probability is estimated by grouping the sample s into response homogeneity groups (RHGs) and compute the ratio between the number of responding and selected units within each RHG.

CLAN allows for two models, *i*) strata and RHGs coincide, or *ii*) RHGs are subgroups of strata. The second alternative implies that CLAN can also be used when the sample design means a two-phase sampling scheme for stratification, see Särndal et al (1992) for details.

Assume that stratum h contains N_h sample units and a sample of size n_h has been taken from stratum h, then for element k in stratum h, $\pi_k=n_h/N_h$. When the sampling unit is cluster no subsampling of elements from the selected clusters is assumed which means that $\pi_k=\pi_{1i}$ if element k belongs to cluster i.

Suppose that the number of respondents in stratum h is m_h out of n_h, when model *i*) is used the response probability is estimated by $\hat{\nu}_k = m_h/n_h$. When model *ii*) is used the sample units in stratum h is divided into G_h subgroups where subgroup hg contains n_{hg} selected units and m_{hg} responding units, then the response probability for element k in subgroup hg is estimated by $\hat{\nu}_k = m_{hg}/n_{hg}$. When the sampling unit is cluster then $\hat{\nu}_k = \hat{\nu}_{1i}$ if element k belongs to cluster i.

2.4. How Is It Done?

The pre-programmed macro %CLAN is the main tool where the user specifies *how* the estimation shall be done. It involves a number of macro parameters for specification of the design and response model. Only a few of these are mandatory and the rest are optional. The user supplies information about which data set to use (DATA), if the sampling units (SAMPUNIT) are elements (default) or clusters, if response homogeneity groups (RHG) are used or not (default), stratification variable (STRATID) if any, the number of units in the sampling frame (NPOP), the number of responding units (NRESP), etc.

In many cases it is practical to be able to group the obtained estimates by the levels of one or two factors. For this purpose the output data set always contains two SAS variables ROW and COL which can be used to structure the output and to identify the computed estimates. The user specifies the number of levels of ROW and COL by setting values to the parameters MAXROW and MAXCOL in %CLAN.

Here is an example of what %CLAN may look like if we want to produce a table (or a number of tables) with, say 3 rows and 4 columns:

```
%CLAN(DATA=s.indata, SAMPUNIT=C, STRATID=stratvar, CLUSTID=clustvar, NPOP=popsize,
    NRESP=mresp, MAXROW=3, MAXCOL=4)
```

It is presumed that the input SAS data set S.INDATA contains the variables STRATVAR, CLUSTVAR, POPSIZE, MRESP and CLUSTVAR, which contains the cluster identity for each element.

The form(s) of the function(s) $\theta=f(t_1, ..., t_j, ..., t_J)$ and the specification of the totals involved, $t_1, ..., t_j, ..., t_J$ must be supplied by the user. To do this the user must write a SAS macro %FUNCTION to specify the elementary estimators of the totals involved, and for which functions of these totals the user wants to compute point and standard error estimates. In %FUNCTION the user specifies *what* shall be estimated. CLAN provides a tool-kit of pre-programmed macros to help the user build %MACRO FUNCTION.

The totals $t_1, ..., t_j, ..., t_J$ in θ are defined and estimated by the pre-programmed macros %TOT and/or %GREG which use a linear estimator of t_j in the form $\hat{t}_j = \sum_s w_k y_{jk}$, where w_k will be defined later. For each %TOT or %GREG that is invoked in %FUNCTION, a table of size MAXROW×MAXCOL with estimates of simple totals is internally set up by CLAN.

To understand how this works, it may be helpful to know that %FUNCTION is internally invoked from %CLAN within a do-loop for every observation read from the input data set. Hence inside %CLAN there is a sequence (which the user does not have to be concerned with in practice):

```
do _i=1 to &maxrow;
do _j=1 to &maxcol;
    %function(_i, _j)
end;
end;
```

The SAS macro language uses '&X' to refer to a *value* of variable X.

As implied by the do-loop, the arguments of %FUNCTION – the arguments are formal parameters, the user can choose any valid SAS names – are associated with the variables ROW and COL that will appear in the output data set, i.e. the first argument takes the values 1, 2,…, MAXROW while the second takes the values 1, 2, …, MAXCOL. The user can control which parameters and/or domains that are to be associated with each combination of ROW, COL by referring to *arg*1 and *arg*2 in macro %FUNCTION(*arg*1*, arg*2).

Since $\theta=f(\mathbf{t})$ must be a rational function of totals, θ can be obtained by using the elementary operations *addition, subtraction, multiplication, division* in a step-wise fashion. There are four pre-programmed "arithmetic" macros in CLAN corresponding to these operations: %ADD, %SUB, %MULT and %DIV. The user can construct any rational function or sets of rational functions of totals by using these elementary macros as building blocks in much the same way as one would use the corresponding operations in elementary algebra. The derivatives needed and the linear transformations of the variable values are automatically provided by these macros.

The "arithmetic" macros operate on cells with identical index in different tables created by %TOT, %GREG or the four "arithmetic" macros. Usually only the final transformation is of interest, the macro %ESTIM is used to tell CLAN which parameter(s) to output and compute of standard errors. In order to sum over cells within the same table the macro %TABSUM is used. We will illustrate this in a small example.

2.5. An Example of %FUNCTION

The objective of this example is to demonstrate how CLAN can be used to estimate a set of partly overlapping domain totals by an estimator which takes the form of a sum of a ratio estimator using an auxiliary variable known in a part of the population and a HT estimator to cover the rest. This compounded estimator is used – slightly modified – at Statistics Sweden to estimate quarterly turnover by industry for the service sector.

The design of the survey is a stratified sample of enterprises with simple random sampling within strata. The stratification is done by size of enterprise and by type of service – 3 to 4 digit levels of the European standard industrial classification NACE.

The quarterly turnover y_k, $k\in s$ is measured for every enterprise in the sample. The auxiliary variable is the yearly turnover of the previous year. This variable x_k, is known for every "old" enterprise in the population, $k\in U_{old}$ but not for the new ones, U_{new}. The sample s is split in the same way in s_{old} and s_{new}.

The objective is to estimate the quarterly turnover for a number of domains d, d=1, 2, …, D. In this example we will use part of the wholesale trade industry as illustration. We will consider the following partly overlapping NACE groups. The overlap stems from the fact that some domains are aggregates of others.

Row	*NACE code*	*Type of activity*
1	51.41+42	Wholesale of textiles, clothing and footwear.
2	51.43+44+47	Wholesale of electrical appliances, glass, china, furniture, leisure goods etc.
3	51.45+46	Wholesale of medicine and cosmetics.
4	51.4	Wholesale of household goods.
5	51.5+6+7	Wholesale of non-agricultural intermediate goods, machinery and other wholesale.
6	51	Wholesale and commission trade except of motor vehicles.

Notice that row 4 is an aggregate of rows 1, 2 and 3. Furthermore, row 6 is an aggregate of rows 1, 2, 3 and 5. We are interested in the quarterly turnover in domains that are defined by the cross classification of NACE code and size classes, 10-49, 50-249, 250-. The margins are also of interest.

The estimator of the quarterly turnover $t_d = \sum_{U_d} y_k$ in domain d can be expressed in the following form, $\hat{t}_d = \hat{t}_{dnew} + \hat{t}_{dold}$, where $\hat{t}_{dnew} = \sum_{s_{new}} w_k y_{dk}$, $\hat{t}_{dold} = \frac{\sum_{U_{old}} x_{dk}}{\sum_{s_{old}} w_k x_{dk}} \sum_{s_{old}} w_k y_{dk}$ and $y_{dk}=c_{dk}y_k$.

For each enterprise in the input data set, INDATA the following variables exist, STRATUM, POP (N_h), RESP (n_h), Y (quarterly turnover), X (last year's turnover if it exists), XSUM ($\sum_{U_{old}} x_{dk}$), NEW (=1 if the enterprise is newly started, =0 otherwise), NACE4 (4 digit NACE), NACE3 (3 digit NACE) SIZE (size class, 1=10-49, 2=50-249, 3=250-). The %FUNCTION may then look like this.

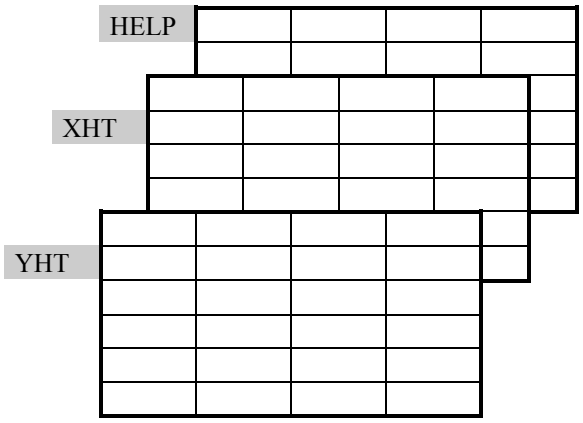
```
%macro function(r,c);
     rcond=0;
     ccond=(&c=size);
     if &r=1 then rcond=(nace4 in('5141','5142'));
     if &r=2 then rcond=(nace4 in('5143','5144','5147'));
     if &r=3 then rcond=(nace4 in('5145','5146'));
     if &r=5 then rcond=(nace3 in('515','516','517'));
     %tot(yht, y, ccond and rcond and new)
     %tot(xht, x, ccond and rcond and not new)
     %tot(help, y*xsum, ccond and rcond and not new)
     %div(yrat, help, xht)
     %add(yest, yht, yrat)
     %tabsum(table=yest,frow=1-3,trow=4)
     %tabsum(table=yest,frow=4 5,trow=6)
     %tabsum(table=yest,fcol=1-3,tcol=4)
     %estim(yest)
%mend;
%clan(data=indata, stratid=stratum, npop=pop, nresp=resp, maxrow=6, maxcol=4)
run;
```

The output is a SAS data set named 'DUT' (Data oUT) that contains 6×4=24 rows in this case and four variables, ROW, COL, PYEST (for the point estimates) and SYEST for the standard error estimates. There may be several %ESTIMs in a %FUNCTION and each will generate two variables with prefix P (for Point-estimate) and S (for Standard error estimate). For example, if the estimated totals for new enterprises had been of interest then by adding %ESTIM(YHT) after the last %ESTIM in %FUNCTION we had obtained two more variables, PYHT and SYHT on the output data set DUT.

The data set may be processed by other procedures in SAS or exported to another program, for example EXCEL for further processing or to create a table in a document.

It is usually a good idea to do some data manipulation such as grouping, recoding etc. in a DATA step before running %CLAN, especially for large tables and large data sets.

In this example CLAN will produce three internal 6×4 tables in the first pass of the data - as illustrated by the following figure - since %TOT is invoked three times (for YHT, XHT and HELP).



In the second pass of the data the transformation defined by %DIV, %ADD and %TABSUM will be done. However only YEST will be used in the computation of the standard error and written to the data set DUT. Note that %DIV will compute HELP/XHT, cell by cell and %ADD will compute YHT+YRAT, cell by cell, while %TABSUM will compute the sum of rows and columns in YEST. The linear transformations needed for the variance estimation will be done according to these operations.

Usually the tricky part in using CLAN is the formulation of the logical conditions in %TOT (and %GREG) that ties each element to the intended cell(s) but this is also the key, we believe, to the high flexibility of the software.

3. THE ESTIMATION OF SIMPLE TOTALS

The simple total $t_j = \sum_U y_{jk}$ is estimated by the estimator $\hat{t}_j = \sum_r w_k y_{jk}$ where r is the set of response elements and w_k is the "sampling weight". By choosing w_k in different ways different estimators are defined. Two estimators for the simple totals, the HT estimator and the GREG estimator are implemented in CLAN.

3.1. The HT estimator of a Simple Total

The well known HT estimator is obtained by using $w_k = 1/(\hat{\nu}_k \pi_k)$. The HT estimator is computed by the pre-programmed SAS macro %TOT(*name*, *variable*, *condition*), where *name* is the users name of the estimate, *variable* is the name of the SAS variable or a SAS expression used as the *y*-value and *condition* is a logical expression that determines to which cell(s), an element belongs.

3.2. The GREG estimator of a Simple Total

The GREG estimator of t_j is obtained by using $w_k = g_k/(\hat{\nu}_k \pi_k)$ in $\hat{t}_j = \sum_r w_k y_{jk}$, where g_k is a function of the auxiliary vector $\mathbf{x}_k$.

It is assumed that the y_j-values are generated by a linear model ξ such that, $E_\xi(y_{jk}) = \mathbf{x}_k' \mathbf{B}_j$ and $V_\xi(y_{jk}) = \sigma^2/q_k$ where q_k is a constant known for each element.

When the sampling units are clusters, two model levels are possible. The auxiliary information may be available for the population of clusters or for the population of elements. Then the model may be specified as an element model ξ or as a cluster model ξ_I, $E_{\xi_I}(t_{yji}) = \mathbf{t}_{xi}' \mathbf{B}_{Ij}$, $V_{\xi_I}(t_{yji}) = \sigma_I^2/q_{Ii}$ where q_{Ii} is a constant known for each cluster, $t_{yji} = \sum_{U_i} y_{jk}$ is the total of y_j in cluster i and $\mathbf{t}_{xi}$ is either defined at the cluster level only or an aggregate of $\mathbf{x}_k$ within each cluster, $\mathbf{t}_{xi} = \sum_{U_i} \mathbf{x}_k$. The unknown constants σ^2 and σ_I^2 disappear in the estimation of $\mathbf{B}_j$ and $\mathbf{B}_{Ij}$.

The g-values are computed by $g_k = 1 + (\mathbf{t}_x - \hat{\mathbf{t}}_x)' \left(\sum_r \mathbf{x}_k \mathbf{x}_k' q_k/(\hat{\nu}_k \pi_k)\right)^{-1} \mathbf{x}_k q_k$ under the model ξ, where $\hat{\mathbf{t}}_x$ is the HT estimator of the known total $\mathbf{t}_x$ in U.

Under the model ξ_I, $g_k = 1 + (\mathbf{t}_x - \hat{\mathbf{t}}_x)' \left(\sum_{r_I} \mathbf{t}_{xi} \mathbf{t}_{xi}' q_{Ii}/(\hat{\nu}_{Ii} \pi_{Ii})\right)^{-1} \mathbf{t}_{xi} q_{Ii}$ $k \in U_i$ and r_I is the set of responding clusters.

The GREG-estimator encompasses as special cases the ratio estimator, the simple regression estimator, the post-stratified estimator and the linear calibration estimator (Deville and Särndal 1992). These cases are obtained by an appropriate choice of the **x**-vector.

The g_k:s should normally have a value around 1. However, for some "unlucky" samples or if there are too many marginal constraints, the g-values may take extreme values. Negative values of g_k is rather undesirable in practice, in fact a value of w_k less than one is hard to explain or interpret since each observation should at least represent itself in the estimation process.

The user may want to restrict the g_k-values within a certain interval. This can be done in CLAN by specifying an upper limit and a lower limit for g_k, such that $L \le g_k \le U$.

The macro %AUXVAR can also compute (linear) calibration weights w_k, such that $\hat{\mathbf{t}}_x = \sum_r w_k \mathbf{x}_k = \mathbf{t}_x$.

Up to nine different %AUXVARs are allowed in %FUNCTION at the same time, which means that nine different models can be treated in the same run. This may for example be used when GREG estimates from different time periods are combined into the estimates of parameters that measures change.

The GREG estimates are computed by the pre-programmed macro %GREG(*name, variable, condition, modelid*), where *name*, *variable* and *condition* has the same meaning as in %TOT, and *modelid* is used to connect %GREG with one of the **x**-vector when several %AUXVARs are used.

The working variable in the estimation of a total or in the z-transformations is $g_k(y_{jk} - \mathbf{x}_k' \hat{\mathbf{B}}_j)$ which itself is a result of Taylor linearization.

3.2.1. The Auxiliary Vector

The **x**-vector of auxiliary variables is defined in much the same way as the *y*-variables, i.e. by a combination of variable(s) and domain indicator(s). Usually the term model group or subpopulation is used instead of domain.

Simple examples of auxiliary vectors are $\mathbf{x}_k = (0, \ldots, 1, \ldots, 0)'$ and $\mathbf{x}_k = (0, \ldots, x_k, \ldots, 0)'$. When $\mathbf{x}_k = (0, \ldots, 1, \ldots, 0)'$ the auxiliary total $\mathbf{t}_x = \sum_U \mathbf{x}_k = (N_1, \ldots, N_j, \ldots, N_J)'$ must be known, where N_j is the population count in the jth subpopulation. In this case, if the sample is taken as an SRS, the corresponding GREG-estimator is the (classical) poststratified estimator. In the case of $\mathbf{x}_k = (0, \ldots, x_k, \ldots, 0)'$ the auxiliary total $\mathbf{t}_x = \sum_U \mathbf{x}_k$ must be known.

A general expression for the $\mathbf{x}_k$ vector is given by, $\mathbf{x}_k = (\delta_{1k}\, \mathbf{x}'_{1k}, \ldots, \delta_{pk}\, \mathbf{x}'_{pk}, \ldots, \delta_{Pk}\, \mathbf{x}'_{Pk})'$ where δ_{pk} is a subpopulation indicator that takes the value $\delta_{pk}=1$ if k belongs to subpopulation p and $\delta_{pk}=0$ otherwise.

In principle the auxiliary vector $\boldsymbol{x}_p$ can be composed differently for different subpopulations. However, the current version of CLAN does not allow for this, $\mathbf{x}_{pk}$ must be identically composed for each p.

The model, the **x**-vector, the known margins $\mathbf{t}_x$, q_k or q_{1i} etc. are defined by the macro %AUXVAR that allows for a variety of different models and **x**-vectors.

3.2.2. An Example

The estimation problem in section 2.5 could also be solved by using of the macros %AUXVAR and %GREG. The composition of the **x**-vector is defined in a separate SAS data set, TOTALS. In this example it contains one observation (or row) and the variables VAR (the name of the x-variable, ='X'), N (number of subpopulations, =12, 4 NACEs by 3 sizes), MAR1-MAR12 (the known totals, =the values of XSUM), XTYPE (type of variable, ='N' for Numerical), XDOM (subpopulation indicator, ='NACESIZE'). The variables X and NACESIZE are assumed to exist in the data set INDATA. The variable NACESIZE takes the values 1-12 for all old enterprises and 0 for new, it is defined by the user through a combination of NEW, NACE3, NACE4 and SIZE. In INDATA there is also a variable, Q that takes the value 1/X for old enterprises and 0 for new ones.

Then, instead of the three %TOTs, the %DIV and the %ADD – the other statements in %FUNCTION are not changed – we could use the following code.

```
.
%auxvar(datax=totals,qk=q)
%greg(yest,y,rcond and ccond)
.
```

The result would be exactly the same as from the code in section 2.5.

4. THE LINEARIZATION VARIANCE ESTIMATOR

CLAN allows arbitrary rational functions of totals, $\theta = f(\mathbf{t})$. We have already mentioned a number such functions that are non-linear in the totals. A natural estimator of θ is to replace the different totals in $f(\mathbf{t})$ by their estimates $\hat{\mathbf{t}} = (\hat{t}_1, \ldots, \hat{t}_j, \ldots, \hat{t}_J)'$, which gives $\hat{\theta} = f(\hat{t}_1, \ldots, \hat{t}_j, \ldots, \hat{t}_J)$.

Notice that $\hat{\theta}$ is generally not an unbiased estimator of θ when $f(\mathbf{t})$ is non-linear, not even when $\hat{\mathbf{t}}$ is an unbiased estimator of $\mathbf{t}$. It is, however, a consistent estimator of θ if $\hat{\mathbf{t}}$ is a consistent estimator of $\mathbf{t}$.

By Taylor's formula we have as an approximation

$$V(\hat{\theta}) \approx V\left(\sum_{j=1}^{J} f'_j(\mathbf{t})\hat{t}_j\right) = \sum_{i=1}^{J}\sum_{j=1}^{J} f'_i(\mathbf{t}) f'_j(\mathbf{t}) C(\hat{t}_i, \hat{t}_j), \tag{4.1}$$

where $f'_j(\mathbf{t})$ is the partial derivatives are evaluated at $\mathbf{t}$ and $C(\hat{t}_i, \hat{t}_j)$ is the covariance between $\hat{t}_i$ and $\hat{t}_j$.

The partial derivatives f' usually depend on the unknown $\mathbf{t}$. For the purpose of variance estimation, we substitute sample based estimates of f'_j and $C(\cdot, \cdot)$. The f'_j is estimated by $f'_j(\hat{\mathbf{t}})$ i.e. the partial derivatives are evaluated at $\hat{\mathbf{t}}$ and $C(\cdot, \cdot)$ is estimated by $\hat{C}(\cdot, \cdot)$ computed from the sample. The result is the estimator,

$$\hat{V}(\hat{\theta}) = \sum_{i=1}^{J}\sum_{j=1}^{J} f'_i(\hat{\mathbf{t}}) f'_j(\hat{\mathbf{t}}) \hat{C}(\hat{t}_i, \hat{t}_j) \tag{4.2}$$

It can be shown by elementary algebra that if we do the following data transformation – due to Woodruff (1971) – for every element in the sample,

$$\hat{z}_k = \sum_{j=1}^{J} f'_j(\hat{\mathbf{t}}) y_{jk} \tag{4.3}$$

and then derive the variance estimator of the total $\hat{t}_{\hat{z}} = \sum_s w_k \hat{z}_k$, which is well known for most designs used in practice, we arrive at the expression (4.2). Hence the computation of variances for *functions of totals* is converted

into computation of variances of *totals* through the Woodruff transformation (4.3). The problem of finding the partial derivatives of all functions allowed is solved by the following observation.

Suppose that $f(\mathbf{t})$ can be written in the form, $f(\mathbf{t}) = G\big(g_1(\mathbf{t}), g_2(\mathbf{t})\big)$, and this is true for all functions allowed in CLAN, i.e. all rational functions.

It is then possible to do the Woodruff transformation (4.3) separately for each of the functions g_1 and g_2, and then use the resulting z-variables as input to the Woodruff transformation for the function G. Thus it is possible to obtain the transformation (4.3) in a stepwise fashion. For a more comprehensive discussion of this, the reader is referred to Andersson and Nordberg (1994).

Hence the Woodruff transformation for a rational function f can be obtained by successive use of the Woodruff transformations corresponding to addition, subtraction, multiplication or division of *two* totals or functions of totals.

4.1. An Algorithm

Since it is possible to compute the derivatives of $\hat{\theta}$ in a stepwise manner, it is not difficult to construct an algorithm that is suitable for a computer program where the user does not have to be concerned with the derivatives.

In the algorithm we only have to worry about the derivatives of functions like t_1 *op* t_2, where *op* is one of the operators +, - , × and /. Although it is rather simple to include other functions, i.e., θ=*func*(t_0), where *func* is, for example, *log*(t_0), *exp*(t_0), *sqrt*(t_0), etc., we do not treat these functions here simply because there has (not yet) been any need for them.

The following table shows the z-transformations needed to estimate the variance of $\theta = t_1$ *op* t_2.

op	*z-transformation*
+	$z_k = y_{1k} + y_{2k}$
-	$z_k = y_{1k} - y_{2k}$
×	$z_k = \theta\,(y_{1k}/t_1 + y_{2k}/t_2)$
/	$z_k = \theta\,(y_{1k}/t_1 - y_{2k}/t_2)$

Table 4.1. *z-transformations for different operators*.

In CLAN a two-pass algorithm is used in order to find the derivatives and compute the variance estimates of an arbitrary rational function of totals. In the first pass all the simple totals needed are estimated by the HT or GREG estimator.

When all totals of interest are estimated, the second pass begins. The z-transformations are then done step by step by using the expression of θ_q which is supplied by the user. The final transformation(s), determined by %ESTIM, is used as input to the standard formula for the computation of the variance estimate of a total. The formula is determined by the users choice of design and non-response model. The point estimate is obtained by transforming the totals from the first pass.

4.2 An illustration

We illustrate the technique by a simple example. Let the parameter of interest be the ratio between two products of totals, (or a ratio between two ratios), i.e. we want to find the z-transformation needed to estimate the variance of $\hat{\theta} = \big(\hat{t}_1 \cdot \hat{t}_2\big) \big/ \big(\hat{t}_3 \cdot \hat{t}_4\big)$.

The totals t_1, ..., t_4 are estimated according to the sample design used. The estimator may be either HT or GREG for each total, if GREG is used y_{jk} is replaced by $e_{jk} = g_k\,(y_{jk} - \mathbf{x}_k' \hat{\mathbf{B}}_j)$.

The computation may be done in the following steps.

step	*estimate*	*z-transformation*
1	$\hat{\theta}_1 = \hat{t}_1 \cdot \hat{t}_2$	$\hat{z}_{1k} = \hat{\theta}_1 \left(y_{1k} / \hat{t}_1 + y_{2k} / \hat{t}_2 \right)$
2	$\hat{\theta}_2 = \hat{\theta}_1 / \hat{t}_3$	$\hat{z}_{2k} = \hat{\theta}_2 \left(\hat{z}_{1k} / \hat{\theta}_1 - y_{3k} / \hat{t}_3 \right)$
3	$\hat{\theta}_3 = \hat{\theta}_2 / \hat{t}_4$	$\hat{z}_{3k} = \hat{\theta}_3 \left(\hat{z}_{2k} / \hat{\theta}_2 - y_{4k} / \hat{t}_4 \right)$

Table 4.2. *Intermediate and final z-transformations when a ratio between two products is of interest*.

The variance of $\hat{\theta}_3$ which is our estimator of interest, $\hat{\theta}$ is estimated by using $\hat{z}_{3k}$ and the formula for the variance of $\hat{t}_z = \sum_r \hat{z}_{3k} / (\hat{\nu}_k \pi_k)$. The operations on $\hat{t}_1, \hat{t}_2, \hat{t}_3$ and $\hat{t}_4$ in steps 1-3 may be taken in different orders, all giving to the same result.

5. LINEARIZATION VS. THE JACKKNIFE

There have been some discussions in the literature about the pros and cons of the linearization and Jackknife techniques for the variance estimation of a complex parameter in survey sampling. It has been pointed out that the computation burden for the Jackknife compared with the linearization is substantial. This is especially so when the GREG estimator is used. In a study by Stukel *et al*. (1996) it is reported that 97% of the computer time was spent by the Jackknife and 3% by the linearization estimator.

It has also been claimed that new expressions have to be developed for the variance of each new estimator when linearization is used, which might be a problem in multipurpose surveys. When the Jackknife is used it is only necessary to express the *form* of the parameter of interest.

In practice most parameters of interest can be expressed as rational functions of (possibly domain) totals. This means that by the CLAN approach it is possible to unite the best of the two techniques, low computation burden and a simple expression for the parameter.

However, there may be other reasons than computational feasibility to choose one of the techniques before the other, for example bias and variation of the variance estimates. When the parameter of interest is highly non-linear the linearization may not work very well for small samples, complex parameters needs larger samples. It is difficult to give any general results on this topic, usually one has to rely on Monte Carlo studies to get some hints of how the two techniques work.

In Yung and Rao (1996) and Stukel et al (1996) two different Monte Carlo studies are reported that shed some light on the issue. The design used in both studies was two-stage sampling where the PSUs were taken with PPS and the SSUs were taken with SRS.

In Yung and Rao (1996) the parameter of interest was the ratio between two totals which was estimated by the GREG estimator with the same **x**-vector in the denominator and the numerator. Two different **x**-vectors were used, defining two different poststratifications, (8 and 5+2 poststrata). They found that the linearization and the Jackknife estimators gave approximately the same results with respect to the relative bias of the variance estimators (which were small, <1%) and the coverage of a 95 % confidence interval (which were good, ≈95%).

In Stukel et al. (1996) the parameter of interest was a total estimated by the GREG estimator with an **x**-vector of length fourteen defining 10+4 poststrata. Besides the linearization and Jackknife estimators of variance they also studied the effect of different restrictions on the *g*-weights. They found that the two techniques produced approximately the same results with respect to the relative bias and the relative variation of the two variance estimators for all restriction cases. Both estimators gave variance estimates that were negatively biased with a larger bias for the linearization estimator (≈-6%) compared with the Jackknife (≈-2%). The coefficient of variation of the variance estimates were slightly smaller for the linearization estimator (≈60%) compared to the Jackknife (≈63%). This pattern remained although the differences diminished when the sample size was doubled.

The conclusions from these two studies indicates that the variance estimates produced by the linearization and the Jackknife techniques seem to be similar with regard to bias, variation and coverage of a 95% confidence interval.

For the computation of variance estimates of complex parameters, especially in multipurpose surveys, the smaller computation burden may then favor the linearization technique the way it is implemented in CLAN.

6. NEW DEVELOPMENTS

At Statistics Sweden there is a need for software that can handle advanced non-response models and calibration in two phase sampling. CLAN can handle auxiliary information in two-phase sampling but only for the case when the same **x**-vector is used in the two phases. A software has been developed at Statistics Sweden, based on the CLAN

concept, that can use different **x**-vectors in phase one and two and also allow for calibration from phases two to one and to the population level. The program is currently (1999) under test.

7. REFERENCES

Andersson, C. and 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**, pp. 395-405.

Andersson, C. and Nordberg, L. (1998), *A User's Guide to CLAN97*, Örebro, Sweden: Statistics Sweden.

Deville, J. C., and Särndal, C. E. (1992), "Calibration Estimators in Survey Sampling," *Journal of the American Statistical Association*, **88**, pp. 1013-1020.

Estevao, E., Hidiroglou, M. A. and Särndal, C. E. (1995), "Methodological Principles for a Generalized Estimation System at Statistics Canada," *Journal of Official Statistics*, **11**, pp. 181-204.

Hidiroglou, M. A., Fuller, W. A. and Hickman, R. D. (1976), *SUPER CARP*, Statistical Laboratory, Iowa State University, Ames, Iowa.

Research Triangle Institute (1989), *SUDAAN: Professional Software for SUrvey DAta ANalysis, Version 5.3*, Research Triangle Park, NC, USA.

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

Schnell, D., Kennedy, W. J., Sullivan, G., Park, J. P. and Fuller, W. A. (1988), "Personal Computer Variance Software for Complex Surveys," *Survey Methodology*, **14**, pp. 59-69.

Stukel, D. M., Hidiroglou, M. A. and Särndal, C. E. (1996), "Variance Estimation for Calibration Estimators: A Comparison of Jackknifing Versus Taylor Linearization," *Survey Methodology*, **22**, pp. 117-125.

Tepping, B. (1968), "The Estimation of Variance in Complex Surveys," *Proceedings of Social Statistics Section, American Statistical Association*, pp. 11-18.

Wolter, K. M. (1985), *Introduction to Variance Estimation*, New York: Springer-Verlag.

Woodruff, R. S. (1971), "A Simple Method for Approximating the Variance of a Complicated Estimate," *Journal of the American Statistical Association*, **66**, pp. 411-414.

Woodruff, R. S. and Causey, B.D. (1976), "Computerized Method for Approximating the Variance of a Complicated Estimate," *Journal of the American Statistical Association*, **71**, pp. 315-321.

Yung, W. and Rao, J. N. K. (1996), "Jackknife Linearization Variance Estimators Under Stratified Multi-Stage Sampling," *Survey Methodology*, **22**, pp. 23-31.

TOWARDS A GENERALIZED WEIGHTING SYSTEM

Nico Nieuwenbroek, Robbert Renssen and Lon Hofman, Statistics Netherlands[*)]
Nico Nieuwenbroek, Statistics Netherlands, P.O. Box 4481, 6401CZ Heerlen, The Netherlands
nnwk@cbs.nl

ABSTRACT

At Statistics Netherlands the software package Bascula version 3.0 has been developed which combines weighting of sample data using auxiliary information with variance estimation based on balanced repeated replication (BRR). Much attention has been paid to implement various techniques in an easy and user-friendly way. In this paper we take a quick glance at a new production process where Bascula will be a part of a general processing and estimation environment in which the whole process of outlier detection and handling, editing, imputation and weighting of the clean records will be integrated.

Key Words: Bascula, Regression estimator, Balanced repeated replication, Outliers

1. INTRODUCTION

Statistics Netherlands has been engaged with developing standard software tools, aimed at facilitating and documenting the production process, and improving the total data quality. The idea is to integrate these software tools in the Blaise system. Blaise is used world-wide by many types of survey organizations. One of the software tools is Bascula for Windows, which combines weighting of sample data using auxiliary information with variance estimation based on balanced repeated replication (BRR). It already links up with Blaise.

Due to political pressure Statistics Netherlands is more and more urged to reduce staff costs and response burden. As a consequence the production process has to be redesigned radically. In the traditional production process data collection, processing and dissemination are organized according to the so-called stovepipe model, i.e. many different surveys are performed more or less independently of each other in the course of which each survey has its own way of processing. In the new setting, similar activities of traditional survey processes are consolidated into new (sub-) divisions in order to increase efficiency. For the same reason, external sources, like the VAT (Value Added Tax) registration, have to be used more intensively and combined with data from existing sample surveys. Also, special attention is given to fulfil an ever-growing demand to compare or relate publication figures from different sources. Particularly, publication tables having some marginal counts in common should have identical estimates for these counts. In this paper we discuss the role of Bascula in the new production process. Being a standard tool, Bascula eventually will belong to a general processing and estimation environment in which the whole process of outlier detection and handling (micro and macro) editing, imputation (for unit and item nonresponse), and weighting of clean records will be integrated.

In section 2 a description of Bascula version 3.0 is given. In section 3 the future developments of the redesign is given in more detail, insofar as it concerns the estimation procedure. In section 4 we elaborate on the weighting procedure of the new production process. In section 5 we take notice of the problem to deal with unexpected influential observations that appear to be well recorded at the editing stage. In section 6 some conclusions are given.

2. BASCULA 3.0

In most sample surveys, estimation procedures are applied which take advantage of the presence of auxiliary information. Auxiliary information is defined as a set of variables that are observed for each unit in the sample, while corresponding population aggregates are known from one or more sources such as administrative registers. Poststratification is a well-known application of auxiliary information. The purpose behind the use of auxiliary information is that the precision of estimators will be improved and/or that a possible bias, e.g. due to nonresponse, will be reduced. Before entering into details of Bascula we briefly sketch the general regression estimator.

[*)] The views expressed in this paper are those of the authors and do not necessarily reflect the policies of Statistics Netherlands.

Consider a finite population of distinguishable (labeled) units, $U = \{1,..., N\}$ where the kth unit is associated with a value y_k of study variable y and with a vector $\mathbf{x}_k = (x_{k1},..., x_{kp})^t$, which contains the values of p auxiliary variables. The corresponding population totals are given by $t_y = \sum_{k\in U} y_k$ and $\mathbf{t}_x = \sum_{k\in U} \mathbf{x}_k$, respectively. The objective is to estimate unknown finite population parameters. To be able to estimate t_y a probability sample $S\subset U$ is drawn according to a given sampling design, inducing strictly positive first and second order inclusion probabilities $\pi_k =$ Pr ($k\in S$) and π_{kl} = Pr ($k\&l\in S$). A convenient way and common practice is to express the estimator $\hat{t}_y$ for t_y as a weighted sample sum, $\hat{t}_y = \sum_{k\in s} w_k y_k$. The Horvitz-Thompson estimator $\hat{t}_{y\mathrm{HT}} = \sum_{k\in s} d_k y_k$ where $d_k = \pi_k^{-1}$ is the sampling weight (design weight or inclusion weight), is a well-known example of such a weighting type estimator. In practice we deal with responding units in which case d_k refers to net sampling weights. Incorporating auxiliary information implies that adjusted weights w_k are determined, such that the estimator $\hat{\mathbf{t}}_x = \sum_{k\in s} w_k \mathbf{x}_k$ exactly reproduces $\mathbf{t}_x$. We then say that the weighted sample is balanced, or calibrated with respect to the used auxiliary variables. There are several estimators that satisfy the imposed restriction; among them the general regression estimator is a very important one. The general regression estimator, $\hat{t}_{y\mathrm{R}}$, is defined as

$$\hat{t}_{y\mathrm{R}} = \hat{t}_{y\mathrm{HT}} + \hat{\mathbf{b}}^t(\mathbf{t}_x - \hat{\mathbf{t}}_{x\mathrm{HT}}), \tag{2.1}$$

with

$$\hat{\mathbf{b}} = (\sum_{k\in s} d_k(\mathbf{x}_k\mathbf{x}_k^t / v_k)^{-1} \sum_{k\in s} d_k(\mathbf{x}_k y_k / v_k),$$

where the v_k are some known non-zero constants, assuming that $\hat{\mathbf{T}}_{xx\mathrm{HT}} = \sum_{k\in S} d_k(\mathbf{x}_k\mathbf{x}_k^t / v_k)$ is nonsingular. In Särndal et al. (1992) the quantity v_k is assumed to be related to the variance structure (by $\sigma_k^2 = v_k\sigma^2$, with at least v_k known) of the linear regression model of y on $\mathbf{x}$, where the y_k are supposed to be realized values of independent random variables. The general regression estimator, given by (2.1) can also be expressed as the weighted sample sum of the observed observations of the study variable, with regression weights

$$w_k = d_k[1 + (\mathbf{x}_k^t / v_k)(\hat{\mathbf{T}}_{xx\mathrm{HT}}^{-1})(\mathbf{t}_x - \hat{\mathbf{t}}_{x\mathrm{HT}})] = d_k g_k. \tag{2.2}$$

It follows from (2.2) that the regression based weighting method can be described as linear weighting given the sample, with the adjusted weight w_k written as the product of sampling weight d_k and correction weight g_k. The general regression estimator thus induces sampling weights to be transformed into regression weights. In a multipurpose survey, for each study variable a different set of auxiliary variables may be used. However, for practical reasons often the same auxiliary information is used for all study variables in which case only one set of adjusted weights is needed.

In Deville and Särndal (1992) a class of calibration estimators is given, where weights are sought which are (given some distance measure) as close as possible to the sampling weights such that balancing with respect to a set of auxiliary variables is achieved. As the general regression estimator generalizes many textbook estimators, the system of calibration estimators can be viewed as a generalization of the general regression estimator. In the view of calibration the calculation of only one set of weights fits very well.

Bascula has been especially developed for computing adjusted weights and aims at the estimation of population totals, means and ratios with corresponding variances. The latest version 3.0 includes:

- Weighting methods as poststratification, ratio estimation, and more generally linear weighting based on the general regression estimator,
- The possibility to achieve an integrated weighting approach for persons and households, as suggested by Lemaître and Dufour (1987),
- A bounding technique based on the Huang Fuller algorithm (1978),

- Multiplicative weighting based on the algorithm in Deming and Stephan (1940) as an alternative for linear weighting when only categorical auxiliary variables are used ,
- A module for estimating variances based on balanced repeated replication (BRR), see Wolter (1985).

The estimation of variances based on BRR offers an attractive alternative for the use of formulas based on Taylor linearization, especially when thought from a weighting perspective. In general both methods are aymptotically equivalent, see Shao and Tu (1995). BRR was originally developed for stratified multistage designs by which in each stratum two primary sampling units (PSUs) are drawn with replacement in the first stage, see McCarthy (1969). The underlying idea is to form a balanced set of half-samples or replicates by deleting one PSU from the sample in each stratum. A minimal set of balanced R half-samples is constructed from an $R \times R$ Hadamard matrix, see Wolter (1985), R being a multiple of 4 with $H \leq R \leq H+3$ and H the number of strata. For notational convenience we limit ourselves to stratified one-stage unit sampling. Let the kth unit of the sample correspond to the ith unit (PSU) of the hth stratum. For each half-sample a replicate estimator can be obtained using the same formula as for the full sample with sampling weights d_{hi} replaced by resampling weights d_{hi}^{α}. Similarly to the full sample adjusted estimate, it is also possible to express its corresponding replicate estimates $\hat{t}_y^{\alpha}$ in terms of adjusted weights. Actually, the weighting procedure transforms the resampling weights d_{hi}^{α} into adjusted resampling weights w_{hi}^{α} and these weights can be used to make the estimates for the resamples. Deleting half the sample correspondents with doubling the original sampling weight for one half-sample and using zero weights for its complement. Especially when using auxiliary information this may give unstable or even undefined results for some replicates. To avoid such a problem, we have followed an idea proposed by Fay (1989) by which the sampling weights are less sharply disturbed. The modified resampling weights thus become

$$d_{hi}^{\alpha} = d_{hi}[1 + q\delta_{\alpha h}(\Delta_{hi}^{1} - \Delta_{hi}^{2})] \ ,$$

with $0<q\leq 1$. The value of $\delta_{\alpha h}$ follows from the relating Hadamard matrix, denoting which PSU in stratum h is in the αth resample; for $\delta_{\alpha h} = 1$ it is the first PSU and for $\delta_{\alpha h} = -1$ it is the second one. Further Δ_{hi}^{1} equals 1 if the ith PSU in stratum h is assigned as the first PSU and 0 otherwise. A similar definition is used for Δ_{hi}^{2} with respect to the second PSU. The value $q = 1$ corresponds with the standard situation. With $q<1$, no units are actually deleted; in this case the units in the not selected half-sample only receive less weights than those in the other half-sample. With Fay's method the BRR variance estimator becomes

$$v_{BRRq} = \frac{1}{q^2 R}\sum_{\alpha=1}^{R}(\hat{t}_y^{\alpha} - \hat{t}_y)^2 \ . \qquad (2.3)$$

Rao and Shao (1999 studied some theoretical properties of the modified BRR estimator. Once the adjusted resampling weights have been calculated, the user can carry out variance estimation without bothering about the design stage and estimation stage. All required information must be properly contained in the replicate weights.

For a broader utilization of Bascula, BRR has been extended to stratified multi-stage sampling where in some or all strata more than two PSUs are drawn (with replacement). The multi-stage design is approximated by a design for which the basic two-per-stratum procedure can be applied, e.g. by randomly forming two groups of PSUs per stratum, or by randomly forming artificial strata, each with two (groups of) PSUs, see Wolter (1985). In these cases BRR is applied to the groups, while artificial strata are viewed as strata. For designs with a relatively small number of strata and large stratum sizes, variance estimation based on grouped BRR may be quite inefficient, while Rao and Shao (1996) showed design-inconsistency. In order to overcome such problems they proposed independently repeating the grouping T times and then taking the average of the resulting T variance estimators; creating artificial strata may be used as an alternative. By means of a simple modification for the finite population, Bascula can also handle stratified one-stage designs without replacement in each stratum. Inspired by Rao and Wu (1988) who suggested bootstrap replications for two-stage designs in which simple random without replacement is carried out at both stages, we modified BRR to be applied to such designs. In Renssen et al. (1997) it is described how half-samples are formed for the implemented sample designs; also the derivation of expressions for the corresponding resampling weights are given.

Bascula internally forms resamples with resampling weights using Fay's (1989) method. Subsequently the weighting procedure for the full sample is repeated for each of the R resamples, where the resampling weights are viewed as sampling weights. In this way R+1 sets of adjusted weights are written to an external file. The first set of weights in this file refers to the full sample and the remaining R sets to the resamples. With these R+1 sets of adjusted weights variances can be estimated for arbitrary study variables using the formula given in (2.3). We have chosen for $q = 0.57$, which simply guarantees that the resampling weights are strictly positive for all implemented sampling designs. Apart from the calculation of the adjusted weights it is also possible to let Bascula itself estimate the point estimates and corresponding standard errors.

Of course, the calculation of adjusted resampling weights may demand a lot of computation time, especially when relatively many resamples are needed. On the other hand, after the determination of the set of auxiliary variables the computation needs to be carried out only once. When these weights have been calculated, variances can be estimated in an easy way using a simple formula for various study variables without knowing them in advance. Nevertheless it has been decided recently to implement the estimation of variance based on first-order Taylor linearization as a welcome alternative. This method can handle the same sampling designs as already implemented for the BRR method and can be applied in simple cases or especially when the user wishes to compare variances for various weighting schemes in order to decide which scheme will be chosen eventually. The implementation of the Taylor linearization is now in its test phase.

Considerable effort has been put into a user-friendly interface resulting in clear windows for the actions to be undertaken, where context sensitive help information is available. Bascula works with tabbed sheets, which are displayed on the desktop. Each sheet is accessible by a tab. Tabs appear while building the setup. The order of the tabs reflects the order in which the various tasks have to be carried out. Consequently it has been decided to make a certain tab only visible and accessible after some minimal preliminary work has been done. The main tasks concern: specification of the sample data file (for a Blaise data file the information is already available in the meta file), specification of population information via tables followed by entering the data, specification of weighting scheme denoting which set of auxiliary variables are involved in the weighting session, selection of the weighting method, specification of design information, adding information for resampling, and the specification of study variables (when the package has to compute point estimates and possibly standard errors).

Actually, Bascula has been developed for individual sample surveys. It is intensively in use for various person and household surveys, and recently also for business surveys. Person and household surveys use mainly categorical auxiliary information (sex, region), while auxiliary information in business surveys is largely continuous (VAT information). However, the complete production process will be redesigned radically, where coherence of statistical output is very important. This new development has fundamental consequences for Bascula. In the following sections we elaborate on this.

3. OUTLINE OF THE NEW PRODUCTION PROCESS

In this section an outline of the new statistical production process of Statistics Netherlands is given. More details can be found in Keller et al. (1999). Roughly, in the new production process we distinguish four stages.

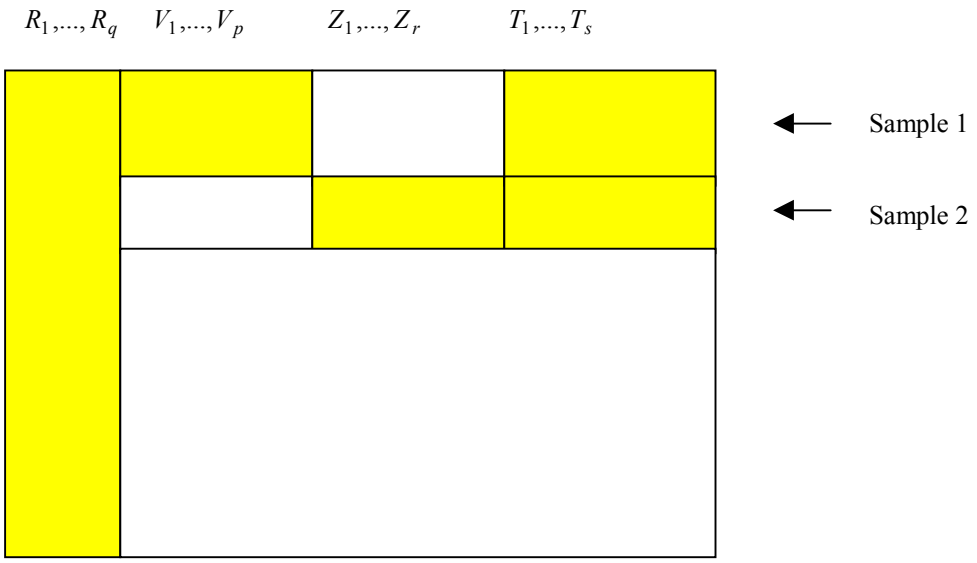
At the first stage, data from different sources such as administrative registers, sample surveys, EDI (Electronic Data Interchange) are put together into an input database called BaseLine. The units of BaseLine may deviate from statistical units and will generally refer to different object-types. Somewhat disrespectful, BaseLine can be considered as an immense reservoir of all input data from different sources.

At the second stage, the input database is used to create one or more micro databases. A micro database can be seen as a rectangular table with objects (statistical units) of one object-type on its rows and scores of variables on its columns. For example, figure 1 exhibits a prototypical micro database. The shaded surfaces are filled with observations; the non-shaded surfaces correspond to missing data.

The third stage can be considered as an estimation stage. Each publication figure (estimate) to be disseminated by Statistics Netherlands should satisfy at least three statistical requirements. Firstly, all point estimates must be approximately (design) unbiased with reasonably small design variances (accuracy) and any interval estimate should be valid in the sense that the "actual interval coverage" should be larger than or equal to the "nominal interval coverage". Secondly, all estimates should be mutually consistent. By comparing two or more estimates or

combinations of estimates no contradictions may occur, even not on account of sampling error. Thirdly, all estimates should pass the rules of disclose control. Especially the second requirement is typical for the new approach and calls for new weighting methods, see section 4. The central database at the third stage is called StatBase. This database can be seen as a storehouse of estimates Statistics Netherlands is willing to publish. That is, all estimates in StatBase should satisfy the statistical requirements stated above. Besides these estimates should be worthwhile, i.e. wanted by users.

Figure 1. A prototypical micro database



The fourth stage can be considered as a presentation stage. The contents of StatBase are presented to the user by StatLine as one giant publication. It is typical for StatLine to condense a large number of "paper tables" into a limited number of "electronic" datacubes. By selecting on specific items StatLine enables the user to view all kinds of cross-sections of such a datacube. Consequently, users are able to confront all kinds of statistical figures, which may very well originate from more several surveys. This modern way of dissemination forces statistical output not only to be comparable, but moreover to be consistent in the sense that each figure supported by StatLine should fit into one of these datacubes.

4. REPEATED WEIGHTING

One of the major methodological challenges is to develop a general estimation strategy to proceed from the micro database to StatBase. In this section we only sketch the procedure, see also Kroese and Renssen (1999). For technical details we refer to Renssen et al. (2000). Emphasis lies on the accuracy and consistency requirement; the requirement with respect to disclosure control is postponed as future research. In section 5 we touch upon the validity of any interval estimate.

Briefly, we consider the problem of estimating a specific *m*-way table, taking into account any related table in StatBase. Two (or more) tables may be related because of some common "marginal" cells. For example, they may describe similar phenomena, yet for different classification variables, or different phenomena for the same classification variable. In the first case, the overall total of both tables should coincide, while in the latter case the class sizes should match. Renssen et al. (2000) also discuss relationships specified by edit rules. However, we only consider the first type of relationships, i.e. we only consider relationships on account of common marginal cells.

We consider the prototypical micro database as exhibited in figure 1 and divide this database into four sub-databases, namely the administrative registration with *R*-variables , two samples (S_1 and S_2) with respectively *V*- and *Z*-variables, and the union of these samples ($S_1 \cup S_2$) with *T*-variables. Estimating a specific *m*-way table involves among others the determination of the proper sub-database. For example, if the estimation concerns only *R*-

variables the proper sub-database is *U*, while a crossing between *V* and *R*-variables should be estimated from $S_1 \cap U = S_1$. Note that a crossing between *V*- and *Z*-variables should be estimated from $S_1 \cap S_2 = \varnothing$, which is empty according to figure 1. Estimates that are based on *U* concern (straightforward) register counts. If estimates are based on S_1, S_2, or $S_1 \cup S_2$ then the general regression estimator as discussed in section 2 is used. Now, utilizing the general regression estimator we need 1) the starting weights of the sampling units, 2) a specification of the weighting scheme, i.e. a specification of the vector of auxiliary variables, and 3) estimated population totals of the vector of specified auxiliary variables.

4.1. Deriving starting weights

An application of the general regression estimator involves, among others, the determination of the starting weights. These starting weights are easily derived if the proper database corresponds to precisely one sample. However, if the proper sub-database consists of the union of two samples these starting weights can be derived in several ways. We discuss two of them.

- Let π_{1k} and π_{2k} denote the first order inclusion probabilities of the *k*th unit, $k \in U$, with respect to S_1 and S_2 respectively. Then the first order inclusion probability of this unit with respect to $S_1 \cup S_2$ equals $\pi_{1k} + \pi_{2k} - \pi_{1k}\pi_{2k}$ (it is conveniently assumed that S_1 and S_2 are independent). Taking the inverse of this inclusion probability we arrive at the starting weight of the *k*th unit with respect to the union of both samples.

The problem of this approach is that both π_{1k} and π_{2k} should be derived for all $k \in S_1 \cup S_2$. Often, only π_{1k} are derived for $k \in S_1$ and π_{2k} for $k \in S_2$. Furthermore, the derivation of the first order inclusion probabilities may be troublesome if the independence assumption no longer holds. A more practical approach is the following.

- Let π_{1k} and π_{2k} denote the first order inclusion probabilities of the *k*th unit, $k \in U$, with respect to S_1 and S_2 respectively, and define $d_k^* = \lambda\pi_{1k}^{-1}$ for $k \in S_1$ and $d_k^* = (1-\lambda)\pi_{2k}^{-1}$ for $k \in S_2$, where $\lambda \in [0,1]$.

The resulting general regression estimator of this second approach resembles the traditional general regression estimator, with the difference that pooled Horvitz-Thompson estimators are used instead of Horvitz-Thompson estimators. The choice of λ may reflect the confidence in the one sample compared to the other. It may depend on indicators for several survey errors, such as sampling errors, nonresponse errors, or measurement errors.

4.2. Specifying the (minimal) weighting scheme

The traditional way to construct estimates, and hence to fill StatBase, is to use one set of weights per survey, or in our case, one set of weights per sub-database, see section 2. When using one set of weight per sub-database, all involved variables are inflated in the same way. The main advantage of such an approach is that once the set of weights has been calculated it can be applied directly to any set of study variables. However, this traditional approach has a striking disadvantage as will be illustrated below.

Consider a register (e.g. the Dutch Municipal Base Administration) with the variables sex (2 categories), age (100 categories), marital status (4 categories), and region (600 categories). The register information for publication purposes is the complete crossing between the register variables resulting into 48×10^4 register counts. Suppose that a sample survey (of size n = 100,000 persons) is matched to this register. Obviously, utilising the complete crossing for calibration will result in many cells with few or no observations, in which case the regression estimator is undefined. So, one is forced to use an incomplete crossing instead. But then the weighted sample gives inconsistent estimates for all sample variables that are crossed with those register variables that are excluded for calibration. On account of the consistency requirement, either these crossings or these register counts should be excluded from StatBase. We note that the deterioration becomes worse as more sample surveys are matched.

In order to avoid unnecessary deterioration, Kroese and Renssen (1999) suggested a new estimation strategy. This estimation strategy differs from traditional weighting in that it no longer sticks to one set of weights per sub-database. For each *m*-way table to be estimated one may look for a weighting scheme that guarantees consistency with all related tables that already have been estimated (and stored in StatBase). If the sample size is large enough to

use the Horvitz-Thompson estimator for estimating a specific *m*-way table (in a view of accuracy), then the sample is also large enough to use the general regression estimator with a minimal set of auxiliary variables, ensuring consistency with all related tables in StatBase. If desired, one may enlarge the minimal set to meet the accuracy requirement. Once the new cell estimates are added to StatBase, the weights are of no use any more. Only the weighting scheme according to which the weights are calculated is stored on behalf of the process information.

Enlarging the weighting model on account of the accuracy requirement has some theoretical advantages, especially for large sample sizes. However, the practical implementation may be troublesome. In addition, the enlarged number of auxiliary variables to be used in the general regression estimator may become too large. Therefore, as an alternative, we suggest the following procedure. Firstly, we derive regression weights according to the traditional approach. That is, per sub-database (sample survey) we derive regression weights according to some overall weighting scheme to adjust for sampling error and nonresponse, noting that this traditional way of weighting already meets some consistency requirements. Secondly, per *m*-way table to be estimated, the overall regression weights are (minimally) adjusted to ensure absolute consistency. This can be accomplished by taking the overall regression weights as starting weights and using the minimal weighting scheme to define the set of auxiliary variables.

5. RESISTANT WEIGHTING METHODS

5.1. Validity of interval estimates; the central limit theorem

Assuming that the general regression estimator is approximately design unbiased and that approximate formulas are available for its design variances (e.g. see Särndal et al. 1992), the first requirement of StatBase estimates is only discussed insofar it concerns the accuracy. No attention is paid to the validity of any interval estimate. Now, justified by the central limit theorem, general regression estimates frequently are assumed to be approximately normally distributed for large sample sizes. For sampling without replacement from finite populations, Hájek (1964) has given sufficient and necessary conditions under which the sampling distribution of the sample mean tends to normality. As a rule of thumb, the percentage points may be taken from the *t*-distribution if sample (cell) sizes are smaller than 50, see Cochran (1977, page 27). However, as Cochran stated, special methods are needed for small samples with very skew distributions. One reason for skewed sample distributions is the presence of outliers in the (finite) population.

To be more precisely, Chambers (1986) distinguished two types of sampled outliers, namely representative and non-representative outliers. A representative outlier is a sample unit with a value that has been recorded correctly and that cannot be regarded as unique. Although some of its values differ substantially from most sample values, such a unit is not unique; the nonsampled part of the population may contain similar units. The second type, an unrepresentative outlier, is typically associated with a sample unit whose values are incorrectly measured (gross errors) or with a sample unit whose values are unique in the sense that there are no other units like them in the finite population. It is assumed that gross errors are already treated according to some editing strategy, see e.g. De Waal et al. (2000). Unique outliers are simply excluded from the estimation process by giving them unit sampling weights. There remains the problem of handling representative outliers.

The inclusion or exclusion of large units (representative outliers) may influence the general regression estimator so much that it becomes unreliable. For this reason, samples are usually designed such that large units are selected with certainty. However, even though such designs can minimize outlier problems, they cannot eliminate them completely. There are two main concerns in dealing with representative outliers: the efficiency of the general regression estimates and the applicability of the central limit theorem. Tukey (1960) first demonstrated the dramatic lack of efficiency of classical estimates, like the arithmetic mean, in the presence of outliers. There is no reason to believe in a better performance of general regression estimates. The central limit theorem states that the sum of a large number of equally small independent errors follows approximately a normal distribution. The presence of a few extreme observations violates the equally assumption.

5.2. Handling representative outliers

Many of the traditional proposals for dealing with skewed populations relate to the choice of the sampling scheme. If prior information about large values is available to the sampler then one obvious strategy is to stratify the population and to put all the large values into a separate stratum. Since these values are influential it is frequently proposed that this stratum should be evaluated completely. If the prior information is not perfect or totally absent, then Kish (1965) proposes the construction of a (post-) stratum for surprises in which all influential observations are

placed. The problem is then to estimate the size of this stratum. As Smith (1987) noted, such procedures for skewed populations all implicitly assume that the target population contains influential observations. The problem is not just to identify such influential observations in the sample, but the additional one of estimating the total number of them in the population. If a population contains some large influential observations then estimates of population totals are often as badly affected by samples, which do not contain any large values as by those that do. Thus in sample surveys the detection of influential values in the sample is only part of the story; influential values which fail to appear in the sample may also be very important.

Given a specific *m*-way table to be estimated on account of StatBase, we divide the outlier problem into two sub-problems. The first concerns the detection of any influential observations of continuous variables per cell, while the second one concerns the determination of the sampling weights of such observations.

5.2.1. Detecting influential observations

For one-dimensional count variables, one may simply detect outlying observations by means of the couple median and median absolute deviation:

$$O_k = \frac{\left| y_k - \underset{i \in S_c}{med}(y_i) \right|}{\underset{i \in S_c}{mad}(y_i)} \quad k \in S,\ S_c = \text{subsample corresponding to specific cell}.$$

As O_k increases, the influence of the *k*th unit on the *c*th cell estimate will be larger. There exist several robust outlier detection methods for multidimensional continuous variables, such as the minimum volume ellipsoid or the minimum covariance determinant. We describe a procedure suggested by Kosinski (1999) and recommended by De Boer and Veltkamp (2000). Assuming that the sample data can be divided into "good" data and "bad" data, the procedure involves the following five steps:

1. Start with a few, say *g*, good points,
2. Calculate the sample mean and the sample covariance matrix,
3. Calculate the traditional Mahalanobis distance of the complete data set,
4. Increase the "good" part with one point by selecting *g*+1 points with the smallest Mahalanobis distance
5. Return to step 2 or stop as soon as the good part of the data contains more than half the data set and the smallest Mahalanobis distance of the remaining points is higher than a predetermined cut off value.

In order to assure that the good part will contain no outliers at the end, it is essential to start the algorithm with good points. After some proper re-scaling, De Boer and Veltkamp (2000) suggest to use co-ordinate-wise a robust estimator for location. We conjecture that this algorithm guarantees a solution if there is just one bulk of good data.

5.2.2. Resistant regression estimators

As is illustrated in subsection 5.1, one reason for the need of resistant estimators in sample surveys is the appearance of representative outliers. Resistant estimation should imply the use of an estimator such that its distribution is essentially unaffected by sample outliers (Chambers, 1986). In general, resistant estimators will be design biased. However, it is hoped that the decrease in design variance will outweigh the increase in design bias. Chambers (1986) and Gwet and Rivest (1992) have elaborated such an approach. The last authors stated that the unconditional bias of their resistant estimator should be viewed as a kind of premium paid for being protected against occasionally wild samples. Furthermore, they stated that the design unbiasedness of most traditional estimators is deceptive, because it hides a high conditional bias that is obtained when the sample proportion of (correctly measured) outliers differs much from the population proportion. Here, the conditioning is with respect to the number of representative outliers in the sample.

Consider the general regression estimator as stated in section 2. If there exists a *p*-vector **a**, such that $\mathbf{a}^t \mathbf{x}_k = v_k$ for all $k \in U$, then this estimator can be rewritten as $\hat{t}_{y\mathrm{R}} = \hat{\mathbf{b}}^t \mathbf{t}_x$. Now, it is typical for least squares estimates as $\hat{\mathbf{b}}$ that they are sensitive to sample outliers, from which it follows that $\hat{t}_{y\mathrm{R}}$ is also sensitive to outliers. To robustify the regression estimator we have to replace $\hat{\mathbf{b}}$ by some robust version. We note that Chambers (1986) and Gwet and

Rivest (1992) have elaborated on such an approach for one-dimensional (continuous) *x*-variables. However, in view of minimal weighting models we need resistant regression estimators for both categorical and continuous **x**-variables.

For categorical **x**-variables we may apply Huber's (1973) approach to robustify the least squares approach without bothering about so-called leverage points. That is, we minimize

$$\sum_{k \in S} \rho\left(\frac{y_k - \theta_y^t \mathbf{x}_k}{S_y}\right)$$

with respect to θ_y, where ρ is some suitable real-valued convex function and S_y is some preliminary robust estimator of scale. This gives residuals, indicating the explanatory power of the auxiliary information. These residuals may be used to define resistant starting weights ω_k. For example one may take $\omega_k = 1$ if the *k*th residual is considered as a unique outlier and not explained by the auxiliary information, and $\omega_k = \pi_k^{-1}$ otherwise (i.e. considered as a bulk value), where it should be noted that records with unit weights are left outside Bascula. For simultaneously using categorical and continuous variables more studies are needed.

6. CONCLUSIONS

Bascula's task is to determine weights that are adjusted for the incorporation of auxiliary information, mainly based on the application of the general regression estimator. In the first instance the estimation of variances is based on BRR. The package forms half-samples from the full sample and determines resampling weights. With these resampling weights adjusted resampling weights are calculated in the same way as for the full sample. Following the concept of estimating population totals for arbitrary study variables, i.e. by means of weighting, their corresponding variances can also be derived by means of resampling weighting. Recently Taylor linearization has been implemented as an alternative for BRR.

Actually, Bascula has been developed for individual sample surveys. In sections 3 and 4 a rough outline of the expected redesign of the production process is described where coherence of statistical output is very important. Statbase will contain accurate and mutually consistent estimates. This database is built by successively adding estimates, taking into account earlier related estimates. It will be clear that Bascula can be used successfully for carrying out repeatedly the general regression estimator. In case of minimal weighting, survey data is calibrated using earlier determined and related estimates as auxiliary information.

Another issue concerns the treatment of representative outliers. Bascula is capable to restrict correction weights. This kind of bounding, however, is merely based on the values of auxiliary variables. For outliers with respect to study variables, we have to resort to other methods. Here the solution may be found in the use of register information in a robust context. If the solution can be found in a restriction of the sampling weights, then Bascula can also usefully be applied to obtain resistant estimates. However, more investigation is needed.

REFERENCES

Chambers, R.L. (1986), "Outlier Robust Finite Population Estimation," *Journal of the American Statistical Association*, 81, pp. 1063-1069.

Cochran, W.G. (1977), *Sampling Techniques*, third edition, New York: Wiley.

De Boer, P., and V. Veltkamp (2000), *Robust Multivariate Outlier Detection*, Research paper no. 0003, The Netherlands: Statistics Netherlands.

Deming, W.E., and F.F. Stephan (1940), "On a least squares adjustment of a sampled frequency table when the expected marginal totals are known," *Annals of Mathematical Statistics*, 11, pp. 427-444.

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

De Waal, T., R.H. Renssen, and F. Van de Pol (2000), "Graphical Macro-editing: Possibilities and Pitfalls," unpublished report, The Netherlands: Statistics Netherlands.

Fay, R.E. (1989), "Theory and application of replicate Weighting for Variance Calculations," *Proceedings of the Social Statistics Session, American Statistical Association 1989*, pp. 212-217.

Gwet, J.P., and L.P. Rivest (1992), "Outlier Resistant Alternatives to the Ratio Estimator," *Journal of the American Statistical Association*, 87, pp. 1174-1182.

Hájek, J. (1960), "Limiting Distribution in Simple Random Sampling from a Finite Population," *Publ. Math. Inst. Hungarian Academy of Science.* 5, pp. 361-374.

Huang, E.T., and W.A. Fuller (1978), "Nonnegative Regression Estimation for Survey data," *Proceedings of the Social Statistics Session, American Statistical Association 1978*, pp. 300-303.

Huber, P.J. (1973), "Robust Regression, Asymptotics, Conjectures, and Monte Carlo," *Annals of Statistics*, 1, pp. 799-821.

Keller, W., J. Bethlehem, A. Willeboordse, and W. Ypma (1999)," Statistical Processing in the Next Millenium," *Proceedings of the XVIth Annual International Methodology Symposium on Combining Data from Different Sources, May 1999 Canada*.

Kish, L. (1965), *Survey Sampling,* New York: Wiley.

Kosinski, A.S. (1999),"A Procedure for the Detection of Multivariate Outliers," *Computational Statistics & Data Analysis*, 29, pp. 145-161.

Kroese, A.H., and R.H. Renssen (1999), "Weighting and imputation at Statistics Netherlands," *Proceedings of the IASS conference on Small Area Estimation, Riga August 1999*, pp. 109-120.

Lemaître, G. and J. Dufour (1987), "An integrated Method for Weighting Persons and Families," *Survey Methodology*, 13, pp. 199-207.

McCarthy, P.J. (1969), "Pseudo-replication: Half samples," *Review of the International Statistical Institute*, 37, pp. 239-264.

Rao, J.N.K., and J. Shao (1996), "On Balanced Half-Sample Variance Estimation in Stratified Random Sampling," *Journal of the American Statistical Association*, 91, pp. 343-348.

Rao, J.N.K., and J. Shao (1999), "Modified Balanced Repeated Replication for Complex Survey Data," *Biometrica*, 86, pp. 403-415.

Rao, J.N.K. and C.F.J. Wu (1988), "Resampling Inference with Complex Survey Data," *Journal of the American Statistical Association*, 83, pp. 231-241.

Renssen R.H., A.H. Kroese, and A.J. Willeboordse (2000), "Aligning Estimates by Repeated Weighting," unpublished report, The Netherlands: Statistics Netherlands.

Renssen, R.H., N.J. Nieuwenbroek, and G.T. Slootbeek (1997*), Variance Module in Bascula 3.0: Theoretical Background*, Research paper no. 9712, The Netherlands: Statistics Netherlands.

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

Shao, J., and D. Tu (1995), *The Jackknife and Bootstrap*, New York: Springer.

Smith, T.M.F. (1987), "Influential Observations in Survey Sampling," *Journal of Applied Statistics*, 14, pp. 143-152.

Tukey, J.W. (1960), "A Survey of Sampling from Contaminated Distributions," in Contributions to Probability and Statistics, I. Olkin, Ed. Stanford University Press, Stanford, Calif., pp. 448-485.

Wolter, K.M. (1985), *Introduction to Variance Estimation*, New York: Springer.

ESTIMATION AND VARIANCE ESTIMATION IN A STANDARD ECONOMIC PROCESSING SYSTEM

Richard Sigman[1], U.S. Census Bureau
ESMPD, Room 3108-4, U.S. Census Bureau, Washington DC 20233, USA rsigman@census.gov

The U.S. Census Bureau has developed software called the Standardized Economic Processing System, or StEPS, that it plans to use to replace 16 separate systems, which are currently used to process over 100 current economic surveys. This paper describes the methodology and design of the StEPS modules for estimation and variance estimation and chronicles our experiences in using these modules to migrate surveys into StEPS. The paper concludes with a discussion of possible future enhancements to the estimation and variance estimation functions in StEPS.

Key Words: Survey Processing, Economic Surveys, StEPS

1. Introduction

The U.S. Census Bureau conducts over one hundred establishment surveys. Many of these are surveys of commercial businesses. A small number are surveys of government institutions. The Census Bureau refers to these surveys as *economic surveys* because they collect quantitative data describing business operations of survey units. Also, these surveys provide economists and other analysts with estimates and data sets needed for macro- and micro-economic analyses. For example, the Bureau of Economic Analysis uses estimates from economic surveys to determine the national income and expense accounts.

Economic surveys can differ widely with respect to characteristics of reporting units and content of survey questions. They are often similar, however, with respect to data-processing requirements, which has prompted the Census Bureau to begin consolidating the survey-processing systems for many of its economic surveys. The development and use of generalized software, called the Standard Economic Processing System (StEPS) has made this possible.

This paper describes the current capabilities of StEPS for calculating survey estimates and associated sampling errors. Sections two through four provide background material. In particular, section two summarizes the characteristics of Census Bureau economic surveys that are relevant to calculating survey estimates and sampling variances. Section three discusses variance estimation methods: those used in the legacy systems, those evaluated for StEPS, and those currently implemented in StEPS. Section four briefly describes the entire StEPS system. Sections five through eight focus on the StEPS Estimates and Variances Module. Section five describes the components of the module, and section six presents and explains two examples of StEPS estimation scripts. Section seven describes our implementation experiences for the Estimates and Variances Module in 1998 and 1999. Finally, section eight describes future activities and possible enhancements.

2. Economic Surveys Conducted by the Census Bureau

The Census Bureau consists of several directorates that conduct censuses and surveys. The most widely known directorate is the Decennial Census Directorate, which conducts the demographic decennial census. Another directorate, called the Economic Programs Directorate, conducts economic censuses every five years and conducts current economic surveys monthly, quarterly, and annually in areas of manufacturing, construction, commercial services, government services, and foreign trade. The directorates, such as the Decennial Census Directorate and the Economic Programs Directorate, are responsible for developing survey methods and associated processing systems for the censuses and surveys they conduct.

[1]This paper reports the results of research and analysis undertaken by Census Bureau staff. It has undergone a more limited review by the Census Bureau than its official publications. This report is released to inform parties and to encourage discussion.

In 1995 the Economic Directorate began to consolidate its processing systems for current surveys. This was preceded, however, by two activities that provided information that was used in planning and directing the consolidation effort. One of these was a compilation of an inventory of the Economic Directorate's statistical practices carried out by King and Kornbau (1994). Some of their findings (along with updates for recent survey changes) are the following:

- Sample designs are primarily single stage, but a number of different sampling methods are used: stratified sampling, cut-off sampling, Poisson sampling, and Tillé sampling.

- Only four types of estimators for (unraked) totals are used: unweighted estimator (used by cut-off surveys), Horvitz-Thompson estimator, ratio estimators (combined and separate), and link-relative estimator. (As a result of a 1997 redesign, the monthly surveys of retail and wholesale trade no longer use composite estimation.)

- A number of different variance estimation methods are used by the economic surveys that calculate sampling variances: jackknife, method of random groups, balanced repeated replication and sampling-theory formulas.

The other activity that provided useful information was the development in 1994 of a processing system for the Farm and Ranch Irrigation Survey (FRIS). This demonstrated the feasibility of the following in developing a production processing system for an economic survey:

- The use of reusable SAS® code configured to individual surveys by analyst-specified parameters,
- The use of interactive screens to allow analysts to specify parameters, and
- The use of a general-purpose variance-estimation program, VPLX, to allow sample designers to specify how to calculate standard errors.

3. Variance Estimation Methods

The Economic Directorate decided to consolidate its current-survey processing systems by developing a Standard Economic Processing System, called StEPS. Because one of the functional requirements for StEPS was the estimation of sampling variances, we carried out several research studies on variance estimation during the development of StEPS. These studies explored two possible development approaches: (1) reduce the number of different variance estimation methods, and (2) use available computer programs for calculating design-based variances.

The available computer programs we studied were VPLX, developed in-house by Fay (1990), and SUDAAN, developed by Research Triangle Institute (RTI). VPLX estimates variances by means of replication . It contains options to estimate variances using random groups, jackknife, stratified jackknife, balanced repeated replication, and generalized replication. (See Wolter 1985 or Rust 1985 for descriptions of these variance estimation methods.) Prior to the use of VPLX in the FRIS system, VPLX had been used very little by the Census Bureau's economic surveys, but it is used extensively by the Census Bureau's demographic surveys of households and by the 2000 population census. RTI has recently added replication-based methods to SUDAAN, but at the time of our investigation these methods were not available.

By performing repeated stratified sampling from a simulated FRIS population, Tremblay and Sigman (1996) compared stratified-jackknife variance estimates calculated by VPLX to variances estimates calculated by SUDAAN using sampling theory formulas--i.e., S^2 formulas with Taylor-series approximations. One objective of this study was to evaluate the two programs as to their suitability for inclusion in StEPS. A second objective, however, was to determine if in stratified-sample designs the use of the stratified jackknife could replace the use of sampling-theory formulas, which would be more difficult to program compared to the stratified jackknife. As expected, Tremblay and Sigman found that for linear estimates, the two programs/methods yielded identical results. They found, however, that "[f]or separate-ratio estimation, the larger absolute bias of SUDAAN and the larger variance of VPLX tend to balance each other when one considers the root mean square errors of calculated standard errors." Tremblay and Sigman concluded that the choice between SUDAAN or VPLX was not obvious in terms of studied statistical properties. They recommended that StEPS use VPLX, however, for the following reasons: "VPLX is more flexible: basically, anything that can be set up in a formula can be done in VPLX. ... VPLX is 'license free', and consulting is more readily available since its developer/maintainer is resident at the Census Bureau."

Tremblay (1996) extended Tremblay and Sigman (1996) by using VPLX to additionally calculate variances using the method of random groups with 16 groups. She compared the SUDAAN results (i.e., S^2 formulas with Taylor-series approximations), the VPLX stratified-jackknife results, and the VPLX-random-group results for two different FRIS survey variables and found that in nearly all cases the estimated relative root-mean-square errors of the VPLX-random-group estimated standard errors were larger than both those from SUDAAN and those from the VPLX stratified jackknife. This difference was particularly pronounced for the case of aggregated multi-stratum estimates.

The Economic Directorate uses the method of random groups to estimate variances for its monthly and annual surveys of retail and wholesale trade and for its annual surveys of other service industries. Rust (1985) states the "[t]he random groups method is most useful in surveys using a large number of PSUs, where either many PSUs are selected per stratum, or few gains are believed to result from the finer levels of stratification." The surveys for which the Economic Directorate uses the random groups method to estimate sampling variances satisfy these requirements. Town (1997) used the random groups option of VPLX to calculate sampling variances for two surveys: FRIS and the Annual Trade Survey (ATS) for wholesale trade. She found that the number of lines of VPLX code that were required to calculate variances for these two surveys were very different. For FRIS only 77 lines of VPLX code were required, whereas for ATS 3591 lines of VPLX code were needed. Some of the reasons for this difference were the following:

- The VPLX program for ATS calculated variances for 17 different types of estimates: 7 type of unadjusted estimates and 10 types of census-adjusted estimates. These included raked and unraked estimates of level, ratios, percentages, trends of level estimates, and trends of ratios. The VPLX programs for FRIS, on the other hand, calculated variances for only one type of estimate: census-adjusted estimates of level.

- The VPLX program for ATS calculated variances for 22,309 estimates; whereas the VPLX program for FRIS calculated variances for only 276 estimates.

- The VPLX program for ATS calculated a number of derived items, whereas the VPLX program for FRIS did not.

At the time of Town's investigation some enhancements to the syntax rules for VPLX code had been completed and more were planned for the future. Dr. Robert Fay, the developer of VPLX, recoded a portion of Town's VPLX program for ATS using the proposed enhancements to the VPLX syntax rules. Using the proposed syntax rules, the VPLX code was 152 lines; using the proposed new syntax rules, however, it was only 31 lines. From her two comparisons--FRIS versus ATS and "old" syntax versus "new" syntax--Town concluded that VPLX could be used by StEPS to calculate variances for surveys that use the random groups method, but the implementation effort of using VPLX for surveys like ATS appeared to be quite high. Town recommended that the new syntax rules for VPLX be used when they become available and that StEPS developers also "investigate alternative software approaches, such as the calculation by StEPS of random-group-level estimates followed by variance estimation using the method of random groups performed by a general-purpose SAS macro."

Dajani (1999) further explored the random group method of variance estimation in order to make recommendations on how it should be used in StEPS. Dajani studied the problem of how the method of random groups should be used to estimate variances for aggregate estimates following the estimation of variances for more detailed estimates. This same problem was studied by Kott (1999) in the context of using the delete-a-group jackknife to estimate sampling variances. In Kott's study the aggregate estimates were national-level estimates, and the detailed estimates were state-level estimates; whereas in Dajani's study the aggregate estimates were for two- and three-digit Standard Industrial Codes (SICs) and the detailed estimates were for four-digit SICs. One approach to estimating the variances for the aggregate estimates is to use the same replication method for the aggregate estimates as was used for the detailed estimates. A second approach, labeled the *hybrid method* by Kott, is for linear estimators to sum the variances of the detailed estimates to the aggregate level. For non-linear estimators one estimates the variance of a first-order Taylor-series approximation to the aggregate estimator, which is a linear combination of variances and covariances of the aggregate totals, calculated by summation of the variances and covariances of detail totals.

Like Town, Dajani studied ATS; but Town used 1995 ATS data (the sample for which was selected in 1990), whereas Dajani used 1997 ATS data (the sample for which selected in 1995). Dajani compared the two different approaches for estimating the variances of aggregate estimates for ATS aggregate totals, ratios of aggregate totals, and trends of

aggregate totals. Dajani found that the resulting differences in the two approaches for calculating variance estimates for all ATS aggregate totals, all trends of aggregate totals, and approximately 87 percent of the ratios of aggregate totals were not statistically significant. Since the use of replication to calculate the variances of aggregate estimates is easier to program (because covariances do not have to be calculated), Dajani recommended that replication "be used to calculate variances for aggregate estimates in ATS and in other surveys that have a similar sample design."

Though the Census Bureau's economic surveys are primarily single-stage surveys, the Census Bureaus's Survey of Construction (SOC) is a multi-stage, multi-frame survey. The SOC reporting unit is a construction project, not an establishment selected from the Census Bureau's business register. Thompson (1998) and Thompson and Sigman (1998) investigated the use of modified half-sample (MHS) replication to estimate variances for SOC. They recommended that StEPS use VPLX to calculate MHS variance estimates for SOC instead of using legacy code that calculated variance estimates with sampling-theory formulas.

Based on the research studies described above, the Economic Directorate decided in 1997 to calculate sampling variances in StEPS using the following "two methods" approach (Sigman 1997):

- For surveys with single-stage Poisson-sampling designs, use appropriate sampling-theory formulas (see Särndal 1996), and

- For all other survey designs, use one of the replication options of VPLX.

Two recent developments, however, have caused the Economic Directorate to abandon this two-method approach. One of these developments was that following the migration of three surveys into StEPS in 1998 we realized that the implementation effort to use VPLX to calculate random group variances for eleven surveys in 1999 would be very high. The second development was the decision by survey designers of several surveys that in the past had used Poisson sampling to instead use Tillé sampling (Tillé 1996, Slanta 1999). As a result of these two developments, we replaced the two-method approach with the following four-method approach:

- For surveys with single-stage Poisson-sampling designs, use appropriate sampling-theory formulas.;

- For surveys with single-stage Tillé-sampling designs, use sampling-theory formulas described in Tillé (1996) and Slanta (1999);

- For all other survey designs that use the random group method to calculate sampling variances, use SAS macros %rg_var1 and %rg_var2, contained in StEPS (described in section 5.3); and

- For all surveys that do not use the random group method to calculate sampling variances, use a replication option in VPLX.

In section 8 we discuss areas of future research, the findings from which may result in additional changes to the StEPS list of methods for calculating sampling variances.

4. Overview of StEPS

StEPS is a generalized survey processing system that the Economic Directorate has developed to replace 16 legacy systems. In addition to reducing resources needed for system maintenance, one of the StEPS objectives is to shift more processing control to survey analysts. StEPS contains integrated modules for data-collection support (e.g., mail-label printing and questionnaire check-in); editing; data review and correction; imputation; calculation of estimates and variances; and system administration (e.g., parameter specification and the submission and monitoring of batch jobs). Functions not in StEPS include: frame development, sample selection, actual data collection, and dissemination.

StEPS is programmed in SAS, and it stores data and parameters in SAS data sets. The Economic Directorate executes StEPS mainly on Compaq® Alpha® machines using UNIX as the operating system. Most users access StEPS via a graphical (X-Windows) communication package loaded on their desktop microcomputer.

Ahmed and Tasky (1999, 2000) provide additional information StEPS. Tasky et al. (1999) describe the StEPS system design and associated programming strategies. In particular, they state that the developers of StEPS "decided on four major design concepts:

1) "Design a set of standard data structures that remain the same, regardless of the survey and the data.
2) "Use parameters (stored in general data structures) to drive the survey-specific processing requirements.
3) "Generate a 'fat' record data set on the fly for certain modules ... .
4) "Standardize field names and possible value for similar concepts."

The next section describes how these design concepts were implemented in the StEPS Estimates and Variances Module.

5. Components of the StEPS Estimates and Variances Module

In a large survey organization, the specification of survey processing operations requires inputs from multiple specialists. This is especially true when specifying the calculation of estimates and sampling errors. Survey analysts know WHAT estimates to calculate with WHAT data. The sample designer knows HOW to calculate the estimates and sampling errors. Thus, one of the functional requirements for the StEPS Estimates and Variances Module was that it permit specification of both WHAT and HOW information. A second functional requirement was that it be able to calculate estimates and variances for many different surveys. A third functional requirement was that the Estimates and Variances Module must be integrated with the other StEPS modules–i.e., it should, where possible, use data sets used by other modules; and its interactive screens should have a similar "look and feel" as those for other StEPS modules.

Like other StEPS modules, the following are the major components of the StEPS Estimates and Variances Module:

- Standard data set structures for micro data, macro data, and processing parameters;
- Interactive screens for specifying parameters, submitting batch jobs, and requesting results listings; and
- SAS macros and scripts for batch calculations.

Each of these is discussed below.

5.1. Standard data set structures

StEPS stores micro data in *control files* and *item files*. Micro data includes data associated with questionnaire items; data associated with survey operations such as sample selection, mailing, collection, or check-in; or auxiliary data available from censuses or administrative sources. The item file can contain only numeric micro data, whereas the control file can contain numeric and character data. Another difference between the control file and the item file is that the control file has a "fat" format, whereas the item file has "skinny" format. In the control file (i.e., fat format) there is one record per reporting unit (ID), and the fields within each record correspond to control-file variables. In the item file (i.e., skinny format) there is one record per ID/item combination, and fields within each record correspond to different *data versions* (plus there is a field containing a data flag).

StEPS stores the following data versions in each record of the item file:

r_{ij} = reported data for item i and reporting unit j
e_{ij} = edited data for item it and reporting unit j
a_{ij} = adjusted data for item it and reporting unit j
w_{ij} = weighted-adjusted data for item it and reporting unit j

The default value for edited data is $e_{ij} = r_{ij}$. StEPS users, however, may change edited data by using the Review and Correction Module, or StEPS can change edited data via the Imputation Module.

Some surveys adjust micro data for data collection effects, such as trading day effects in monthly surveys or in annual surveys the effect on reported inventories of ending inventory dates other than December 31. One way that StEPS adjusts micro data is

$$a_{ij} = f(t_i, \boldsymbol{B}_j)\, e_{ij} ,$$

where

t_i = the value for item i of a variable, called *adjustment type* stored on the *item data dictionary* file;

$\boldsymbol{B}_j$ = a vector of *BY variables* --i.e., categorical variables--associated with reporting unit j; and

f( ) = a SAS format that StEPS creates to map the vector $(t_i, \boldsymbol{B}_j)$ into user-provide adjustment factors.

Another way StEPS adjusts micro data is to use user-provided SAS code stored in the *adjust/derive definitions file*. Many surveys do not adjust their micro data, however, in which case $a_{ij} = e_{ij}$.

StEPS calculates weighted-adjusted data using the following formula:

$w_{ij} = \omega_j \, g_{n(i),j} \, a_{ij}$.

The quantity ω_j is the sampling weight for reporting unit j. The control file stores three *g weights*, g_{1i} , g_{2i} , and g_{3i} , for each reporting unit. We had planned to use the g-weights in the manner described in Estavao, et al. (1995), in which if they are chosen appropriately the resulting weighted totals (or weighted means) are generalized regression estimators. To date, we have not used the g-weights for this purpose. One way we have used the g-weights was in our annual retail trade survey, which collects some items for only a subsample of the survey, we let ω_j store the first-phase sampling weight and let the g-weight store the second-phase weight. In the future we plan to use the g-weights to store non-response adjustment factors for surveys that use weight adjustment to handle unit nonresponse. The g-weight is equal to 1.0 for unweighted and Horvitz-Thompson estimators. The quantity *n(i)* is the *g-weight number* and indicates which g-weight, g_{1i} , g_{2i} , or g_{3i} , is associated with item i. If *n(i)*=0 then item i has a g-weight of 1.0. The g-weight number, like the adjustment type, t_i, is stored in the item data dictionary, which contains one record for each item-data variable.

The item file's skinny format can be difficult to use for estimation and variance calculations. Consequently, StEPS can create an *estimation fat file*, which has one record per ID, and the fields within each record can be any of the following: control file variable; adjusted or weighted-adjusted version of an item file variable; constant data; or *recode*, which is a variable created at the time of fat-file creation via a user-provided SAS expression involving other fat-file variables. When StEPS creates an estimation fat file, a variable on the control file, called the *weighting switch,* selects for each ID the adjusted or weighted-adjusted version of the item file variables. Certain values of the weighting switch zero out item data in the fat file or delete an entire record from the estimation fat file. By setting the weighting switch to a particular value for each ID, one can control the contents of each estimation-fat-file record, for purposes such as handling deaths by zeroing out or deleting data or handling outliers by deleting or down-weighting to self-representing.

StEPS stores macro data in *estimation results files* (ERFs). One ERF corresponds to one *table*, which is the result of StEPS performing calculations on *analysis variables* for individual values of categorical *BY variables*. The types of results StEPS stores in ERFs include: totals, ratios, trends, other derived estimates (i.e., functions of totals), standard errors, CVs, covariances, t-tests, imputation rates and disclosure-avoidance information. The ERF has a skinny format--each ERF record contains only one calculated result, with other variables in the record identifying the type of result, the name(s) of the analysis variable(s), and the value(s) of any BY variable(s).

Two files store estimation processing information: the *estimation specification file* (ESF) and the *estimation formulas file* (EFF). The ESF stores parameters used by the SAS macros described in section 5.3; the EFF stores SAS expressions and SAS code, also used by the SAS estimation macros. Both the ESF and EFF are populated via interactive screens. Developing a file layout for the ESF was challenging. We rejected a skinny-record format of one record per parameter because of the complexity of file updating from screens displaying multiple parameters. Instead, we decided the ESF would have one record per *specification*, which is a vector of parameters displayed together on the same screen. In the ESF, sets of specifications (i.e., records) associated with the same type of screen and processing action are called *objects*. For example, the "BY object" contains specifications for BY variables, whereas the derived object contains specifications for the calculation of derived estimates.

5.2. Interactive screens

Interactive screen in the Estimates and Variances Module allow StEPS users to do the following:

- Calculate weighted data for all items and IDs in the item file.

- Run Quicktab program, which calculates weighted totals, year-to-year trends, imputation rates, unweighted counts, and disclosure-avoidance information. The Quicktab program requires analysis variables to be item file variables and any BY variables to be control file variables. Quicktab does not calculate standard errors, CVs, or derived estimates. The possible outputs from Quicktab are a SAS data set, an ASCII file (for downloading), printer output, or the SAS Output Window.

- Enter and modify specifications and formulas for use by batch jobs. Specifications and formulas tell StEPS WHAT to estimate with WHAT data. The StEPS user can select analysis and BY variables (from item data, control data, recodes, or constants); specify the method of calculating standard errors (random groups, VPLX replication, or formulas for Poisson or Tillé samplig); enter formulas for derived estimates and the derivatives of non-linear estimators); copy results from one ERF to another ERF; and remove results from an ERF.

- Submit estimation scripts to run in batch. Scripts are described in more detail in section 5.3. A screen displays the available scripts, and the user selects one of the displayed scripts to run immediately or at a scheduled time.

- Review estimation results. A screen displays a list of ERFs, and the user can select an ERF for interactive viewing with SAS/FSVIEW® or for formatting by StEPS into a printed listing.

5.3. SAS macros and scripts

StEPS scripts execute SAS code that is part of StEPS or has been generated by StEPS. For the Estimates and Variances Module, scripts execute SAS code that is part of StEPS. In particular, estimation scripts execute one or more of the following SAS macros:

%extract — Creates estimate fat file.
%totals — Calculates totals and imputation rates.
%derive — Calculates derived estimates.
%erfrmt — Reformats an ERF.
%rtsumvar — Aggregates totals, standard errors, and imputation rates.
%copy1 — Copies results between ERFs.
%remove — Removes results from an ERF.
%round — Rounds totals and standard errors
%vpl2stp — Stores VPLX-calculated estimates and standard errors in an ERF (Dajani 1999a).\
%rgvar_1 — Calculates replicate totals from random group totals.
%rgvar_2 — Calculates replicate-based standard errors from replicate estimates.
%vrncs_p — Calculates standard errors for Poisson-sampling designs.
%vrncs_t — Calculates standard errors for Tillé sampling designs.
%cvrncs_t — Calculates covariances for Tillé sampling designs.
%taylor — Calculates standard errors of non-linar estimates using Taylor approximation.

Many of these macros are individually controlled by parameters analysts have entered into the ESF and EFF. Parameters specify WHAT to estimate and WHAT data to use. The estimation script controls the overall logic of HOW to calculate estimates and variances. Because this depends on the sample design, surveys with different sample designs require different scripts. Also, the sample designer should be involved in developing an estimation script--either as an advisor or as the person who produces the script.

6. Examples of estimation scripts

StEPS has two type of scripts: generic scripts and complete scripts. A *generic script* is a SAS program that executes SAS code contained in StEPS or generated by StEPS. In a generic script, macro variables are used to refer to the survey, statistical period, and other job submission conditions. StEPS users (or StEPS implementation staff) prepare generic scripts using the SAS Editor or other word processing package. A *complete script*, on the other hand, is created by StEPS, and it links to a corresponding generic script (via a %include statement). A complete script defines the macro variables that are used in the corresponding generic script.

Example 1: Create fat file, calculate totals and derived estimates, and calculate CVs using method of random groups.

Generic script:

```
01   *sc_no=F001  sc_desc=Example1;
02    %include '/steps/central/autocall.sas';
03    %setlibs(survey=&survey,statp00=&statp00);
04    %getstime;
05    %let ntbles=1; %let table1=F401;
06    %extract;
07    %totals(rg=y);
08    %rg_var1(intab=F401, outerf=F401);
09    %derive;
10    %rg_var2(inerf=F401,outerf=F401);
11     %applog(module=estimate, submod=&sc_no,
                  starttme=&startme,prgnme=&sc_no,
                  otherinfo=&sc_no);
```

Discussion: Since **line 1** begins with an "*", it is a SAS comment and is ignored by the batch processing. The information in line 1 is used, however, by the interactive script-submission screen to identify the script number (F001) and the script description ("Example 1"). This information appears on the script-submission screen in the list of scripts available for submission. **Line 2** makes StEPS SAS code available to the batch program. **Line 3** creates all needed SAS LIBNAMEs and UNIX environment variables. **Line 4** puts the starting time into the macro variable &startme. **Lines 5 and 6** create the estimation fat file needed to calculate totals. **Line 7** calculates random group totals and stores the results in ERF F401. **Line 8** converts the random group totals in ERF F401 into replicate totals. **Line 9** updates ERF F401 with derived estimates calculated for each replicate. **Line 10** calculates standard errors from the replicate estimates and stores them in ERF F401 along with the full-sample estimates and the associated CVs. **Line 11** puts information in the production log about the completed batch job.

Example 2 : Calculate totals and imputation rates for XSALES00 and XESALE00 with BY1=state and BY2=NAICS6 (i.e. six digit NAICS code) and store in ERF F301. File ERF F302 contains census totals CSALES with BY1=NAICS2 (i.e., two digit NAICS code). Adjust XESALES00(state,NAICS6) by multiplying by

F(NAICS2) = CSALES(NAICS2) / XSALES00(NAICS2) .

Generic script:

```
01-04 ... (comment, %setlibs, %getstime)
05    %let ntbles=2;
06    %let table1=F301; %let table2=F302;
07    %extract;
08    %let ntbles=1;
09    %totals(imprate=y);
10    %rtsumvar(intn=F301, outn=F302);
11    %let table1=F302;
12    %derive; ** Calculate F(NAICS2) **;
13    %erfrmt(intn=F302, outn=F301,
                  incndtn=%str(ITEM EQ F));
14    %let table1=F301;
15    %derive; **Calculate adj XESALE00 **;
16    .... (%applog)
```

Discussion: **Lines 1 through 4** are the same as Example 1. **Lines 5. 6 and 7** create an estimation fat file containing all the variables needed to calculate totals (in line 9), aggregate results (in line 10), and reformat an ERF (in line 13). **Lines 8 and 9** calculate totals and imputation rates for ERF F301 with BY1=state and BY2=NAICS6. **Line 10** aggregates results in ERF F301 and puts the aggregated results in ERF F302, with BY1=NAICS2. **Lines 11 and 12** calculate the adjustment factors and stores them in ERF F302. **Line 13** reformats the adjustment factors in ERF F302 into the structure of ERF F301, where the reformatted adjustment factors are stored with BY1=state and BY2=NAICS6. **Lines 14 and 15** calculate adjusted XESALE00 values and updates ERF F301 with these results.

7. Implementation Experiences

In 1998 the Economic Directorate used StEPS for production processing of three annual surveys. The largest of these was the (wholesale) Annual Trade Survey (ATS), with a stratified sample of approximately 7000 reporting units. The other two surveys were small industrial-product surveys; one was a Poisson sample with less that 600 reporting units

and the other was a cut-off sample with less that 200 reporting units. In 1998 interactive screens for entering estimation specifications had not yet been developed. Thus, development staff typed estimation parameters into fixed-field ASCII files for the three surveys. This was tedious and error prone, which motivated the development of interactive screens that were used in 1999. In 1998 the StEPS developers created the estimation script files for these three surveys.

In 1998 we used VPLX to calculate variances for ATS using the method of random groups. The length of the VPLX program used to calculate these variances was 513 lines. Since in 1999 there would be an additional ten StEPS surveys using random groups to calculate variances, we concluded that the implementation effort involved in continuing to use VPLX for these surveys was unacceptably high. Hence, we decided to develop the StEPS macros %rg_var1, %rg_var2, %rtsumvar, and %erfrmt for calculating random group variances in 1999.

The production of variances for ATS was improved considerably in 1998 over what was possible from the legacy system. ATS calculates census-adjusted estimates, and the legacy system incorrectly treated the adjustment factors as constants. StEPS, however, was able to correctly calculate the variances of the ATS census-adjusted estimates by treating them as ratio estimates. Another improvement over the legacy system was that the interactive script-submission capability of StEPS permitted survey analysts to obtain estimates and variances upon demand, which was not possible from the legacy system.

In 1999 the Economic Directorate used StEPS for the production processing of fifty surveys. Eleven of these surveys were service-sector surveys that used random groups to calculate variances. These surveys ranged in size from as small as 4,000 reporting units to as large as 27,000 reporting units. Two of the surveys processed by StEPS in 1999 used Poisson sampling; and one survey, the Manufacturing Energy Consumption Survey, used Tillé sampling. The remaining 30+ surveys were industrial-products surveys that used cut-off sampling.

In 1999, survey analysts for the eleven service-sector surveys and for one of the Poisson-sample surveys used interactive screens to enter estimation specifications into StEPS. As in 1998, however, estimation scripts were for the most part created by StEPS development staff. In 1999, we did not use VPLX to calculate random-group variances–instead, we used StEPS macros to calculate random group variances for the eleven service-sector surveys. For the Poisson-sample and Tillé-sample surveys, we will use the StEPS macros %vrncs_p and %vrncs_t, respectively, to calculate variances. For the 30+ surveys that were cut-off samples, analysts used Quicktab to calculate estimates, so it was not necessary for scripts to be written or for estimation specifications to be entered into StEPS.

8. Future Activities

The Economic Directorate plans to migrate additional surveys into StEPS. One of these is the Survey of Construction (SOC), which will use the Estimates and Variances Module of StEPS (and not the other modules in StEPS). Because SOC is a multi-stage survey, we will use VPLX to calculate SOC variances. Another survey migrating into StEPS is the Annual Capital Expenditures Survey (ACES), which has a stratified sample design. We plan to investigate the stratified jackknife for estimating variances for ACES.

The use of standard data structures in StEPS facilitates comparative methodological research. We plan on comparing replication-based variances with those calculated using the sampling-theory formulas for Poisson and Tillé sampling. Another study we plan on conducting will compare random group variances to those from a delete-a-group jackknife.

Finally, we observe that the development of a survey processing system is a journey, not a destination. Lessons learned from today's processing suggest enhancements to the system to perform tomorrow's processing. One possible enhancement to the Estimates and Variances Module is to provide a graphical user interface for developing estimation scripts. Another possible enhancement is to interface estimation results files from StEPS to online analysis tools such as SAS/EIS® and SAS/INSIGHT®.

References

Ahmed, S. and Tasky, D. (1999), "The Standard Economic Processing System: A Generalized Integrated System for Survey Processing, " *Proceedings of the Survey Research Methods Section*, Alexandria VA: American Statistical Association, to appear.

Ahmed, S. and Tasky, D. (2000), "Standardized Economic Processing System," *Proceedings of the International Conference on Establishment Surveys*, Alexandria, VA: American Statistical Association, to appear.

Dajani, A. (1999a), *Version 1.0 of %vpl2stp: A SAS macro for Creating SAS Data Sets from VPLX Display-step Output*, Technical Report #ESM-9901, Washington DC: Bureau of the Census.

_____ (1999b), *Comparison of Variance Estimation Methods for Aggregate Estimates*, Technical Report #ESM-9902, Washington, DC: Bureau of the Census.

Estavao, V.; Hidiroglou, M; and Särndal, C. (1995), "Methodological Principles for Generalized Estimation System at Statistics Canada," *Journal of Official Statistics*, **11**, pp. 181-204.

Fay, R.E. (1990), "VPLX: Variance Estimates for Complex Samples," *Proceedings of the Survey Research Methods Section*, Alexandria VA: American Statistical Association, pp. 266-271.

King, C. and Kornbau, M. (1994), *Inventory of Economic Area Statistical Practice, Phase 2: Editing, Imputation, Estimation, and Variance Estimation*, Technical Report #ESMD-9401, Washington DC: Bureau of the Census, March 1994.

Kott, P. (1999), "Some Problems and Solutions with a Delete-A-Group Jackknife", *Proceedings of the Federal Committee on Statistical Methodology Research Conference*, Tuesday B Sessions, Washington, DC: Council of Professional Associations on Federal Statistics, pp. 129-135.

Rust, K. (1985), "Variance Estimation for Complex Estimators in Sample Surveys," *Journal of Official Statistics*, **1**, pp. 381-397.

Särndal, C. (1996), "Efficient Estimators with Simple Variance in Unequal Probability Sampling," *Journal of the American Statistical Association*, **91**, pp. 1289-1300.

Sigman, R. (1997), "How Should We Proceed to Develop Generalized Software for Survey Processing Operations Such as Editing, Imputation, etc?", unpublished paper presented at the Meeting of Census Advisory Committee of Professional Associations, Washington DC: U.S. Bureau of the Census, May 1-2, 1997.

Slanta, J. (1999), "Implementation of Modified Tillé Sampling Procedure in the MECS and R&D Surveys," *Proceedings of the Survey Research Section*, Alexandria, VA: American Statistical Association, to appear.

Tasky, D.; Linonis, A.; Ankers, S; Hallam, D., Altmayer, L.; and Chew, D. (1999), "Get in Step with StEPS: Standard Economic Processing System," *Proceedings of the North East SAS Users Group*, pp. 167-178.

Thompson, K. (1998), *Evaluation of Modified Half-Sample Replication for Estimating Variances for the Survey of Construction (SOC)*, Technical Report ESM-9801, Washington DC: Bureau of the Census.

_____ and Sigman, R. (1998), "Modified Half Sample Variance Estimation for Median Sales Prices of Sold Houses: Effects of Data Grouping Methods," *Proceedings of the Survey Research Methods Section*, Alexandria, VA: American Statistical Association, pp. 698-703.

Tillé, Y (1996), "An Elimination Procedure for Unequal Probability Sampling Without Replacement," *Biometika,* **83**, pp. 238-241.

Town, G. (1997), *The Use of VPLX to Calculate Variances for Economic Surveys Using the Method of Random Groups: A Tale of Two Surveys*, Technical Report #ESM-9701, Washington DC: Bureau of the Census.

Tremblay, T. (1996), "Estimates from VPLX Random Groups Option," unpublished memorandum, Washington DC: Bureau of the Census, September 26, 1996.

_____ and Sigman, R. (1996), "Comparison of Two Variance-Estimation Methods for a Standardized Economic Processing System," *Proceedings of the Survey Research Methods Section*, Alexandria VA: American Statistical Association, pp. 204-208.

Wolter, K. (1985), *Introduction to Variance Estimation*, New-York: Springer-Verlag.

GENERALISED ESTIMATION SYSTEM AND FUTURE ENHANCEMENTS

Michel A. Hidiroglou, Victor M. Estevao and Charlie Arcaro, Statistics Canada
Michael A. Hidiroglou, R.H. Coats Building 11th floor, Ottawa, Ontario, K1A 0T6, Canada
hidirog@statcan.ca

ABSTRACT

This paper presents an overview of the methodology of the Generalised Estimation System (GES) developed at Statistics Canada and some plans for future enhancements. The GES uses auxiliary information to produce domain estimates for one-stage and two-phase designs. The methodology is based on the generalised regression (GREG) estimator and its extension to the wider family of calibration estimators. Most estimators in survey practice, including the post-stratified and raking ratio estimators are members of this family. GES produces estimates of totals, means and ratios over any domains of interest. Variance estimation in GES is done using either the design-based Taylor method or the jack-knife procedure. The calculation of the Taylor variance uses a linearised residual or z-score obtained from the first-order approximation of the functions of totals. We are planning to extend this approach to handle more complex non-linear functions of totals and to allow for tests of hypothesis. The idea is to eventually extend GES as a data analysis package.

Key Words: Calibration; Auxiliary Variables; One-Stage Sampling; Two-Phase Sampling

1. INTRODUCTION

Estimation procedures at Statistics Canada have seen a growing use of auxiliary data for a variety of sampling designs. The need to automate estimation procedures to take advantage of the increased availability of auxiliary data was recognised in the mid-eighties. The rationale for the development of generalised systems is described in Outrata and Chinnappa (1989). As attested by practice, generalised estimation software has several advantages over customised software. These include reduced maintenance costs, a unified methodology and a single systems architecture. It has given methodologists the flexibility to try out different estimation procedures and the capability of including new system and methodological advances.

Several estimation packages have been developed elsewhere using slightly different approaches for the methodological framework. These include LINWEIGHT (Bethlehem and Keller 1987), PC-CARP (Schnell et al. 1988), SUDAAN (Shah et al. 1989), CLAN (Andersson and Norberg 1994) and WESVAR (Brick et al. 1997. These packages have several features in common with respect to the sampling designs that they accommodate and the available parameters for estimation. For instance, they all handle stratified clustered sampling designs with and without replacement sampling. They all provide methods for the estimation of population totals, means and ratios. The packages differ on the availability of analytic features (regression, and two-way table analysis) and the procedures used for variance estimation (design-based variance of the Taylor approximation, or replication methods such as the jack-knife and balanced repeated replication).

GES provides estimation procedures for cross-sectional surveys. The framework adopted for GES is based on the use of auxiliary information for calibration. A detailed description of the methodology used in GES is given in Estevao, Hidiroglou and Särndal (1995). GES is built around four concepts: (1) the sampling design; (2) the auxiliary information; (3) the domains of interest; and (4) the population parameters for estimation. The methodology for one-stage element and cluster designs is described in sections 2 and 3 respectively. GES has recently been extended to include two-phase estimation with auxiliary information at each phase. This is presented in section 4. In section 5, we briefly mention the general direction of future developments in GES.

2. ONE-STAGE ELEMENT SAMPLING DESIGNS

The set up for one-stage element sampling designs is as follows. Let the population of elements be given by $U = \{1,...,k,...,N\}$. A probability sample s is selected from U using either sampling without replacement (WOR) or probability proportional to size with replacement (PPSWR). In WOR sampling, s is an unordered sample and element k has inclusion probability π_k. The sample design weights are simply $w_k = 1/\pi_k$ for $k \in U$. In PPSWR, we consider s as an ordered sample of n fixed draws with replacement and the element corresponding to selection k is assigned the design weight $w_k = 1/(n\, p_k)$ where p_k is the probability of selecting the element on each draw. Let $Y = \Sigma_{k \in U} y_k$ be the population total for a variable of interest y. A design-based estimator is the Horvitz-Thompson estimator $\hat{Y}_{HT} = \Sigma_{k \in s} w_k y_k$. It is design unbiased but usually not very efficient. We obtain more efficient estimators by using the available auxiliary information.

2.1. Auxiliary Information and Calibration

Suppose we know $\{\boldsymbol{x}_k\}$ for $k \in s$ and $\boldsymbol{X} = \Sigma_{k\in U}\boldsymbol{x}_k$, where $\boldsymbol{x}_k' = (x_{1k},...,x_{jk},...x_{Jk})$ is a vector of auxiliary data for element *k*. This auxiliary information can be used to produce calibrated weights $\tilde{w}_k$ using the calibration approach proposed by Deville and Särndal (1992) or Huang and Fuller (1978). Deville and Särndal obtain these weights by minimising a distance function between $\tilde{w}_k$ and the design weights w_k subject to the restriction that $\Sigma_{k\in s}\,\tilde{w}_k\boldsymbol{x}_k = \Sigma_{k\in U}\,\boldsymbol{x}_k$. The Deville-Särndal procedure has been implemented in GES using the least squares distance function $\Sigma_{k\in s}\,c_k(\tilde{w}_k - w_k)^2/2w_k$. This distance function is minimised with respect to the $\tilde{w}_k$, subject to the calibration equation

$$\Sigma_{k\in s}\tilde{w}_k\boldsymbol{x}_k = \Sigma_{k\in U}\boldsymbol{x}_k \tag{2.1}$$

and $\ell < \tilde{w}_k < u$. Here ℓ and u are reasonable lower and upper bounds chosen to prevent negative weights or extremely large weights. The positive values c_k allow weighting of the individual terms of the distance function. If there are no bounds $(\ell = -\infty,\, u = +\infty)$ on the calibrated weights, then the solution to the above problem is $\tilde{w}_k = w_k\{1 + (\Sigma_{k\in U}\boldsymbol{x}_k - \Sigma_{k\in s}w_k\boldsymbol{x}_k)'(\Sigma_{k\in s}\,w_k\boldsymbol{x}_k\boldsymbol{x}_k'/c_k)^{-1}\,\boldsymbol{x}_k/c_k\}$. These are the weights of the GREG estimator $\hat{Y}_{GREG} = \hat{Y}_{HT} + \Sigma_{k\in s}w_k(y_k - \boldsymbol{x}_k'\hat{\boldsymbol{B}})$ with $\hat{\boldsymbol{B}} = (\Sigma_{k\in s}\,w_k\boldsymbol{x}_k\boldsymbol{x}_k'/c_k)^{-1}(\Sigma_{k\in s}\,w_k\boldsymbol{x}_k y_k/c_k)$. Many traditional estimators are obtained by assuming no bounds. For a single auxiliary variable, putting $c_k = x_k$ leads to the simple ratio estimator when $x_k > 0$. Table 1 shows how some well-known estimators can be obtained as calibration estimators for fixed sample size designs. There is no closed form expression for $\tilde{w}_k$ when bounds are part of the problem specification. The weights $\tilde{w}_k$ are calculated using the non-linear programming algorithm described by Estevao (1994). In either case, the calibration estimator is written as $\hat{Y}_{CAL} = \Sigma_{k\in s}\tilde{w}_k y_k = \Sigma_{k\in s}w_k g_k y_k$, where g_k is the calibration factor for element *k*.

Table 1: Horvitz-Thompson, Hàjek and Ratio estimators for fixed sample size designs.

Estimator	$\hat{Y}_{CAL}$	$\boldsymbol{x}_k$	c_k	g_k
Horvitz-Thompson	$\hat{Y}_{HT}$	π_k	π_k	1
Hàjek	$\frac{N}{\hat{N}_{HT}}\hat{Y}_{HT}$	1	1	$\frac{N}{\hat{N}_{HT}}$
Ratio	$\frac{X}{\hat{X}_{HT}}\hat{Y}_{HT}$	x_k	x_k	$\frac{X}{\hat{X}_{HT}}$

The calibrated weights $\tilde{w}_k$ should be close to the design weights w_k to minimise the bias and permit the estimation of variance of the calibration estimator. This is shown in the next section.

2.2. Bias and Variance Estimation

The bias and variance of $\hat{Y}_{CAL}$ are obtained by representing y_k as a linear function of $\boldsymbol{x}_k$ over the population *U*. We write

$$y_k = \boldsymbol{x}_k'\boldsymbol{B} + E_k \text{ for } k \in U \tag{2.2}$$

where $\boldsymbol{B}$ is defined as

$$\boldsymbol{B} = (\Sigma_{k\in U}\,\boldsymbol{x}_k\boldsymbol{x}_k'/c_k)^{-1}(\Sigma_{k\in U}\,\boldsymbol{x}_k y_k/c_k)\,. \tag{2.3}$$

It is important to note that this is a conceptual representation over the population. It does not have to provide a proper linear fit. We do not generally expect *y* to be linearly related to the auxiliary variables $\boldsymbol{x}$. Using (2.2) we

have $Y = \Sigma_{k\in U} y_k = \Sigma_{k\in U} \boldsymbol{x}_k' \boldsymbol{B} + \Sigma_{k\in U} E_k$ and $\hat{Y}_{CAL} = \Sigma_{k\in s} \tilde{w}_k \boldsymbol{x}_k' \boldsymbol{B} + \Sigma_{k\in s} \tilde{w}_k E_k$. From the calibration equation (2.1), we then obtain $\hat{Y}_{CAL} = \Sigma_{k\in U} \boldsymbol{x}_k' \boldsymbol{B} + \Sigma_{k\in s} \tilde{w}_k E_k$. The design bias of $\hat{Y}_{CAL}$ is $\mathrm{E}_s\{\Sigma_{k\in s} E_k(\tilde{w}_k - w_k)\}$ which can be written as

$$\mathrm{Bias}(\hat{Y}_{CAL}) = \Sigma_{k\in U} E_k \mathrm{E}_s(g_k - 1) \tag{2.4}$$

where $\mathrm{E}_s(g_k - 1)$ is the expected value of $(g_k - 1)$ over all samples containing element *k*. The bias will be small if g_k is close to 1 or equivalently, $\tilde{w}_k$ is close to w_k. This provides a rationale for the distance minimisation approach. Since $\Sigma_{k\in U} \boldsymbol{x}_k' \boldsymbol{B}$ is a constant, the variance of $\hat{Y}_{CAL}$ is simply the variance of $\Sigma_{k\in s} \tilde{w}_k E_k$. By assuming $g_k \doteq 1$ for $k \in s$ over all samples, we obtain

$$\mathrm{Var}(\hat{Y}_{CAL}) = \mathrm{Var}_s(\Sigma_{k\in s} \tilde{w}_k E_k) \doteq \mathrm{Var}_s(\Sigma_{k\in s} w_k E_k). \tag{2.5}$$

An estimate of $\mathrm{Var}(\hat{Y}_{CAL})$ can be obtained by replacing the E_k with the sample residuals $e_k = y_k - \boldsymbol{x}_k' \hat{\boldsymbol{B}}$ where $\hat{\boldsymbol{B}} = (\Sigma_{k\in s} w_k \boldsymbol{x}_k \boldsymbol{x}_k' / c_k)^{-1} (\Sigma_{k\in s} w_k \boldsymbol{x}_k y_k / c_k)$. We can also incorporate the calibration factors in the variance estimate by using $u_k = g_k e_k$ instead of e_k. For stratified WOR designs, an estimate of the variance is given by

$$\hat{\mathrm{V}}\mathrm{ar}_{WOR}(\hat{Y}_{CAL}) = \Sigma_{h=1}^{H} \Sigma_{k\in s_h} \Sigma_{l\in s_h} \frac{(\pi_{kl} - \pi_k \pi_l)}{\pi_{kl} \pi_k \pi_l} (u_k u_l) \tag{2.6}$$

where *H* is the number of strata and π_{kl} is the joint inclusion probability of elements *k* and *l* in stratum *h*. In stratified SRSWOR selection of n_h elements from the N_h in stratum *h*, this variance is $\hat{\mathrm{V}}\mathrm{ar}_{SRSWOR}(\hat{Y}_{CAL}) = \Sigma_{h=1}^{H} \{N_h^2 (1 - f_h) / (n_h (n_h - 1))\} \Sigma_{k\in s_h} (u_k - \bar{u}_h)^2$ with $\bar{u}_h = \Sigma_{k\in s_h} u_k / n_h$ and $f_h = n_h / N_h$.

For stratified PPSWR sampling, the variance of $\hat{Y}_{CAL}$ is estimated by

$$\hat{\mathrm{V}}\mathrm{ar}_{PPSWR}(\hat{Y}_{CAL}) = \Sigma_{h=1}^{H} \frac{n_h}{n_h - 1} \Sigma_{k\in s_h} \{(u_k / (n_h p_k)) - (\hat{U}_h / n_h)\}^2 \tag{2.7}$$

where $\hat{U}_h = \Sigma_{k\in s_h} u_k / (n_h p_k)$ and n_h is the number of selections with replacement in stratum *h*. In stratified SRSWR, $p_k = 1/N_h$ for $k \in s_h$ and the variance reduces to $\hat{\mathrm{V}}\mathrm{ar}_{SRSWR}(\hat{Y}_{CAL}) = \Sigma_{h=1}^{H} \{N_h^2 / (n_h (n_h - 1))\} \Sigma_{k\in s_h} (u_k - \bar{u}_h)^2$ with $\bar{u}_h = \Sigma_{k\in s_h} u_k / n_h$. The jack-knife analogue is $\hat{\mathrm{V}}\mathrm{ar}_{JK}(\hat{Y}_{CAL}) = \Sigma_{h=1}^{H} ((n_h - 1)/n_h) \Sigma_j (\hat{Y}_{CAL\,hj} - \hat{Y}_{CAL})^2$ where $\hat{Y}_{CAL\,hj}$ is the estimate of *Y* after deleting element *j* in stratum *h*. Hidiroglou (1991) notes that the similarity between $\hat{\mathrm{V}}\mathrm{ar}_{SRSWOR}(\hat{Y}_{CAL})$ and $\hat{\mathrm{V}}\mathrm{ar}_{SRSWR}(\hat{Y}_{CAL})$ suggests (2.6) can be approximated by $\Sigma_{h=1}^{H} \{(1 - f_h) n_h / (n_h - 1)\} \Sigma_{k\in s_h} \{(u_k / (n_h p_k)) - (\hat{U}_h / n_h)\}^2$ where f_h is an appropriate "sampling fraction" such as $n_h / (\sum_{k\in s_h} 1/\pi_k)$. Other approximations have been given by Rao (1963), Ardilly (1994) and Rosén (1991). GES currently computes variance estimates for stratified designs under SRSWOR and PPSWR. The approximations for without replacement PPS schemes are not yet implemented in GES.

2.3. Calibration Using Sub-Populations Totals

Calibration can be carried out within groups of any partition of the population with auxiliary information. We assume the population is partitioned into *P* mutually exclusive and exhaustive calibration groups $U_1 \ldots U_p \ldots U_P$ and that the following auxiliary information is available within each group:

(i) $\boldsymbol{x}_k' = (x_{1k}, \ldots, x_{jk}, \ldots x_{Jk})$ for $k \in s_p$ where s_p is the sample in U_p

(ii) $\boldsymbol{X}_p = \Sigma_{k\in U_p} \boldsymbol{x}_k$

The partition may be based on strata or post-strata. When there is only one group, the entire population, we obtain a combined estimator. Otherwise, we obtain a separate estimator over strata or post-strata.

2.4. Estimation of Ratios

Many parameters of interest can be expressed as functions of totals. One of the simplest functions is the ratio $R = Y_1/Y_2 = \Sigma_{k\in U} y_{1k} / \Sigma_{k\in U} y_{2k}$ of the totals of variables y_1 and y_2. We estimate this ratio by $\hat{R}_{CAL} = \hat{Y}_{1,CAL} / \hat{Y}_{2,CAL}$ where $\hat{Y}_{1,CAL} = \Sigma_{k\in s} \tilde{w}_k y_{1k}$ and $\hat{Y}_{2,CAL} = \Sigma_{k\in s} \tilde{w}_k y_{2k}$ are the estimated totals for y_1 and y_2 respectively. The mean or average of y is a special ratio with $y_{1k} = y_k$ and $y_{2k} = 1$ for $k \in U$. The estimated mean of y is given by $\hat{\bar{Y}}_{CAL} = \hat{Y}_{CAL} / \hat{N}_{CAL}$ where $\hat{N}_{CAL} = \Sigma_{k\in s} \tilde{w}_k$.

The estimated variance of $\hat{R}_{CAL} = \hat{Y}_{1,CAL} / \hat{Y}_{2,CAL}$ involves a linearisation of this ratio. We obtain a linearised statistic or z-score $z_k = (y_{1k} - \hat{R}_{CAL}\, y_{2k}) / \hat{Y}_{2,CAL}$, which we then substitute for e_k in the formulas in section 2.2. In fact, e_k can be viewed as the z-score for a total. A similar approach is used for the estimation of the variance of a mean.

2.5. Estimation for Domains

A domain U_d is an arbitrary subset of the population of elements *U*. Domains can cut across strata. In many surveys, the strata are domains of interest. It is important to note that for cluster designs, a domain can include elements from different clusters.

For a given domain U_d, the estimator of the domain total for variable *y* is $\hat{Y}_{(d)CAL} = \Sigma_{k\in s} \tilde{w}_k y_{(d)k}$ where $y_{(d)k}$ is equal to y_k if $k \in U_d$ and zero otherwise. A ratio within domain U_d is estimated by $\hat{R}_{(d)CAL} = \hat{Y}_{1,(d)CAL} / \hat{Y}_{2,(d)CAL}$ where y_1 and y_2 are the two variables of interest and $\hat{Y}_{i,(d)CAL} = \Sigma_{k\in s} \tilde{w}_k y_{i,(d)k}$ for $i = 1, 2$. Similarly, a domain mean is estimated by $\hat{\bar{Y}}_{(d)CAL} = \hat{Y}_{(d)CAL} / \hat{N}_{(d)CAL}$ with $\hat{N}_{(d)CAL} = \Sigma_{k\in s} \tilde{w}_k y_{2,(d)k}$ where $y_{2,(d)k}$ is equal to 1 if $k \in U_d$ and zero otherwise. The calibrated weights $\tilde{w}_k$ do not depend on the domain. They are computed within the calibration groups.

3. ONE-STAGE CLUSTER SAMPLING DESIGNS

For one-stage cluster sampling designs, the population of clusters is denoted by $U^{(C)} = \{1,..., i,..., N\}$ and the population of elements among the clusters is given by $U^{(E)}$. For simplicity, we consider only WOR sampling although the extension to PPSWR sampling can be done as in section 2.2. A sample $s^{(C)}$ is selected from $U^{(C)}$ with cluster *i* having inclusion probability π_i. All elements in the selected clusters form the sample of elements $s^{(E)}$. The cluster sample design weights are $w_i = 1/\pi_i$ for $k \in U^{(C)}$. In view of the design, the element sample design weights are $w_k = 1/\pi_k = 1/\pi_i = w_i$ for $k \in i$. Since there are two populations, auxiliary information may be known for clusters or elements. For the clusters, we may have a vector of auxiliary variables $\mathbf{z}_i$ for $i \in s^{(C)}$ and corresponding totals $\boldsymbol{Z} = \Sigma_{i\in U^{(C)}} \mathbf{z}_i$. Similarly, for the elements, we may have a vector of auxiliary variables $\boldsymbol{x}_k$ for $k \in s^{(E)}$ and corresponding totals $\boldsymbol{X} = \Sigma_{k\in U^{(E)}} \boldsymbol{x}_k$. This gives rise to calibration on either the cluster or element information and two different families of calibration estimators.

3.1. Estimators Based on Element Auxiliary Information

We use the least squares approach described in section 2.1 to obtain weights $\tilde{w}_k = w_k g_k$ for the elements in $s^{(E)}$ by minimising $\Sigma_{k\in s^{(E)}} c_k (\tilde{w}_k - w_k)^2 / 2w_k$ subject to the calibration equation

$$\Sigma_{k\in s^{(E)}} \tilde{w}_k \boldsymbol{x}_k = \Sigma_{k\in U^{(E)}} \boldsymbol{x}_k \tag{3.1}$$

where c_k is a positive coefficient and g_k is the calibration factor for $k \in s^{(E)}$. The total $Y = \Sigma_{k\in U} y_k$ is estimated by $\hat{Y}^{(E)}_{CAL} = \Sigma_{k\in s^{(E)}} \tilde{w}_k y_k$. We examine the properties of this estimator by considering the following linear representation.

$$y_k = \boldsymbol{x}_k' \boldsymbol{B}^{(E)} + E_k \text{ for } k \in U^{(E)} \tag{3.2}$$

where $\boldsymbol{B}^{(E)}$ is defined as

$$\boldsymbol{B}^{(E)} = (\Sigma_{k\in U^{(E)}} \boldsymbol{x}_k \boldsymbol{x}_k' / c_k)^{-1} (\Sigma_{k\in U^{(E)}} \boldsymbol{x}_k y_k / c_k) . \tag{3.3}$$

Using (3.2) and (3.1), we write $Y = \Sigma_{k\in U^{(E)}} y_k = \Sigma_{k\in U^{(E)}} \boldsymbol{x}_k' \boldsymbol{B}^{(E)} + \Sigma_{k\in U} E_k$ and $\hat{Y}_{CAL}^{(E)} = \Sigma_{k\in U^{(E)}} \boldsymbol{x}_k' \boldsymbol{B}^{(E)} + \Sigma_{k\in s^{(E)}} \widetilde{w}_k E_k$. Since $\Sigma_{k\in s^{(E)}} \widetilde{w}_k E_k = \Sigma_{i\in s^{(C)}} w_i \Sigma_{k\in i}\ g_k E_k$, it follows that the design bias of $\hat{Y}_{CAL}^{(E)}$ is

$$\text{Bias}\,(\hat{Y}_{CAL}^{(E)}) = \Sigma_{k\in U^{(E)}} E_k \,\text{E}_{s^{(C)}} \Sigma_{k\in i}(g_k - 1) \tag{3.4}$$

where $\text{E}_{s^{(C)}} \Sigma_{k\in i}(g_k - 1)$ is the expected value of $\Sigma_{k\in i}(g_k - 1)$ over all cluster samples containing cluster i (and the elements $k \in i$). If $g_k \doteq 1$ for $k \in i \in s^{(C)}$ over all cluster samples, we obtain the following approximation to the variance.

$$\begin{aligned} \text{Var}\,(\hat{Y}_{CAL}^{(E)}) &= \text{Var}(\Sigma_{i\in s^{(C)}} w_i \Sigma_{k\in i}\ g_k E_k) \\ &\doteq \Sigma_{i\in U^{(C)}} \Sigma_{j\in U^{(C)}} \frac{(\pi_{ij} - \pi_i \pi_j)}{\pi_i \pi_j} (\Sigma_{k\in i} E_k)(\Sigma_{k\in j} E_k) . \end{aligned} \tag{3.5}$$

An estimate of $\text{Var}\,(\hat{Y}_{CAL}^{(E)})$ which includes the calibration factors is given by,

$$\hat{\text{Var}}\,(\hat{Y}_{CAL}^{(E)}) = \Sigma_{i\in s^{(C)}} \Sigma_{j\in s^{(C)}} \frac{(\pi_{ij} - \pi_i \pi_j)}{\pi_{ij} \pi_i \pi_j} (\Sigma_{k\in i}\ g_k e_k)(\Sigma_{k\in j}\ g_k e_k) \tag{3.6}$$

where $e_k = y_k - \boldsymbol{x}_k' \hat{\boldsymbol{B}}^{(E)}$ and $\hat{\boldsymbol{B}}^{(E)} = (\Sigma_{k\in s^{(E)}}\ w_k \boldsymbol{x}_k \boldsymbol{x}_k' / c_k)^{-1} (\Sigma_{k\in s^{(E)}}\ w_k \boldsymbol{x}_k y_k / c_k)$. Under stratified SRSWOR selection of n_h clusters from the N_h in stratum h, we have $\hat{\text{Var}}_{SRSWOR}(\hat{Y}_{CAL}^{(E)}) = \Sigma_{h=1}^{H} \{N_h^2 (1 - f_h) / (n_h (n_h - 1))\}\, \Sigma_{i\in s_h^{(C)}} (u_i - \overline{u}_h)^2$ where $u_i = \Sigma_{k\in i} g_k e_k$, $\overline{u}_h = \Sigma_{i\in s_h^{(C)}} u_i / n_h$ and $f_h = n_h / N_h$.

3.2. Estimators Based on Cluster Auxiliary Information

We use a similar approach when we have auxiliary information for the clusters. We obtain weights $\widetilde{w}_i = w_i g_i$ for the clusters in $s^{(C)}$ by minimising $\Sigma_{i\in s^{(C)}} c_i (\widetilde{w}_i - w_i)^2 / 2 w_i$ subject to the calibration equation

$$\Sigma_{i\in s^{(C)}} \widetilde{w}_i \mathbf{z}_i = \Sigma_{i\in U^{(C)}} \mathbf{z}_i \tag{3.7}$$

where c_i is a positive coefficient for cluster i in the function and g_i is the calibration factor for $i \in s^{(C)}$. In this case, the total $Y = \Sigma_{k\in U} y_k$ is estimated by $\hat{Y}_{CAL}^{(C)} = \Sigma_{i\in s^{(C)}} \widetilde{w}_i \Sigma_{k\in i} y_k = \Sigma_{i\in s^{(C)}} \widetilde{w}_i y_{i\bullet}$ where $y_{i\bullet} = \Sigma_{k\in i} y_k$. The bias and variance of this estimator are obtained by considering the following linear representation.

$$y_{i\bullet} = \mathbf{z}_i' \boldsymbol{B}^{(C)} + E_i \text{ for } i \in U^{(C)} \tag{3.8}$$

where $\boldsymbol{B}^{(C)}$ is defined as

$$\boldsymbol{B}^{(C)} = (\Sigma_{i\in U^{(C)}} \mathbf{z}_i \mathbf{z}_i' / c_i)^{-1} (\Sigma_{i\in U^{(C)}} \mathbf{z}_i y_{i\bullet} / c_i) . \tag{3.9}$$

Using (3.8) and (3.7), we then write $\hat{Y}_{CAL}^{(C)} - Y = \Sigma_{i\in s^{(C)}} \widetilde{w}_i E_i - \Sigma_{i\in U^{(C)}} E_i$. It follows that the design bias is

$$\text{Bias}\,(\hat{Y}_{CAL}^{(C)}) = \Sigma_{i\in U^{(C)}} E_i \,\text{E}_{s^{(C)}} (g_i - 1) \tag{3.10}$$

where $\text{E}_{s^{(C)}} (g_i - 1)$ is the expected value of $(g_i - 1)$ over all cluster samples containing cluster i. Assuming $g_i \doteq 1$ for $i \in s^{(C)}$ over all samples, the approximate variance of $\hat{Y}_{CAL}^{(C)}$ is given by

$$\begin{aligned} \mathrm{Var}(\hat{Y}_{CAL}^{(C)}) &= \mathrm{Var}\,(\Sigma_{i\in s^{(C)}} w_i\ g_i E_i) \\ &\doteq \Sigma_{i\in U^{(C)}} \Sigma_{j\in U^{(C)}} \frac{(\pi_{ij}-\pi_i\pi_j)}{\pi_i\pi_j} E_i E_j\ . \end{aligned} \tag{3.11}$$

An estimate of $\mathrm{Var}(\hat{Y}_{CAL}^{(C)})$ which includes the calibration factors is given by,

$$\hat{\mathrm{Var}}\,(\hat{Y}_{CAL}^{(C)}) = \Sigma_{i\in s^{(C)}} \Sigma_{j\in s^{(C)}} \frac{(\pi_{ij}-\pi_i\pi_j)}{\pi_{ij}\pi_i\pi_j}(g_i e_i)(g_j e_j) \tag{3.12}$$

where $e_i = y_{i\bullet} - \mathbf{z}_i'\hat{\boldsymbol{B}}^{(C)}$ and $\hat{\boldsymbol{B}}^{(C)} = (\Sigma_{i\in s^{(C)}}\ w_i\mathbf{z}_i\mathbf{z}_i'/c_i)^{-1}(\Sigma_{i\in s^{(C)}}\ w_i\mathbf{z}_i y_{i\bullet}/c_i)$. Under stratified SRSWOR selection of n_h clusters from the N_h in stratum h, we have $\hat{\mathrm{Var}}_{SRSWOR}(\hat{Y}_{CAL}^{(C)}) = \Sigma_{h=1}^{H}\{N_h^2(1-f_h)/(n_h(n_h-1))\}\Sigma_{i\in s_h^{(C)}}(u_i-\bar{u}_h)^2$ where $u_i = g_i e_i$, $\bar{u}_h = \Sigma_{i\in s_h^{(C)}} u_i/n_h$ and $f_h = n_h/N_h$.

The framework in sections 3.1 and 3.2 can be readily extended to handle calibration groups based on element or cluster auxiliary information. We also note that these two families generally produce different estimators since they are based on different auxiliary information.

4. TWO-PHASE SAMPLING DESIGNS

Two-phase sampling is increasingly being used at Statistics Canada due to the wealth of timely administrative data. This is especially the case in business surveys where this procedure has been used for several annual and sub-annual surveys. Two-phase sampling for annual surveys are described in Choudhry et al. (1989), and Armstrong and St-Jean (1994), whereas for sub-annual surveys, they are described in Hidiroglou et al. (1995), and Binder et al. (2000).

The use of auxiliary information in two-phase element designs is described in Hidiroglou and Särndal (1998). Estevao and Särndal have extended this to a general framework for estimation in two-phase designs with auxiliary information at each phase. The main ideas in their paper are described in this section assuming WOR sampling. The population $U = \{1, ..., k, ..., N\}$ is first divided into strata U_h. The first-phase probability sample s_1 is selected within each of these strata. The probability of selecting element k in stratum U_h is denoted by π_{1k} and its first-phase sampling weight is defined as $w_{1k} = 1/\pi_{1k}$. The first-phase sample is then divided into second-phase strata s_{1i} with $s_1 = \bigcup_i s_{1i}$. A second-phase sample s_2 is obtained by selecting within each of these strata. The conditional probability of selecting the first-phase element k in s_2 is denoted by π_{2k} and the conditional second-phase weight is given by $w_{2k} = 1/\pi_{2k}$. The final sampling weight for element $k \in s_2$ is $w_k = w_{1k}w_{2k}$. We assume that we have two sets of auxiliary variables for the first phase elements. These are denoted by $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$. The population totals $\boldsymbol{X}_1 = \Sigma_{k\in U}\boldsymbol{x}_{1k}$ are known for the auxiliary variables $\boldsymbol{x}_1$, whereas $\boldsymbol{x}_2$ have no known population totals. In view of the nested nature of the design, the values of $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ for the first-phase sample are also known for the second-phase sample. In this presentation, it is useful to define a general auxiliary vector $\boldsymbol{x}$ that may contain any of the variables in $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$. This vector is defined quite generally, but we are particularly interested in the cases $\boldsymbol{x} = \boldsymbol{x}_1$ and $\boldsymbol{x} = (\boldsymbol{x}_1', \boldsymbol{x}_2')'$. The auxiliary vector $\boldsymbol{x}_k$ is known for each element k in the second-phase sample along with the value y_k for the variable of interest y. Here, we consider only the estimation of the total $Y = \Sigma_{k\in U} y_k$. The estimation of domain parameters follows from the discussion in sections 2.4 and 2.5.

4.1. Calibration for Two-Phase Sampling

We obtain calibration estimators by determining calibrated weights $\tilde{w}_{1k}$ and $\tilde{w}_k$ using the available auxiliary information at the two phases. We start by obtaining first-phase calibrated weights $\tilde{w}_{1k} = w_{1k}g_{1k}$ for each element k in the first-phase sample, where g_{1k} is the first-phase calibration factor. We then determine the overall calibrated weights $\tilde{w}_k = w_{1k}w_{2k}g_{2k}$ for the elements in the second phase sample, where g_{2k} is the second-phase calibration

factor. The first-phase and overall calibrated weights are $\tilde{w}_{1k}$ and $\tilde{w}_k$. These weights satisfy the following two calibration equations.

$$\begin{aligned} \Sigma_{k\in s_1}\tilde{w}_{1k}\boldsymbol{x}_{1k} &= \Sigma_{k\in U}\boldsymbol{x}_{1k} \\ \Sigma_{k\in s_2}\tilde{w}_{k}\boldsymbol{x}_{k} &= \Sigma_{k\in s_1}\tilde{w}_{1k}\boldsymbol{x}_{k} \end{aligned} \tag{4.1}$$

The first-phase calibrated weights $\tilde{w}_{1k}$ are obtained by minimising $\Sigma_{k\in s_1} c_{1k}(\tilde{w}_{1k}-w_{1k})^2/w_{1k}$ with respect to $\tilde{w}_{1k}$ subject to $\Sigma_{k\in s_1}\tilde{w}_{1k}\boldsymbol{x}_{1k}=\Sigma_{k\in U}\boldsymbol{x}_{1k}$ and bounds $\ell_1<\tilde{w}_{1k}<u_1$. As in section 2.1, c_{1k} are positive values that provide arbitrary weighting coefficients in the objective function. Once the first-phase weights $\tilde{w}_{1k}$ have been determined, the final calibrated weights $\tilde{w}_k$ are obtained by minimising $\Sigma_{k\in s_2} c_k(\tilde{w}_k-w_{1k}w_{2k})^2/(w_{1k}w_{2k})$ with respect to $\tilde{w}_k$ subject to $\Sigma_{k\in s_2}\tilde{w}_k\boldsymbol{x}_k=\Sigma_{k\in s_1}\tilde{w}_{1k}\boldsymbol{x}_k$ and bounds $\ell<\tilde{w}_k<u$.

The two-phase calibration estimator for Y is then given by $\hat{Y}_{CAL}=\Sigma_{k\in s_2}\tilde{w}_k y_k$. As in one-phase sampling, the calibrated weights $\tilde{w}_k$ are sample dependent. They are a function of the auxiliary variables $\boldsymbol{x}_{1k}$ and $\boldsymbol{x}_k$ for $k\in s_1$ and the known totals $\boldsymbol{X}_1=\Sigma_{k\in U}\boldsymbol{x}_{1k}$. The calibration estimator is not design unbiased. It is close to unbiased if $g_{1k}\doteq 1$ for $k\in s_1$ and $g_{2k}\doteq 1$ for $k\in s_2$. This is shown in the next section.

4.2. Bias and Variance Estimation

The following derivation of the bias and variance of the two-phase estimator is given by Estevao and Särndal. We assume that y is linked to $\boldsymbol{x}$ by $y_k=\boldsymbol{x}_k'\boldsymbol{B}+E_k$, and that $\boldsymbol{x}'\boldsymbol{B}$ is then linked to $\boldsymbol{x}_1$ by $\boldsymbol{x}_k'\boldsymbol{B}=\boldsymbol{x}_{1k}'\boldsymbol{B}_1+E_{1k}$ for $k\in U$ as shown below.

$$\begin{aligned} y_k &= \boldsymbol{x}_k'\boldsymbol{B}+E_k \\ \boldsymbol{x}_k'\boldsymbol{B} &= \boldsymbol{x}_{1k}'\boldsymbol{B}_1+E_{1k} \end{aligned} \tag{4.2}$$

As in section 2, this is a conceptual representation over the population. Given this representation, $Y=\Sigma_{k\in U}y_k$ can be written as $Y=\boldsymbol{X}_1'\boldsymbol{B}_1+\Sigma_{k\in U}(E_{1k}+E_k)$. Using (4.1) and (4.2), we have $\hat{Y}_{CAL}=\boldsymbol{X}_1'\boldsymbol{B}_1+\Sigma_{k\in s_1}\tilde{w}_{1k}E_{1k}+\Sigma_{k\in s_2}\tilde{w}_k E_k$. It can be shown that the bias of $\hat{Y}_{CAL}$ is given by

$$\text{Bias}(\hat{Y}_{CAL})=\Sigma_{k\in U}E_{1k}\,\text{E}_{s_1}(g_{1k}-1)+\text{E}_{s_1}\{\Sigma_{k\in s_1}w_{1k}E_k\,\text{E}_{s_2|s_1}(g_{2k}-1)\} \tag{4.3}$$

where E_{s_1} is the expectation over the first phase samples containing element k, and $\text{E}_{s_2|s_1}$ is the conditional expectation over the second phase samples given s_1 has been selected and element k is in s_2. From this expression, it is clear that the bias should be relatively small if $g_{1k}\doteq 1$ for $k\in s_1$ and $g_{2k}\doteq 1$ for $k\in s_2$.

The representation given by (4.2) leads to an approach similar to the one given by Axelson (1998) to provide alternative variance estimators for two-phase sampling. The approximate variance of the two-phase calibration estimator is obtained by assuming $g_{1k}\doteq 1$ for $k\in s_1$ and $g_{2k}\doteq 1$ for $k\in s_2$.

$$\begin{aligned} \text{Var}(\hat{Y}_{CAL}) &= \text{Var}_{s_1}\left(\Sigma_{k\in s_1}w_{1k}g_{1k}(E_{1k}+E_k)\right)+\text{E}_{s_1}\text{Var}_{s_2|s_1}\left(\Sigma_{k\in s_2}w_{1k}w_{2k}g_{2k}E_k\right) \\ &\doteq \text{Var}_{s_1}\left(\Sigma_{k\in s_1}w_{1k}(E_{1k}+E_k)\right)+\text{E}_{s_1}\text{Var}_{s_2|s_1}\left(\Sigma_{k\in s_2}w_{1k}w_{2k}E_k\right) \\ &\doteq \Sigma_{k\in U}\Sigma_{l\in U}\frac{(\pi_{1kl}-\pi_{1k}\pi_{1l})}{\pi_{1k}\pi_{1l}}(E_{1k}+E_k)(E_{1l}+E_l)+\text{E}_{s_1}\left(\Sigma_{k\in s_1}\Sigma_{l\in s_1}\frac{(\pi_{2kl}-\pi_{2k}\pi_{2l})}{\pi_{1k}\pi_{2k}\pi_{1l}\pi_{2l}}E_kE_l\right) \end{aligned} \tag{4.4}$$

where π_{1kl} is the first-phase joint inclusion probability $\Pr\{(k,l)\in s_1\}$ and π_{2kl} is the conditional joint inclusion probability $\Pr\{(k,l)\in s_2\mid s_1\}$.

An estimate of the variance (4.4) is obtained by replacing E_k and E_{1k} by the corresponding sample-based values. These are $e_k = y_k - x_k'\hat{\boldsymbol{B}}$ and $e_{1k} = \boldsymbol{x}_k'\hat{\boldsymbol{B}} - x_{1k}'\hat{\boldsymbol{B}}_1$ where $\hat{\boldsymbol{B}} = \left(\Sigma_{k\in s_2} w_{1k}w_{2k}\boldsymbol{x}_k\boldsymbol{x}_k'/c_{2k}\right)^{-1}\left(\Sigma_{k\in s_2} w_{1k}w_{2k}\boldsymbol{x}_k y_k/c_{2k}\right)$ and $\hat{\boldsymbol{B}}_1 = \left(\Sigma_{k\in s_1} w_{1k}\boldsymbol{x}_{1k}\boldsymbol{x}_{1k}'/c_{1k}\right)^{-1}\left(\Sigma_{k\in s_1} w_{1k}\boldsymbol{x}_{1k}\boldsymbol{x}_k'\hat{\boldsymbol{B}}/c_{1k}\right)$.

Incorporating the calibration factors, we obtain the following estimator of the variance

$$\hat{\mathrm{V}}\mathrm{ar}(\hat{Y}_{CAL}) = \Sigma_{k\in s_2}\Sigma_{k\in s_2}\frac{(\pi_{1kl}-\pi_{1k}\pi_{1l})}{(\pi_{1k}\pi_{1l}\pi_{1kl}\pi_{2kl})}(g_{1k}e_{1k}+g_{2k}e_k)(g_{1l}e_{1l}+g_{2l}e_l) + \Sigma_{k\in s_2}\Sigma_{k\in s_2}\frac{(\pi_{2kl}-\pi_{2k}\pi_{2l})}{(\pi_{1k}\pi_{2k}\pi_{1l}\pi_{2l}\pi_{2kl})}(g_{2k}e_k)(g_{2l}e_l)\,. \tag{4.5}$$

This framework for the two-phase calibration estimator is quite general. The expressions for the bias, variance and estimated variance are valid for any definition of $\boldsymbol{x}$ including the case of no auxiliary variables - $\boldsymbol{x}_1 = \phi$, $\boldsymbol{x}_2 = \phi$ and $\boldsymbol{x} = \phi$. This case gives the two-phase expansion estimator. We note that when $\boldsymbol{x} = \boldsymbol{x}_1$, the population representation (4.2) gives $\boldsymbol{B}_1 = \boldsymbol{B}$ and $E_{1k} = 0$ for $k \in U$. We also obtain the correct correspondence $\hat{\boldsymbol{B}}_1 = \hat{\boldsymbol{B}}$ and $e_{1k} = 0$ for $k \in s_2$. Furthermore, in the special case of $s_2 = s_1$, we can verify that the formulas reduce to those for a one-phase or a one-stage design. In this case, we note $\pi_{2k} = \pi_{2l} = \pi_{2kl} = 1$ and $c_{2k} = c_{1k}$ lead to $w_k = w_{1k}$ and $g_{2k} = g_{1k}$.

The estimator of variance (4.5) can be simplified for stratified SRSWOR and PPSWR. For these two designs, Binder et al. (2000) express it as the sum of three terms, each involving single sums. For example, let us consider a design with SRSWOR at each phase. We have $\pi_{1k} = N_h/n_h$ for $k \in s_{1h}$ and $\pi_{2k} = M_g/m_g$ for $k \in s_{2g}$ where N_h and n_h are the respective first phase population and sample sizes in stratum *h*, and M_g and m_g are the respective second phase population and sample sizes in stratum *g*. The corresponding sampling fractions are $f_h = n_h/N_h$ and $f_g = m_g/M_g$. Arcaro (1998) shows that for stratified SRSWOR at each phase, we can express (4.5) as

$$\hat{\mathrm{V}}\mathrm{ar}(\hat{Y}_{CAL}) = \Sigma_h N_h^2(1-f_h)\frac{s_h^2}{n_h} + \Sigma_h\Sigma_g\frac{N_h^2(1-f_h)M_g^2(1-f_g)}{n_h^2(n_h-1)}\frac{s_{1hg}^2}{m_g} + \Sigma_g\frac{M_g}{m_g}(1-f_g)\frac{s_{2g}^2}{m_g}\,. \tag{4.6}$$

In the description of the 3 terms in this formula, it is convenient to define $\tilde{e}_{1k} = g_{1k}e_{1k} + g_{2k}e_k$ and $\tilde{e}_k = g_{2k}e_k$. In the first term of this expression, $s_h^2 = \left\{\Sigma_g\Sigma_{k\in s_{2g}}(M_g/m_g)\tilde{e}_{1k}^2 - [\Sigma_g\Sigma_{k\in s_{2g}}(M_g/m_g)\tilde{e}_{1k}]^2/n_h\right\}/(n_h-1)$. In the second term, $s_{1hg}^2 = \Sigma_{k\in s_{2g}}(\tilde{e}_{1(h)k} - \bar{\tilde{e}}_{1(h)})^2/(m_g-1)$ with $\tilde{e}_{1(h)k} = \tilde{e}_{1k}$ if $k \in s_{1h}$, 0 otherwise and $\bar{\tilde{e}}_{1(h)} = \Sigma_{k\in s_{2g}}\tilde{e}_{1(h)k}/m_g$. The last term has $s_{2g}^2 = \Sigma_{k\in s_{2g}}(\tilde{e}_k - \bar{\tilde{e}})^2/(m_g-1)$ with $\bar{\tilde{e}} = \Sigma_{k\in s_{2g}}\tilde{e}_k/m_g$.

It is possible to extend this framework when auxiliary totals are known for subgroups of the population. This allows us to define first-phase and second-phase calibration groups as described by Hidiroglou and Särndal (1998). In addition, it is possible to have different stratification for the first-phase and second-phase samples. This creates a general family of two-phase calibration estimators. Estevao and Särndal analyse the properties of these estimators and provide a general discussion on the use of auxiliary information in two-phase designs.

5. FUTURE PLANS

Some new methodologies are being considered for implementation in GES. Linear and non-linear parameters of interest, such as population means, ratios, linear and logistic regression coefficients can be expressed as solutions to population estimating equations. Parameter estimates can be obtained by solving the corresponding sample estimating equations involving the design weights and the calibration factors determined from the auxiliary information. As Hidiroglou et al. (1999) have shown, the standard errors of such estimates can be obtained by regressing the components of the estimating functions on the auxiliary variables. The automation of the required computations is relatively simple, only requiring the form of the derivatives to linearise the functions of interest. The estimating equation approach allows the estimation of functions of totals which are not available in GES.

Another area of research is variance estimation in the presence of imputation. Current variance estimation procedures do not take imputation into account. That is, they treat imputed data the same as actual responses. This typically results in an underestimate of the variance. Several papers have addressed this problem. These include Särndal et al. (1992), Rao and Shao (1992), Rancourt et al. (1994), Rao and Sitter (1995) and Rao (1996). The push to incorporate these variance correction procedures in the GES is emphasised in Lee et al. (1994), Gagnon et al. (1996) and Gagnon et al. (1997). A preliminary version of such a system, called SIMPVAR has been developed but is not currently part of GES. SIMPVAR requires the additional specification of the imputation process: the imputation method, the imputation groups, the auxiliary data used for imputation, respondent flags, and donor identifiers for the donor imputation methods.

6. REFERENCES

Andersson, C., and 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**, pp. 395-405.

Arcaro, C. (1998), "GES Estimation Specifications for Two-Phase Sampling with Auxiliary Information," report, Statistics Canada.

Ardilly, P. (1994), *Les techniques de sondages*, Éd. Tecknip - Paris.

Axelson, M. (1998), "Variance Estimation for the Generalized Regression Estimator Under Two-Phase Sampling: A Modified Approach," *Proceedings of the 3rd International Conference on Methodological Issues in Official Statistics*, Stockholm, October 12-13, 1998, pp. 85-90.

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

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

Binder, D.A., Babyak, C., Brodeur, M., Hidiroglou, M., and Jocelyn, W. (2000), "Variance Estimation for Two-Phase Stratified Sampling," to appear in the *Canadian Journal of Statistics*.

Brick, J.M., Broene, P., James P., and Severynse, J. (1997), *A Users's Guide to WesVar PC*, Westat.

Choudhry, G.H., Lavallée, P., and Hidiroglou, M. (1989), "Two-Phase Sample Design for Tax Data," *Proceedings of the Section on Survey Research Methods, American Statistical Association*, pp. 646-651.

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

Estevao, V. (1994), "Calculation of G-Weights Under Calibration and Bound Constraints," report, Statistics Canada.

Estevao, V., Hidiroglou, M.A., and Särndal, C.-E. (1995), "Methodological Principles for a Generalized Estimation System at Statistics Canada," *Journal of Official Statistics*, **11**, pp. 181-204.

Estevao, V. and Särndal, C.-E., "Calibration Estimators in Two-Phase Sampling,", to be published.

Gagnon, F., Lee, H., Provost, M., Rancourt, E., and Särndal, C.-E. (1997), "Estimation of Variance in the Presence of Imputation," *Proceedings: Symposium 97, New orientations*.

Gagnon, F., Lee, H., Rancourt, E., and Särndal, C.-E. (1996), "Estimating the Variance of the Generalised Regression Estimator in the Presence of Imputation for the Generalised Estimation System," *Proceedings of the Survey Methods Section, Statistical Society of Canada*, pp. 151-156.

Hidiroglou M.A. (1991), "Using Stratified SRSWOR in GES to Approximate Stratified PPS (WR or WOR) Designs," Internal memo, Statistics Canada.

Hidiroglou, M.A., and Särndal C.-E. (1998), "Use of Auxiliary Information for Two-Phase Sampling," *Survey Methodology,* **24**, pp. 11-20.

Hidiroglou, M.A., Latouche, M., Armstrong, B., and Gossen, M. (1995), "Improving Survey Information Using Administrative Records: The Case of the Canadian Employment Survey," *Annual Research Conference Proceedings*, pp. 171-197.

Hidiroglou, M.A., Rao, J.N.K., and Yung, W. (1999), "Variance Computations for Complex Surveys Using Estimating Equations," *Proceedings of the Survey Methods Section, Statistical Society of Canada*, *Regina, Alberta, Canada*, pp. 3-9.

Huang, E., and Fuller, W.A. (1978), "Nonnegative Regression Estimators," *Proceedings of the Social Statistics Section*, *American Statistical Association*, pp. 300-305.

Lee, H., Rancourt, E., and Särndal, C.-E. (1994), "Experiments with Variance Estimation from Survey Data with Imputed Values," *Journal of Official Statistics*, **10**, pp. 231-243.

Outrata, E., and Chinnappa, B.N. (1989), "General Survey Functions at Statistics Canada," *Bulletin of the International Statistical Institute*, **53**, 2, pp. 219-238.

Rancourt, E, Särndal, C.-E., and Lee, H. (1994), "Estimation of Variance in Presence of Nearest Neighbour Imputation," paper presented at the annual meeting of the American Statistical Association, Toronto.

Rao, J.N.K. (1963), "On Three Procedures of Unequal Probability Sampling Without Replacement," *Journal of the American Statistical Association,* **58**, pp. 202-215.

Rao, J.N.K. (1996), "On Variance Estimation with Imputed Survey Data," *Journal American Statistical Association*, **91**, pp. 499-506.

Rao, J.N.K., and Shao, J. (1992), "Jack-knife Variance Estimation with Survey Data under Hotdeck Imputation," *Biometrika*, **79**, pp. 811-822.

Rao, J.N.K., and Sitter, R.R. (1995), "Variance Estimation under Two-Phase Sampling with Application to Imputation for Missing Data," *Biometrika*, **82**, pp. 453-460.

Rao, J.N.K., Kovar, J.G., and Mantel, H.J. (1990), "On Estimating Distribution Function and Quantile from Survey Data Using Auxiliary Information," *Biometrika*, **77**, pp. 365-375.

Rosén (1991), "Variance Estimation for Systematic PPS-sampling," SCB R&D report, Statistics Sweden.

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**, pp. 527-537.

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

Schnell, D., Kennedy, W. J., Sullivan, G., Park, J. P. and Fuller, W. A. (1988), "Personal Computer Variance Software for Complex Surveys," *Survey Methodology*, **14**, pp. 59-69.

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