Component of Statistics Canada
Catalogue no. 12-001-X Business Survey Methods Division

Article

The use of auxiliary information in design-based estimation for domains

by Victor M. Estevao and Carl-Erik Särndal

December 1999

Statistics Canada Statistique Canada

Canada

Survey Methodology, December 1999
Vol. 25, No. 2, pp. 213-221
Statistics Canada, Catalogue No. 12-001

The use of auxiliary information in design-based estimation for domains

Victor M. Estevao and Carl-Erik Särndal [1]

Abstract

This paper examines some important issues in the use of auxiliary information to produce design-based estimates for domains. We identify three types of design-based estimators and discuss two of these in detail. Both are defined as linear weighted sums of the observed values y_k of the variable of interest. The first is the linear prediction estimator, which is built on a principle of model fitting and good predictions of the unobserved y_k. The second is the uni-weight estimator, which applies the same weight to y_k in the calculation of all estimates for those domains containing unit *k*. The latter approach has practical advantages for large-scale productions of statistics because it does not require the calculation of different weight systems for the many variables of interest. It is used in Statistics Canada's Generalized Estimation System (GES), which produces point estimates and corresponding design-based variance estimates for any domains. The auxiliary information used to create the weight system determines the precision of the domain estimates. For the uni-weight estimator in particular, a crucial factor for its variance is the level (domain level, population level, or some intermediate level) for which the auxiliary information is known. We define information groups as the subpopulations with known auxiliary totals. These should be as close as possible to the domains of interest in order to produce efficient estimates. We prove that under certain conditions, the variance of the domain total increases monotonically as the information group moves from the domain to the entire population.

Key Words: Design-based domain estimation; Auxiliary information; Level of auxiliary information; Information groups; Prediction estimator; Uni-weight estimator.

1. Introduction

The estimation for domains of various sizes is an important requirement in the production of statistics for most government surveys. In most statistical agencies, the estimation and the associated measurement of precision rest on design-based principles as far as possible. As Singh, Gambino and Mantel (1994) point out: "Most producers of survey data are accustomed to design estimators and the corresponding design-based inferences. They interpret the data in the context of repeated samples selected using a given probability sampling design, and use estimated design-based cv's (coefficients of variation). Where possible, samples should be designed to produce small area estimates of adequate precision, and sample designs should be fashioned with this in mind. Auxiliary data should be used, where possible, to improve the precision of direct small area estimates."

Let $U = \{1, ..., k, ..., N\}$ denote the survey population. A probability sample s is drawn from U. The inclusion probability and the sampling weight of unit k are denoted by π_k and $a_k = 1/\pi_k$ respectively. Let U_d be a domain of interest. It can be an arbitrary subpopulation $U_d \subseteq U$. The variable of interest is denoted by y, and y_k is its value for unit k. We want to estimate the population total $Y = \sum_U y_k$. In dealing with a domain, U_d, it is convenient to use the domain specific variable y_d, defined as $y_{dk} = y_k$ if $k \in U_d$, and $y_{dk} = 0$ if $k \notin U_d$. Similarly, if x is an auxiliary variable, we have $x_{dk} = x_k$ if $k \in U_d$, and $x_{dk} = 0$ if $k \notin U_d$. Then we can write $Y_d = \sum_{U_d} y_k = \sum_U y_{dk}$ and $\sum_{s_d} x_k = \sum_s x_{dk}$, where $s_d = s \cap U_d$ denotes the part of the sample s that falls in domain U_d.

2. An example to introduce the issue

The following example illustrates how different levels of auxiliary information can cause large differences in the variance of an estimator of a domain total. Suppose we have a one-dimensional auxiliary variable x which is strictly positive and positively correlated with the variable of interest y. Simple random sampling without replacement (SRSWOR) is used to draw a sample s of size n from U, so $a_k = N/n$ for all $k \in U$. Consider a domain of interest U_d for which we need to estimate $Y_d = \sum_{U_d} y_k$. The following three design-based estimators come to mind.

$$\hat{Y}_{d1} = (X_d/\hat{X}_{d\pi})\,\hat{Y}_{d\pi} = X_d\hat{R}_d$$

$$\hat{Y}_{d2} = (X/\hat{X}_{\pi})\,\hat{Y}_{d\pi} = X\,\hat{R}_{(d)}$$

$$\hat{Y}_{d3} = \hat{Y}_{d\pi}$$

where $\hat{Y}_{d\pi} = (N/n)\sum_s y_{dk}$, $\hat{X}_{d\pi} = (N/n)\sum_s x_{dk}$, $\hat{X}_{\pi} = (N/n)\sum_s x_k = N\bar{x}_s$, $X_d = \sum_U x_{dk}$, $X = \sum_U x_k$, $\hat{R}_d = \sum_s y_{dk}/\sum_s x_{dk}$ and $\hat{R}_{(d)} = \sum_s y_{dk}/\sum_s x_k$. All three estimators are design

1. Victor M. Estevao and C.-E. Särndal, Statistics Canada, Ottawa, Ontario K1A 0T6, Canada.

consistent. Estimator $\hat{Y}_{d1}$ uses the auxiliary total X_d at the domain level, whereas $\hat{Y}_{d2}$ uses the auxiliary total X at the population level. In practice, this distinction comes into play when the auxiliary information is derived from a source other than the current survey, such as an administrative data source. The total at the domain level is not always available to construct $\hat{Y}_{d1}$, but we often know the total at the population level required for $\hat{Y}_{d2}$. Since $\hat{Y}_{d1}$ uses more detailed information, we intuitively expect that its variance should be smaller than that of $\hat{Y}_{d2}$. Finally, $\hat{Y}_{d3}$ is the Horvitz-Thompson (HT) estimator. Although design-unbiased, it is usually less efficient since auxiliary information is not used.

The variance of $\hat{Y}_{dj}$ for $j = 1, 2, 3,$ is approximately

$$V(\hat{Y}_{dj}) = N^2(1/n - 1/N)\, P_d\, S^2_{yU_d} H_{dj} \qquad (2.1)$$

where $P_d = N_d/N$ is the relative size of the domain, $S^2_{yU_d} = \sum_{U_d}(y_k - \bar{y}_{U_d})^2/(N_d - 1)$ is the variance of y in the domain and H_{dj} is the only factor differentiating the three variances. Let $K_d = \mathrm{cv}_{xU_d}/\mathrm{cv}_{yU_d}$ where $\mathrm{cv}_{yU_d} = S_{yU_d}/\bar{y}_{U_d}$ and $\mathrm{cv}_{xU_d} = S_{xU_d}/\bar{x}_{U_d}$ are the coefficients of variation of y and x within the domain, and let $r_d = S_{xyU_d}/S_{xU_d}S_{yU_d}$ denote the corresponding correlation coefficient. Then we have

$$H_{d1} = 1 + K_d^2 - 2r_dK_d$$

$$H_{d2} = 1 - 2P_dM_dr_dK_d + \{1 + P_d[1 - 2M_d + (\mathrm{cv}_{xU})^2]\}/(\mathrm{cv}_{yU_d})^2$$

$$H_{d3} = 1 + (1 - P_d)/(\mathrm{cv}_{yU_d})^2$$

where $M_d = \bar{x}_{U_d}/\bar{x}_U$ and $\mathrm{cv}_{xU} = S_{xU}/\bar{x}_U$. These expressions can be obtained using, for example, Särndal, Swensson and Wretman (1992), Chapters 6 and 9. The terms H_{d1} and H_{d2} follow from the Taylor variance, that is, the variance of the linearized statistic. The expression for H_{d3} follows from the exact HT variance. The approximations $(N_d - 1)/(N - 1) \cong N_d/N = P_d$ and $(N_d - 1)/N_d \cong 1$ were used in all three cases.

Since the variances $V(\hat{Y}_{dj})$, $j = 1, 2, 3,$ depend on several parameters, it is not so easy to compare them. In Table 1 we compare the three variances for different values of r_d and P_d under the assumption $M_d = \bar{x}_{U_d}/\bar{x}_U = 1$ and $\mathrm{cv}_{yU_d} = \mathrm{cv}_{xU_d} = \mathrm{cv}_{xU} = 1$. Roughly speaking, we assume that y and x have the same variability in the domain and that x has a similar distribution in the domain as in the population. For domains of size $P_d \leq 0.5$, Table 1 shows the following.

1. $\hat{Y}_{d1}$ has considerably smaller variance than $\hat{Y}_{d2}$, particularly as the domain size decreases and the correlation increases. It is not surprising that $\hat{Y}_{d2}$ is less efficient than $\hat{Y}_{d1}$. What is surprising is the rapid rate at which this occurs.

2. $\hat{Y}_{d2}$ has only marginally smaller variance than the HT estimator $\hat{Y}_{d3}$. This is particularly striking for smaller domains (P_d = 0.1 and 0.3), but even if the domain is as large as half the population ($P_d = 0.5$), $\hat{Y}_{d2}$ is only moderately more efficient than $\hat{Y}_{d3}$ when the correlation is 0.9 or larger.

Table 1
Variance ratios comparing $\hat{Y}_{d1}$, $\hat{Y}_{d2}$ and $\hat{Y}_{d3}$ under SRSWOR; $V_{dj} = V(\hat{Y}_{dj})$, $j = 1, 2, 3$; r_d is the correlation between x and y in the domain; $P_d = N_d/N$ is the relative domain size

	Variance Ratio							
	V_{d2}/V_{d1}				V_{d3}/V_{d2}			
	r_d				r_d			
P_d	0.70	0.80	0.90	0.95	0.70	0.80	0.90	0.95
0.1	3.10	4.60	9.10	18.10	1.02	1.03	1.04	1.05
0.3	2.63	3.80	7.30	14.30	1.08	1.12	1.16	1.19
0.5	2.17	3.00	5.50	10.50	1.15	1.25	1.36	1.43
1.0	1.00	1.00	1.00	1.00	1.67	2.50	5.00	10.00

In fact, under conditions other than those of Table 1, it is easy to see from (2.1) that $\hat{Y}_{d2}$ can have a larger variance than $\hat{Y}_{d3}$. This can happen for example, when $M_d = \bar{x}_{U_d}/\bar{x}_U > 1$ and $\mathrm{cv}_{xU_d}/\mathrm{cv}_{xU} < 1$. Therefore, even though $\hat{Y}_{d2}$ uses a highly correlated auxiliary variable, an estimator which does not, namely $\hat{Y}_{d3}$, may be a better choice. The poor performance of $\hat{Y}_{d2}$ may seem disappointing but what counts is not so much the use of a highly correlated auxiliary variable as the level at which we have information about this variable. Estimator $\hat{Y}_{d2}$ uses auxiliary information at the population level, and this is not very efficient for estimation at the domain level.

In a survey where the frame provides a positive measure of size x_k for every $k \in U$, it is possible to calculate an auxiliary total at any level – for the domain, the whole population, or any other subpopulation in between. We can use any of these totals to form a ratio estimator for the domain. The example suggests that the estimator with the auxiliary total at the domain level is better than one with an auxiliary total at a level above the domain. The gain from using the highly correlated auxiliary variable diminishes rapidly as the level of the known auxiliary total moves from the domain to the entire population.

Note that $\hat{Y}_{d1}$ is not the only design consistent estimator that can be constructed with the auxiliary information at the domain level, X_d. Another possibility is $\hat{Y}_{d4} = \hat{Y}_{d\pi} + (X_d - \hat{X}_{d\pi})\hat{R}$, where the slope estimate $\hat{R} = \sum_s y_k / \sum_s x_k$ is based on the entire sample s, not only on the domain part of the sample as in $\hat{Y}_{d1}$. The difference is in the underlying regression model: a common slope for the whole population in the case of $\hat{Y}_{d4}$ and a separate slope for the domain in the case of $\hat{Y}_{d1}$. The variances $V(\hat{Y}_{d1})$ and $V(\hat{Y}_{d4})$ are equal (to the same order of approximation as in (2.1)) if $R = \sum_U y_k/\sum_U x_k = \sum_{U_d} y_k/\sum_{U_d} x_k = R_d$. They will not be very different even when $R_d \neq R$. That is, the choice of model is relatively unimportant. By contrast, the level of auxiliary information leads to considerable differences in the variance.

3. Issues raised by the example

The results in Table 1 raise several issues. In this paper, we examine three of these, in the general case of a multidimensional auxiliary vector $\boldsymbol{x}$:

1. The ratio estimators $\hat{Y}_{d1}$ and $\hat{Y}_{d2}$ in the example use the same auxiliary variable x. Thus both should benefit from a positive correlation between x and y. But they are different by construction, and they behave very differently, as the example shows. What are the two construction principles, in the general setting with multidimensional auxiliary vectors? Do these two principles yield identical estimators in some situations? These issues are discussed in sections 4, 5 and 6.
2. For a given domain, we define the information group to be the subpopulation for which the $\boldsymbol{x}$-total is known. In $\hat{Y}_{d1}$, the domain is the information group and for $\hat{Y}_{d2}$ it is the entire population. The example shows that the level of the information group is an important factor for the variance. Are there conditions for which a lower level group will yield a strictly smaller variance than a higher level group? This issue is discussed in section 7.
3. The domain size (number of units in the domain), is another component of auxiliary information. How are the domain sizes incorporated into $\hat{Y}_{d1}$ and $\hat{Y}_{d2}$ in addition to the auxiliary information on $\boldsymbol{x}$? This issue is discussed in section 8.

4. Construction by predicted values obtained by model fitting

4.1 The prediction argument in estimation for the entire population

Suppose the target of estimation is the entire population total, $Y = \sum_U y_k$. A sample s is drawn, giving unit k the sampling weight $a_k = 1/\pi_k$. The data $\{y_k: k \in s\}$ are observed. For non-sampled units, y_k is unknown but suppose we can find a value μ_k that approximates y_k for all units in the population, even if only rather crudely. Then there is strong incentive to build the estimator by "shifting the origin" of unit k from 0 to μ_k, because the residuals $y_k - \mu_k$ are smaller on average than the y_k values and have smaller design-based variance. Now $Y = \sum_U \mu_k + \sum_U (y_k - \mu_k)$, where the known sum $\sum_U \mu_k$ is the dominant term, and the smaller residual sum $\sum_U (y_k - \mu_k)$ requires estimation. Conceptually, two choices must now be made:

(i) Treating the μ_k as non-random, we must choose an estimator for the residual sum $\sum_U (y_k - \mu_k)$.
(ii) We need to find values μ_k close to the y_k. There are two parts to this choice: (a) the model relating y_k to μ_k, and (b) the technique used to fit this model: (generalized) least squares, GLIM, maximum likelihood or some other alternative.

The usual choice in step (i) is the HT estimator, leading to $\hat{Y} = \sum_U \mu_k + \sum_s a_k (y_k - \mu_k)$. This choice is made by convention and is not optimal. No minimum variance unbiased choice exists. Alternatives are the estimators considered by Raj (1956) and Murthy (1957). Auxiliary information is important in step (ii). Let $\boldsymbol{x}$ be auxiliary vector of dimension $J \geq 1$, and let $\boldsymbol{x}_k$ be its value for unit k. Suppose $\boldsymbol{x}_k$ is on the sampling frame for every $k \in s$. Predicted values $\hat{y}_k$ are obtained from the auxiliary information by fitting a model so that $E_m(y_k | \boldsymbol{x}_k, \boldsymbol{\beta}) = f(\boldsymbol{x}_k | \boldsymbol{\beta})$, where E_m is the expectation operator under the model m, $f(\bullet | \boldsymbol{\beta})$ is a specified function, and $\boldsymbol{\beta}$ is an unknown vector of model parameters. The model is **linear** if $f(\boldsymbol{x}_k | \boldsymbol{\beta}) = \boldsymbol{x}_k' \boldsymbol{\beta}$, otherwise **non-linear**.

Using the sample data $\{(y_k, \boldsymbol{x}_k): k \in s\}$, we obtain $\hat{\boldsymbol{B}}$ as an estimate of $\boldsymbol{\beta}$. Then we calculate a predicted value $\hat{y}_k = f(\boldsymbol{x}_k | \hat{\boldsymbol{B}})$, for every $k \in U$. This is feasible because $\boldsymbol{x}_k$ is known for all $k \in U$. Using $\hat{y}_k$ and the HT estimator for the residual sum, we have

$$\hat{Y}_{\text{PRED}} = \sum_U \hat{y}_k + \sum_s a_k (y_k - \hat{y}_k). \qquad (4.1)$$

This is the Generalized Regression (GREG) estimator. We use the subscript PRED rather than GREG to emphasize that the construction is based on predicted values. It is an asymptotically design unbiased (ADU) estimator, regardless of whether or not the model m is "true". Hence, it is called model assisted as opposed to model based. It is not known how to obtain an optimal (minimum variance unbiased) estimator of Y under the twofold choice (i) and (ii). The model is linear if $f(\boldsymbol{x}_k | \boldsymbol{\beta}) = \boldsymbol{x}_k' \boldsymbol{\beta}$, otherwise non-linear.

4.2 Linear model

The generalized least squares method is usually used to estimate the parameters of the linear model. Find $\boldsymbol{B}$ to minimize $\sum_s a_k (y_k - \boldsymbol{x}_k' \boldsymbol{B})^2 / c_k$, where the c_k are suitable positive constants. This leads to

$$\hat{\boldsymbol{B}} = \boldsymbol{T}_s^{-1} \sum_s a_k \boldsymbol{x}_k y_k / c_k \qquad (4.2)$$

where $\boldsymbol{T}_s = \sum_s a_k \boldsymbol{x}_k \boldsymbol{x}_k' / c_k$. The predicted values are $\hat{y}_k = \boldsymbol{x}_k' \hat{\boldsymbol{B}}$, and the construction principle (4.1) gives the linear GREG estimator,

$$\hat{Y}_{\text{LINPRED}} = \boldsymbol{X}' \hat{\boldsymbol{B}} + \sum_s a_k (y_k - \boldsymbol{x}_k' \hat{\boldsymbol{B}}) = \hat{Y}_\pi + (\boldsymbol{X} - \hat{\boldsymbol{X}}_\pi)' \hat{\boldsymbol{B}} \qquad (4.3)$$

where $\boldsymbol{X} = \sum_U \boldsymbol{x}_k$, $\hat{Y}_\pi = \sum_s a_k y_k$ and $\hat{\boldsymbol{X}}_\pi = \sum_s a_k \boldsymbol{x}_k$. The choice of c_k influences the variance but only in a mild way. For some designs, we can find optimal c_k that minimize the Taylor variance of $\hat{Y}_{\text{LINPRED}}$. The specification of $\boldsymbol{x}_k$ should include information from the sample design. For example, consider a stratified SRS design with H strata and

sampling fractions $f_h = n_h/N_h$, for $h = 1, ..., H$. Then, to obtain the minimum Taylor variance, we take $1/c_k = 1/f_h - 1$ for all k in stratum h and we let $\boldsymbol{x}_k = (\xi_{1k}, ..., \xi_{hk}, ..., \xi_{Hk}, \boldsymbol{x}'_{0k})'$, where $\xi_{hk} = 1$ if k is in stratum h, otherwise $\xi_{hk} = 0$ and x_{0k} includes all of the other auxiliary variables with known total $\sum_U \boldsymbol{x}_{0k}$. Then, except for the factor $(n_h - 1)/n_h \cong 1$, (4.3) coincides with an asymptotically optimal solution derived in a different manner by Montanari (1987).

4.3 Non-linear model

In the nonlinear case, the model m can be fitted by GLIM or some other technique. Generalized least squares remains an expedient approach: minimize the weighted sum of squares $\sum_s a_k(y_k - f(\boldsymbol{x}_k|\boldsymbol{\beta}))^2/c_k$ as discussed by Fuller (1996). This produces an estimator $\hat{\boldsymbol{B}}$ of $\boldsymbol{\beta}$. We then calculate the predicted values $\hat{y}_k = f(\boldsymbol{x}_k|\hat{\boldsymbol{B}})$ for all $k \in U$, and the estimator is built, as in the linear case, according to (4.1). In some recent research, the nonlinear case is compared to the linear case. For a categorical variable of interest taking m possible values, Lehtonen and Veijanen (1998) fit a multinomial logistic model $f(\boldsymbol{x}_k|\boldsymbol{\beta}) = P(y_k = i) = \exp(\boldsymbol{x}'_k\boldsymbol{\beta}_i)/\sum_{r=1}^m \exp(\boldsymbol{x}'_k\boldsymbol{\beta}_r)$ for $i = 1, ..., m$. They use weighted log-likelihood to estimate $\boldsymbol{\beta} = (\boldsymbol{\beta}'_1, ..., \boldsymbol{\beta}'_m)'$ and from the resulting predicted values $\hat{y}_k$ they build the estimator as in (4.1). Their empirical investigations indicate that this estimator realizes modest efficiency gains compared to the linear GREG estimator. Unlike the linear fit, the $\hat{y}_k$ for the multinomial logistic model are guaranteed to fall in the unit interval. This model is more realistic and provides better fit for many survey data. However, it requires more detailed auxiliary information since the auxiliary values $\boldsymbol{x}_k$ must be known individually for all k. Unlike the linear model, it is not sufficient to simply know the population total of $\boldsymbol{x}_k$ at some level of aggregation. Firth and Bennett (1997) examine the fitting of GLIM models, producing predicted values $\hat{y}_k = G^{-1}(\hat{B}_1\boldsymbol{x}_{1k} + ... + \hat{B}_J\boldsymbol{x}_{Jk})$ where $G(\bullet)$ is the link function. In an empirical study involving tax auditing data, a binary y-variable, and maximum likelihood fit of the simple logistic function, they find that the improvement over the linear GREG estimator is at most, a few percent. These differences are insignificant compared to the very large effects in Table 1 caused by the level of the auxiliary information.

From a variance perspective, it is important that the model m fits well, because the variance depends on the size of the squared residuals $(y_k - \hat{y}_k)^2$. Lehtonen and Veijanen (1998), and Firth and Bennett (1997) show that when we estimate the entire population total Y, there is only a modest decrease in variance in fitting a non-linear model over a linear model. However, in the estimation of a domain total, described below, this decrease in variance is more pronounced. This seems to be especially true as the domain gets smaller, as suggested by the study of Lehtonen and Veijanen (1998).

4.4 The prediction argument in estimating for a domain

Most surveys require estimates for a large number of domains. There are two simple techniques for constructing a design-based domain estimator. Both take the GREG estimator $\hat{Y}_{\text{PRED}}$ given by (4.1) as the starting point, but the resulting domain estimators are not in general identical. In this section, we present the **predictive argument**, leading up to the **predictive estimator** of a domain total. In section 5.3, we present the **unique weighting argument** leading to the **uni-weight estimator** of a domain total.

The predictive domain estimator is constructed as follows: predicted values $\hat{y}_k$ have been determined for the entire finite population U, under some appropriate model. When the target is a specified domain, it is natural to use the $\hat{y}_k$ and the residuals $(y_k - \hat{y}_k)$ from within that domain only. Replacing U by U_d and s by s_d in (4.1) we get

$$\hat{Y}_{d\,\text{PRED}} = \sum_{U_d} \hat{y}_k + \sum_{s_d} a_k(y_k - \hat{y}_k). \qquad (4.4)$$

By construction, estimator (4.4) meets the objective of additivity over a set of domains that partition the population. That is, the sum of the estimators (4.4) over a set of these domains is equal to $\hat{Y}_{\text{PRED}}$, the entire population estimator given by (4.1). In the special case of linear prediction, $\hat{y}_k = \boldsymbol{x}'_k\hat{\boldsymbol{B}}$, where $\hat{\boldsymbol{B}}$ is given by (4.2). Then (4.4) becomes (see Särndal *et al.* 1992, Chapter 9)

$$\hat{Y}_{d\,\text{LINPRED}} = \boldsymbol{X}'_d\hat{\boldsymbol{B}} + \sum_{s_d} a_k(y_k - \boldsymbol{x}'_k\hat{\boldsymbol{B}}). \qquad (4.5)$$

The construction of (4.5) requires that the total $\boldsymbol{X}_d = \sum_{U_d} \boldsymbol{x}_k$ be known, either from an exterior source, or from the microdata as when both $\boldsymbol{x}_k$ and domain membership are specified in the sampling frame. If the model produces accurate predictions, $\hat{Y}_{d\,\text{LINPRED}}$ can be very precise.

Estimators for domains are frequently classified as either **direct** or **indirect**. In the terminology of Schaible (1992) and Federal Committee on Statistical Methodology (1993), an estimator for a domain is called direct only if it uses values of the variable of interest over the domain and for the time period in question. Otherwise, it is indirect. It follows that $\hat{Y}_{d\,\text{LINPRED}}$ is indirect when $\hat{\boldsymbol{B}}$ is based in part on y-data from outside the domain. For some structures of the auxiliary vector $\boldsymbol{x}_k$, the expression (4.5) for $\hat{Y}_{d\,\text{LINPRED}}$ is of the direct variety, requiring only $\{y_k: k \in s_d\}$. An example is shown in section 6.

5. Construction via a supply of weights

5.1 The linear GREG estimator as a weighting procedure

The linear form of $\hat{Y}_{\text{LINPRED}}$ given by (4.3), invites an alternative view. We can regard $\hat{Y}_{\text{LINPRED}}$ as a linear weighted sum of the observed y_k values. This gives

$$\hat{Y}_{\text{LINPRED}} = \sum_s w_k y_k \tag{5.1}$$

where $w_k = a_k g_k$, with

$$g_k = 1 + \lambda' \boldsymbol{x}_k / c_k \tag{5.2}$$

and $\lambda' = (\boldsymbol{X} - \hat{\boldsymbol{X}}_\pi)' \boldsymbol{T}_s^{-1}$ where $\boldsymbol{T}_s = \sum_s a_k \boldsymbol{x}_k \boldsymbol{x}_k' / c_k$. Note that the g_k depend on both k and s (hence they are random) and their calculation requires the auxiliary total $\boldsymbol{X} = \sum_U \boldsymbol{x}_k$. The domain prediction estimator $\hat{Y}_{d\,\text{LINPRED}}$ given by (4.5) can also be expressed by linear weighting as

$$\hat{Y}_{d\,\text{LINPRED}} = \sum_s w_{dk} y_k \tag{5.3}$$

where the weight is now $w_{dk} = a_k g_{dk}$ with $g_{dk} = \delta_{dk} + \lambda_d' \boldsymbol{x}_k / c_k$ where $\delta_{dk} = 1$ if $k \in s_d$, $\delta_{dk} = 0$ if $k \in s - s_d$ and $\lambda_d' = (\boldsymbol{X}_d - \hat{\boldsymbol{X}}_{d\pi})' \boldsymbol{T}_s^{-1}$. We note two features of the weights w_{dk} in (5.3): (i) each domain requires a separate weight system, because g_{dk} depends not only on k and s but also on the domain U_d; (ii) all units in the sample s (not just those in the domain of interest) may receive non-zero weights g_{dk}. Hence, the estimation may be indirect. Exceptions to this occur when $\boldsymbol{x}_k$ has a structure such that $g_{dk} = 0$ for $k \in s - s_d$. In this case, (5.3) becomes a direct estimator. An example is given in theorem 6.1 of section 6.

5.2 The case for a unique set of weights

The use of separate weight systems for different domains, as in $\hat{Y}_{d\,\text{LINPRED}}$ given by (5.3), is usually efficient but may be considered a drawback in large-scale production of statistics. Most government surveys require estimates for many y-variables and for each domain of interest we need to create a domain specific variable y_d as defined in section 1. Timeliness in the dissemination of survey results is important. Estimates must be produced routinely and rapidly for all the y_d-variables, including the corresponding estimates of variance. It may not be practical to produce and manage separate weight systems for every y_d-variable. These factors speak in favour of a multi-purpose weight system that can be applied, with good results, to all y_d-variables.

5.3 The uni-weight system estimator for domains

The weights in (5.1) are $w_k = a_k g_k$, where the g_k, given by (5.2), depend on the $\boldsymbol{x}_k$ for the sampled units and on the known total $\boldsymbol{X} = \sum_U \boldsymbol{x}_k$ but not on the y_k values. They can be computed once and then applied to any domain specific y_d-variable. The information carried by this weighting system is based on the known total $\boldsymbol{X}$ which we assume includes all the information that is available or that we wish to use. Let us apply the weight system $w_k = a_k g_k$ to the data for domain U_d. We obtain the uni-weight estimator of the domain total Y_d defined by

$$\hat{Y}_{d\,\text{WEIT}} = \sum_{s_d} w_k y_k. \tag{5.4}$$

This is a direct estimator, because it uses y-data only from within the domain. By construction, it is additive over mutually exclusive and exhaustive domains since $\sum_{d=1}^{D} \hat{Y}_{d\,\text{WEIT}} = \sum_s w_k y_k$. Alternatively, we can write $\hat{Y}_{d\,\text{WEIT}}$ as $\sum_{s_d} w_k y_k = \sum_s w_k y_{dk}$. This permits us to determine the basic statistical properties (asymptotic unbiasedness and variance) of $\hat{Y}_{d\,\text{WEIT}}$ from the entire sample s. The subscript dWEIT in (5.4) emphasizes the construction in terms of weighted y-values. If the domain is the entire population U, then the linear prediction estimator (4.5) and the uni-weight estimator (5.4) are identical. Both are equal to the linear GREG estimator (4.3). In general however, they differ for a domain that is a proper subset of U.

The idea of a uni-weight system is the basis for the methodology in GES as given by Hidiroglou (1991) and Estevao, Hidiroglou and Särndal (1995). A single weight system creates economies of scale in many surveys. The uni-weight system is not the most efficient for each of the y_d-variables, but this simple approach can often be used to provide good estimates on a timely basis. Sometimes, there is little to be gained by searching for an "optimal" weighting scheme for each of the y_d-variables. Furthermore, when the $\boldsymbol{x}_k$ values are not available for all units on the frame, there is no choice but to compute the weight system using whatever totals are known from administrative sources. Table 2 provides a summary of the features of $\hat{Y}_{d\,\text{PRED}}$, $\hat{Y}_{d\,\text{LINPRED}}$ and $\hat{Y}_{d\,\text{WEIT}}$.

Table 2
Comparison of non-linear $\hat{Y}_{d\,\text{PRED}}$, linear $\hat{Y}_{d\,\text{PRED}}(\hat{Y}_{d\,\text{LINPRED}})$ and $\hat{Y}_{d\,\text{WEIT}}$

	Estimator		
	non-linear $\hat{Y}_{d\,\text{PRED}}$	$\hat{Y}_{d\,\text{LINPRED}}$	$\hat{Y}_{d\,\text{WEIT}}$
Auxiliary information requirement	$\boldsymbol{x}_k$ for all $k \in U$	$\sum_{U_d} \boldsymbol{x}_k$	$\sum_U \boldsymbol{x}_k$
Linear weights (for y_k)	No	Yes	Yes
Uni-weights (for all y_d-variables)	No	No	Yes
Additivity over domains that partition U	Yes	Yes	Yes

5.4 Calibration as a procedure for creating weights

Calibration is a computational procedure designed to produce a system of weights based on the known total $\boldsymbol{X} = \sum_U \boldsymbol{x}_k$. The calibration procedure starts with the basic weights $a_k = 1/\pi_k$ and modifies them through the use of auxiliary information $\boldsymbol{X}$. We define and minimize a measure of distance between the original weights a_k and the new weights w_k, subject to the calibration constraint $\sum_s w_k \boldsymbol{x}_k = \boldsymbol{X}$ which states that the new weight system must produce an exact estimate of the known vector total $\boldsymbol{X}$. When the distance function is defined as the generalized least squares measure $\sum_s c_k (w_k - a_k)^2 / a_k$, we get the

weights $w_k = a_k g_k$ of $\hat{Y}_{\text{LINPRED}} = \boldsymbol{X}'\hat{\boldsymbol{B}} + \sum_s a_k (y_k - \boldsymbol{x}_k'\hat{\boldsymbol{B}})$. GES uses a procedure that permits individual bounding of weights, see Estevao (1994). The function is minimized subject to the calibration constraint and the bounding constraint $u_k \le w_k \le l_k$ for $k \in s$, where the u_k and the l_k are the user specified bounds. This avoids negative weights and large positive weights whenever a solution exists. Several alternative distance measures have been considered and recently evaluated by Stukel, Hidiroglou and Särndal (1996), and Singh and Mohl (1996). Each distance measure produces a slightly different weight system, but these do not generally lead to significantly different point estimates.

5.5 Information groups

The efficiency of the uni-weight system depends on the availability of auxiliary information for suitable population groups. In general, totals known for smaller groups produce more efficient $\hat{Y}_{d\,\text{WEIT}}$ estimates than totals known for larger groups. This is illustrated by theorem 7.1. We are led to examine the structure of the auxiliary vector total $\boldsymbol{X} = \sum_U \boldsymbol{x}_k$ used to compute the uni-weight system $\{w_k = a_k g_k : k \in s\}$ for $\hat{Y}_{d\,\text{WEIT}}$. The vector total $\boldsymbol{X}$ includes known totals of one or more x-variables either for the whole population U or for a set of population subgroups. The x-variables form a vector that we denote by $\boldsymbol{x}_0$ and call the core vector. We define an information group as a subpopulation with a known core vector total. The groups establish the level of the auxiliary information. The efficiency of the uni-weight system is a function of the core vector variables and the level of the information groups.

Classical post-strata are information groups with $\boldsymbol{x}_{0k} = 1$ for all k. As another example, consider a business survey where $\boldsymbol{x}_{0k} = (x_{1k}, x_{2k})$ is the value for enterprise k of a two-dimensional core vector, where $x_{1k} =$ Number of Employees and $x_{2k} =$ Gross Business Income. If the estimation is based on $\boldsymbol{X} = (\sum_U x_{1k}, \sum_U x_{2k})$, then the information for the core vector is at the entire population level. If $\boldsymbol{X} = (\sum_{U_1} x_{1k}, \sum_{U_1} x_{2k}, \sum_{U_2} x_{1k}, \sum_{U_2} x_{2k})$, then the information about the core vector is at the more disaggregated level defined by U_1 and U_2, for example, a geographical subdivision of the population.

Information groups and domains of interest are different concepts. An information group may be a domain, but in general, domains will cut across information groups. In the above example, the domains of interest may be industry classes. In business surveys, some units change classification. As a result, information groups based on the frame information may not be the same as the domains of interest.

If the core vector has dimension $J \ge 1$, and there are $P \ge 1$ groups, then the auxiliary vector $\boldsymbol{x}_k$ has dimension $J \times P$ and is given by $\boldsymbol{x}_k = (\gamma_{1k}\boldsymbol{x}_{0k}', ..., \gamma_{pk}\boldsymbol{x}_{0k}', ..., \gamma_{Pk}\boldsymbol{x}_{0k}')'$ where γ_{pk} is the group identifier such that $\gamma_{pk} = 1$ if unit k is a member of group p and $\gamma_{pk} = 0$ otherwise. The vector total that must be known is $\boldsymbol{X}' = \sum_U \boldsymbol{x}_k' = (\boldsymbol{X}_{01}', ..., \boldsymbol{X}_{0p}', ..., \boldsymbol{X}_{0P}')$, where $\boldsymbol{X}_{0p} = \sum_{U_p} \boldsymbol{x}_{0k}$ is the known core vector total for information group p. In the special case where the information groups form a partition of U, Estevao, Hidiroglou and Särndal (1995) point out that the factors g_k given by (5.2), can be computed group by group. The factors for units in group p will depend on the known total for that group, $\boldsymbol{X}_{0p}$, but not on the other $P - 1$ known totals. Letting $\boldsymbol{T}_{s_p} = \sum_{s_p} a_k \boldsymbol{x}_{0k}\boldsymbol{x}_{0k}'/c_k$ and $\hat{\boldsymbol{X}}_{0p\pi} = \sum_{s_p} a_k \boldsymbol{x}_{0k}$, we have from (5.2), for units k in group p,

$$g_k = 1 + (\boldsymbol{X}_{0p} - \hat{\boldsymbol{X}}_{0p\pi})' \boldsymbol{T}_{s_p}^{-1} \boldsymbol{x}_{0k}/c_k. \qquad (5.5)$$

6. Equivalence of the prediction estimator and the uni-weight estimator for particular cases

In sections 4 and 5, we examined two arguments for constructing a linear weighted estimator of a domain total. They lead to two possibly different estimators of a domain total, the linear prediction estimator $\hat{Y}_{d\,\text{LINPRED}}$ given by (4.5) and the uni-weight estimator $\hat{Y}_{d\,\text{WEIT}}$ given by (5.4). The motivating factor in the prediction approach is to obtain close predictions $\hat{y}_k = f(\boldsymbol{x}_k|\hat{\boldsymbol{B}})$ of the y_k. In the uni-weight approach, we apply one set of weights to all the y_d-variables in the survey. The motivation here is to construct a unique weight system that uses auxiliary information to the fullest extent possible. Model fitting and getting the closest possible predictions are not the primary concerns.

It is important to emphasize the distinction between the auxiliary vector and the amount of auxiliary information used in the estimation. Both $\hat{Y}_{d\,\text{LINPRED}}$ and $\hat{Y}_{d\,\text{WEIT}}$ use the same auxiliary vector $\boldsymbol{x}_k$. However, the amount of auxiliary information is not necessarily the same: $\hat{Y}_{d\,\text{LINPRED}}$ requires $\sum_{U_d} \boldsymbol{x}_k$, a total at the domain level, whereas $\hat{Y}_{d\,\text{WEIT}}$ requires $\sum_U \boldsymbol{x}_k$, a total at the population level. The estimators $\hat{Y}_{d\,\text{LINPRED}}$ and $\hat{Y}_{d\,\text{WEIT}}$ are not in general identical, but they are equal for certain structures of the auxiliary vector $\boldsymbol{x}_k$ as we now show.

We consider D domains, U_1, ..., U_d, ..., U_D, forming a partition of U. Let $\delta_k = (\delta_{1k}, ..., \delta_{dk}, ..., \delta_{Dk})'$ be the domain identifier vector for unit k and $\boldsymbol{x}_k = (\delta_{1k}\boldsymbol{x}_{0k}', ..., \delta_{dk}\boldsymbol{x}_{0k}', ..., \delta_{Dk}\boldsymbol{x}_{0k}')'$ where $\boldsymbol{x}_{0k}$ is the known core vector for unit k. Then the requirement for $\hat{Y}_{d\,\text{WEIT}}$, "$\sum_U \boldsymbol{x}_k$ must be known", is equivalent to "$\sum_{U_d} \boldsymbol{x}_{0k}$ must be known for each domain". Next, the requirement for $\hat{Y}_{d\,\text{LINPRED}}$, "$\sum_{U_d} \boldsymbol{x}_k$ must be known" is equivalent to "$\sum_{U_d} \boldsymbol{x}_{0k}$ must be known", since the domains are non-overlapping. Thus, when we estimate for all D domains, the use of $\hat{Y}_{d\,\text{LINPRED}}$ requires that "$\sum_{U_d} \boldsymbol{x}_{0k}$ must be known for each domain." Both approaches require the D core vector totals $\sum_{U_d} \boldsymbol{x}_{0k}$, $d = 1, ..., D$. Each domain is an information group for the core vector $\boldsymbol{x}_0$. Are $\hat{Y}_{d\,\text{LINPRED}}$ and $\hat{Y}_{d\,\text{WEIT}}$ identical in this situation? Although they use the same information, this can not be taken for granted because they are constructed differently. However, the following statement shows that they are in fact identical.

Theorem 6.1 Let $U_1, ..., U_d, ..., U_D$ be domains that form a partition of U. Let the auxiliary vector be $\boldsymbol{x}_k = (\delta_{1k}\boldsymbol{x}'_{0k}, ..., \delta_{dk}\boldsymbol{x}'_{0k}, ..., \delta_{Dk}\boldsymbol{x}'_{0k})'$ such that $\boldsymbol{X}_{0d} = \sum_{U_d}\boldsymbol{x}_{0k}$ is a known core vector total for $d = 1, ..., D$. Then the prediction estimator $\hat{Y}_{d\,\mathrm{LINPRED}}$ given by (5.3) and the uni-weight estimator $\hat{Y}_{d\,\mathrm{WEIT}}$ given by (5.4) are identical, and $\hat{Y}_{d\,\mathrm{LINPRED}} = \hat{Y}_{d\,\mathrm{WEIT}} = \sum_{s_d} a_k g^*_{dk} y_k$ where $g^*_{dk} = 1 + (\boldsymbol{X}_{0d} - \hat{\boldsymbol{X}}_{0d\pi})'\boldsymbol{T}^{-1}_{s_d}\boldsymbol{x}_{0k}/c_k$ with $\hat{\boldsymbol{X}}_{0d\pi} = \sum_{s_d} a_k \boldsymbol{x}_{0k}$ and $\boldsymbol{T}_{s_d} = \sum_{s_d} a_k \boldsymbol{x}_{0k}\boldsymbol{x}'_{0k}/c_k$.

The proof follows by showing $g_{dk} = g_k = g^*_{dk}$ for $k \in s_d$ and $g_{dk} = 0$ for all $k \in s - s_d$. The details are omitted. Theorem 6.1 suggests that when possible, we should determine the uni-weight system by using an important set of domains as information groups. The theorem does not in general hold when the domains overlap. For example, consider two domains U_1 and U_2 with a non-empty intersection, $U_{12} = U_1 \bigcap U_2$. Let $\boldsymbol{x}_k = (\delta_{1k}x_k, \delta_{2k}x_k)'$. Then to estimate the first domain total, $Y_1 = \sum_{U_1} y_k$, $\hat{Y}_{1\,\mathrm{WEIT}}$ requires the auxiliary information $\sum_U \boldsymbol{x}_k = (\sum_{U_1} x_k, \sum_{U_2} x_k)'$, while $\hat{Y}_{1\,\mathrm{LINPRED}}$ requires the more detailed information $\sum_{U_1} \boldsymbol{x}_k = (\sum_{U_1} x_k, \sum_{U_{12}} x_k)'$. The two estimators are not identical.

7. A monotonic property of the Taylor variance

The level of the information groups will greatly influence the variance of the uni-weight estimator. The example in section 2 illustrated this. Loosely speaking, the closer a domain is to an information group, the smaller the variance. For specific cases, it can be shown that the variance is a monotonic function of the information group. Theorem 7.1 illustrates this.

Consider a domain of interest, U_d, of the sampled population U, such that U_d is wholly contained in a given information group, U_1, with known auxiliary total $\boldsymbol{X}_{01} = \sum_{U_1}\boldsymbol{x}_{0k}$. Consider also an alternative, larger information group, U_2, with known auxiliary total $\boldsymbol{X}_{02} = \sum_{U_2}\boldsymbol{x}_{0k}$, and such that $U_d \subseteq U_1 \subseteq U_2 \subseteq U$. Thus U_1 provides information at a more detailed level than U_2, and U_1 is "closer" than U_2 to the domain of interest U_d. The uni-weight estimator of Y_d based on the group U_j, $j = 1$ or 2, is given by

$$\hat{Y}_{d\,\mathrm{WEIT}\,j} = \sum\nolimits_s a_k g_{kj} y_{dk} \qquad (7.1)$$

where g_{kj} is given by the right hand side of (5.5) if we replace the index p by j. Theorem 7.1 deals with two designs, Poisson sampling (with inclusion probability $\pi_k \propto z_k$, where z_k is a measure of size) and stratified simple random sampling (STSRS). Note that the choice of the c_k is important to obtain the result. In the theorem, $V(\hat{Y}_{d\,\mathrm{WEIT}\,j})$ denotes the Taylor variance of $\hat{Y}_{d\,\mathrm{WEIT}\,j}$.

Theorem 7.1 Let $U_1 \subseteq U_2$ be any information groups such that $U_d \subseteq U_1 \subseteq U_2 \subseteq U$, where U_d is the domain of interest and $\boldsymbol{X}_{0j} = \sum_{U_j}\boldsymbol{x}_{0k}$ is a known core auxiliary total for U_j, $j = 1, 2$. Then the following holds: (a) under Poisson sampling, $V(\hat{Y}_{d\,\mathrm{WEIT}1}) \le V(\hat{Y}_{d\,\mathrm{WEIT}2})$ provided $c_k = 1/(a_k - 1)$; (b) under STSRS $V(\hat{Y}_{d\,\mathrm{WEIT}1}) \le V(\hat{Y}_{d\,\mathrm{WEIT}2})$ if $\boldsymbol{x}_{0k}$ is defined to include, in addition to other auxiliary variables, the stratum identifier $(\delta_{1k}, ..., \delta_{hk}, ..., \delta_{Hk})$ where $\delta_{hk} = 1$ if unit k belongs to stratum h and $\delta_{hk} = 0$ otherwise, and $c_k = 1$ for all k.

The proof of (a) is given in Appendix A. The proof of (b) is similar. In practice, we usually settle (or have to settle) for a single set of information groups and calculate the uni-weight system as a consequence of this choice. Theorem 7.1 requires rather special conditions but it suggests that in general, we can obtain efficient estimates for important domains by using them as information groups. However, other domains if interest may cut across these information groups and for these domains, the conditions for precise estimates may not be as favourable.

8. Incorporating information about domain sizes

In this section, we return to the example of section 2 with more auxiliary information. In addition to the unidimensional, always positive core auxiliary variable x with a total known either at the population level or at the domain level, we assume now that there is also information about the sizes, N_d, $d = 1, ..., D$, for D domains of interest forming a partition of U.

By formulating the auxiliary vector x_k as in Case A or Case B below, we can incorporate the known domain sizes into the estimator through either of the two construction arguments in sections 4.4 and 5.3. We use the following notation: $S_{xyd} = \sum_{s_d} a_k (x_k - \tilde{x}_{s_d})(y_k - \tilde{y}_{s_d})$, $S_{xxd} = \sum_{s_d} a_k (x_k - \tilde{x}_{s_d})^2$, $S_{yyd} = \sum_{s_d} a_k (y_k - \tilde{y}_{s_d})^2$, where $\tilde{x}_{s_d} = \sum_{s_d} a_k x_k / \hat{N}_d$, $\tilde{y}_{s_d} = \sum_{s_d} a_k y_k / \hat{N}_d$, and $\hat{N}_d = \sum_{s_d} a_k$. We also use the domain identifier $\delta_k = (\delta_{1k}, ..., \delta_{dk}, ..., \delta_{Dk})'$.

Case A Specify the auxiliary vector as $\boldsymbol{x}_k = (\delta'_k, x_k)'$ and let $c_k = 1$ for all k. Then, the predictive estimator (Case A1) and the uni-weight estimator (Case A2) are not identical.

Case Al Jointly, the set of D predictive domain estimators (4.5) requires the information (N_d, X_d), $d = 1, ..., D$, where $X_d = N_d\bar{x}_{U_d} = \sum_{U_d} x_k$ is a known total at the domain level. The prediction estimator for domain U_d becomes

$$\hat{Y}_{d\,\mathrm{LINPRED}} = N_d\tilde{y}_{s_d} + \hat{B}_{\mathrm{COMB}}(X_d - N_d\tilde{x}_{s_d}) \qquad (8.1)$$

where $\hat{B}_{\mathrm{COMB}} = \sum_{d=1}^{D} S_{xyd} / \sum_{d=1}^{D} S_{xxd}$. The underlying regression has a common regression slope for all D domains, whereas the intercepts are allowed to vary between domains. This estimator is indirect since the sampled units in all domains (and not only those in U_d) receive non-zero weights. Each domain has its own weight system.

Case A2 Jointly, the set of uni-weight estimators (5.4) requires the information $(N_1, ..., N_d, ...N_D, X)$, where $X = \sum_U x_k$ is a known total at the population level. The uni-weight estimator for domain U_d becomes

$$\hat{Y}_{d\,\mathrm{WEIT}} = N_d \tilde{y}_{s_d} + \hat{B}_{(d)}(X - \textstyle\sum_{d=1}^{D} N_d \tilde{x}_{s_d}) \qquad (8.2)$$

where $\hat{B}_{(d)} = S_{xyd} / \sum_{d=1}^{D} S_{xxd}$. This direct estimator is the one that we would have to use if the x-total is not available at any level lower than the entire population. The predictive estimator (8.1) will ordinarily have much lower variance than the uni-weight estimator (8.2). Although they are both based on the same auxiliary vector, the information content is higher for (8.1).

Case B Specify the auxiliary vector as $\boldsymbol{x}_k = (\delta_{1k}\boldsymbol{x}'_{0k}, ..., \delta_{dk}\boldsymbol{x}'_{0k}, ..., \delta_{Dk}\boldsymbol{x}'_{0k})'$ and let $c_k = 1$ for all k. Here the core vector is $\boldsymbol{x}_{0k} = (1, x_k)'$, and each domain is an information group. Because $\boldsymbol{x}_k$ has this structure, theorem 6.1 tells us that the prediction estimator and the uni-weight estimator are identical. By either approach, the required information for the D estimates is (N_d, X_d), $d = 1, ..., D$. We have

$$\hat{Y}_{d\,\mathrm{LINPRED}} = \hat{Y}_{d\,\mathrm{WEIT}} = N_d \tilde{y}_{s_d} + \hat{B}_{d\,\mathrm{SEP}}(X_d - N_d \tilde{x}_{s_d}) \quad (8.3)$$

where $\hat{B}_{d\,\mathrm{SEP}} = S_{xyd}/S_{xxd}$. This is a direct estimator, allowing a separate slope and a separate intercept to be fitted in each domain. We can now compare $\hat{Y}_{d\,\mathrm{LINPRED}}$ and $\hat{Y}_{d\,\mathrm{WEIT}}$ from two perspectives: (i) the prediction perspective: (ii) the uni-weight perspective.

The prediction perspective: Both (8.1) and (8.3) are linear prediction estimators. They use the same amount of auxiliary information, but differ in the underlying regression model. The model for (8.1) is a regression with a common slope but with separate intercepts for each domain. The model for (8.3) is one in which each domain has a separate slope as well as a separate intercept. It follows that (8.3) has a better fitting model since more parameters are fitted. Consequently, it has a smaller average squared regression residual and usually a smaller variance, compared to (8.1). However, the variance advantage of (8.3) will be highly limited, often one or two percentage points, depending on the population data. For a small domain, (8.1) may in fact be preferred since the separate slope estimate is unstable when based on few data points. Thus, when both (8.1) and (8.3) are available choices (their common auxiliary information is available), the choice between them (which is a choice between two regression models) is not one of crucial importance.

The uni-weight perspective: Both (8.2) and (8.3) are uni-weight estimators. They require different amounts of auxiliary information. An x-total at the population level suffices for (8.2), but the x-total must be known at the domain level for (8.3). Because the information is much stronger for (8.3), it will usually have considerably smaller variance than (8.2). It would in fact be a mistake to choose (8.2) when the information is available to use (8.3). The amount of auxiliary information is more essential than the choice of model.

9. Discussion and recommendations

We have argued that the prediction approach is not always practical in a survey with many domains and variables of interest, because the search for the best fitting model will often require a lot of effort. In the much simpler uni-weight approach, we attempt to construct a unique set of weights that give good efficiency for all domains and variables of interest. We have given some formal evidence (theorem 7.1) that it is in our interest to have the information groups as close as possible to the principal domains of interest. The selection of the x-variables and, above all, the specification of the information groups are crucial factors in obtaining high overall efficiency in the uni-weight approach. In this paper we have not addressed a question of considerable importance, namely, how to make sure that high overall efficiency is realized, given the multi-purpose use of the uni-weight system.

Acknowledgements

For the formulation of the ideas in this paper, we have benefited from discussions with M.A. Hidiroglou and A. Singh of Statistics Canada, R. Lehtonen of Statistics Finland and S. Lundström of Statistics Sweden. We also acknowledge the helpful comments and suggestions from the Associate Editor and two referees.

Appendix A

Proof of Theorem 7.1. For $j = 1, 2$, denote by V_j the Taylor variance of $\hat{Y}_{d\,\mathrm{WEIT}\,j} = \sum_s a_k g_{kj} y_{dk}$ where $g_{kj} = 1 + (\boldsymbol{X}_{0j} - \hat{\boldsymbol{X}}_{0j\pi})' \boldsymbol{T}_{s_j}^{-1} \boldsymbol{x}_{0k}/c_k$ is based on the information group U_j, with $\hat{\boldsymbol{X}}_{0j\pi} = \sum_{s_j} a_k \boldsymbol{x}_{0k}$ and $\boldsymbol{T}_{s_j} = \sum_{s_j} a_k \boldsymbol{x}_{0k}\boldsymbol{x}'_{0k}/c_k$ where $s_j = s \cap U_j$. Then we have $V_j = \sum\sum_{(k,l)\in U_j} (a_k a_l/a_{kl} - 1) E_{dkj} E_{dlj}$ where E_{dkj} is a regression residual explained below, and $a_{kl} = 1/\pi_{kl}$ where π_{kl} denotes the probability that units k and l are both included in the sample, and $a_{kk} = a_k = 1/\pi_k$. In general, V_j is a quadratic form in the E_{dkj}, but the expression simplifies for Poisson sampling, where $a_{kl} = a_k a_l$ for all $k \neq l$. Then with only squared terms remaining, we have $V_j = \sum_{U_j} E_{dkj}^2/Q_k$ where $Q_k = (a_k - 1)^{-1}$. Now since $c_k = Q_k$, we get $E_{dkj} = y_{dk} - \boldsymbol{x}'_{0k}\boldsymbol{B}_{(d)j}$ for $k \in U_j$, where $\boldsymbol{B}_{(d)j} = (\sum_{U_j} \boldsymbol{x}_{0k}\boldsymbol{x}'_{0k}/Q_k)^{-1} \sum_{U_j} \boldsymbol{x}_{0k} y_{dk}/Q_k$. Let $D_{dk} = E_{dk2} - E_{dk1} = \boldsymbol{x}'_{0k}(\boldsymbol{B}_{(d)1} - \boldsymbol{B}_{(d)2})$. Then

$$V_2 - V_1 = \sum_{U_2} E_{dk2}^2/Q_k - \sum_{U_1} E_{dk1}^2/Q_k$$
$$= \sum_{U_1} \{(E_{dk1} + D_{dk})^2 - E_{dk1}^2\}/Q_k + \sum_{U_2 - U_1} E_{dk2}^2/Q_k.$$

It follows from the normal equations that $\sum_{U_1} D_{dk} E_{dk1}/Q_k = (\boldsymbol{B}_{(d)1} - \boldsymbol{B}_{(d)2})' \sum_{U_1} \boldsymbol{x}_{0k} E_{dk1}/Q_k = 0$, so that $V_2 - V_1 = \sum_{U_1} D_{dk}^2/Q_k + \sum_{U_2 - U_1} E_{dk2}^2/Q_k$. Because both terms on the right hand side are non-negative, we conclude that $V_2 \geq V_1$.

References

Estevao, V.M. (1994). Calculation of *g*-weights Under Calibration and Bound Constraints. Report, Statistics Canada.

Estevao, V.M., Hidiroglou, M.A., and Särndal, C.-E. (1995). Methodological principles for a generalized estimation system at Statistics Canada. *Journal of Official Statistics*, 11, 2, 181-204.

Federal Committee on Statistical Methodology (1993). Indirect estimators in Federal programs. Washington, D.C.: Office of Management and Budget, Statistical Policy Office.

Firth, D., and Bennett, K.E. (1997). Robust models in probability sampling. To appear, *Journal of the Royal Statistical Society*, Series B.

Fuller, W. A. (1996). *Introduction to Statistical Time Series*. New York: Wiley.

Hidiroglou, M.A. (1991). Structure of the Generalized Estimation System. Report, Statistics Canada.

Lehtonen, R., and Veijanen, A. (1998). Logistic generalized regression estimators. *Survey Methodology Journal*, 24, 51-55.

Montanari, G.E. (1987). Post-sampling efficient prediction in large-scale surveys. *International Statistical Review*, 55, 191-202.

Murthy, M.N. (1957). Ordered and unordered estimators in sampling without replacement. *Sankhyā*, 18, 379-390.

Raj, D. (1956). Some estimators in sampling with varying probabilities without replacement. *Journal of the American Statistical Association,* 51, 269-284.

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

Schaible, W.L. (1992). Use of small area estimators in U.S. Federal programs. *Proceedings International Scientific Conference on Small Area Statistics and Survey Design*, 1, 95-114. Warsaw: Central Statistical Office of Poland.

Singh, M.P., Gambino, J., and Mantel, H.J. (1994). Issues and strategies for small area data. *Survey Methodology Journal*, 20, 3-14.

Singh, A.C., and Mohl, C.A. (1996). Understanding calibration estimators in survey sampling. *Survey Methodology Journal*, 22, 107-115.

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 Journal*, 22, 117-125.