Journal of Official Statistics, Vol. 12, No. 1, 1996, pp. 3–32

Robust Case-Weighting for Multipurpose Establishment Surveys

R.L. Chambers[1]

Case-weighting or assigning a unique weight to each sample unit is a popular method of sample weighting when internal consistency of the survey estimates is paramount. If in addition external constraints on key variables (the survey benchmarks) must also be met, then case-weights computed via generalised least squares, based on an assumed linear regression model for the survey variables, can be used. Unfortunately, this method of weighting can lead to negative case-weights. It is also susceptible to bias if the linear model is misspecified. This article proposes a modified method of linear regression-based case-weighting which ensures positive weights via use of a ridging procedure, and model misspecification robustness via the inclusion of a nonparametric regression bias correction factor. Empirical results which illustrate the gains from the new method of weighting are presented.

Key words: Sample surveys; sample weighting; model-based approach; ridge regression; nonparametric regression.

1. Motivation and Summary

Consider the following scenario, a not uncommon one for an establishment survey. There are N units in the target population, and a population frame is available which contains, for each of these N units, a unique identifier, a geographical code (R), an industry code (A) and a measure of size (D). In addition, the population totals $T(X_1), T(X_2), \ldots, T(X_p)$ of a set of p non-negative economic activity variables, $X_1, X_2, \ldots, X_p$, are known for this population. These benchmark or control totals may have been collected in a previous census (or larger survey) of the population, or may be produced as a byproduct of administrative data collection processes. In any case, we assume that the survey data analyst has access to the sample values of $X_1, X_2 \ldots, X_p$, though the corresponding nonsample values of these benchmark variables may well be unavailable.

The aim of the survey is to estimate the population totals of a set of m non-negative economic performance variables, say $Y_1, Y_2, \ldots, Y_m$. Suppose further that there is

[1] Department of Social Statistics, University of Southampton, Highfield, Southampton SO17 1BJ, U.K. The work reported in this article was carried out while the author worked with the Department of Statistics, Australian National University, Canberra A.C.T. 0200, Australia.

Dedication and Acknowledgment: This article is dedicated to the memory of E.K. Foreman, whose enthusiasm for the application of modelling ideas in sample surveys was one of the main reasons for the author's interest in the subject, and whose strong support over many years served to allow the author to develop those interests. The author also wishes to acknowledge a number of illuminating conversations with K.R.W. Brewer during the process of developing and writing this paper. Comments from him on previous versions were very helpful in preparing the final draft of this article.

good reason to believe that each Y-variable is approximately proportionately related to one of the X-variables, in the sense that it is reasonable to assume that the regression of this Y-variable on this X-variable is linear and passes through the origin. Furthermore, suppose that the variance of a residual from this population regression is proportional to the corresponding value of the X-variable.

This scenario, of course, is one traditionally identified as where estimation of the population totals of the Y-variables is best carried out by ratio estimation. That is, if we use $X_{(k)}$ to denote the benchmark X-variable 'linked' to the survey variable Y_k, then our estimator of the population total of Y_k is the classical ratio estimator

$$\hat{T}_R(Y_k) = \frac{T_s(Y_k)}{T_s(X_{(k)})} T(X_{(k)}) \tag{1}$$

where $T_s(Y_k)$ and $T_s(X_{(k)})$ denote the sample totals of Y_k and $X_{(k)}$. This estimator is known to have "nice" properties provided the sample s (of size $n \ll N$, say) is drawn using equal probabilities of inclusion. If the sample is drawn using some form of an unequal probability sampling scheme (e.g., stratified sampling, or probability proportional to size sampling), then a simple and approximately design-unbiased alternative to the ratio estimator is the inverse-π-weighted ratio estimator

$$\hat{T}_{R\pi}(Y_k) = \frac{\hat{T}_\pi(Y_k)}{\hat{T}_\pi(X_{(k)})} T(X_{(k)}) \tag{2}$$

where $\hat{T}_\pi(Y_k)$ and $\hat{T}_\pi(X_{(k)})$ are now the Horvitz-Thompson estimators of the population totals of Y_k and $X_{(k)}$, respectively. That is

$$\hat{T}_\pi(Y_k) = \sum_s Y_{ki}\pi_i^{-1} \tag{3}$$

where π_i denotes the sample inclusion probability of sample unit i and $\hat{T}_\pi(X_{(k)})$ is defined similarly. Observe that this inverse-π-weighted ratio estimator reduces to the usual ratio estimator when the π-values in the population are constant. When the π_i are generated by a stratified random sampling scheme, this estimator is usually referred to as the combined ratio estimator.

There are two basic problems with practical implementation of this ratio-based survey estimation strategy. The first is due to the need to identify an appropriate estimation benchmark $X_{(k)}$ to use with each survey variable Y_k. If the number of such survey variables is large (as will be the case in many surveys), then the analyst has to be prepared to carry out a fairly lengthy modelling exercise which matches an appropriate benchmark variable with each of these survey variables.

The second problem relates to the internal consistency of this estimation strategy. This is best illustrated by an example. Suppose Y_1 and Y_2 are two survey variables, with associated estimation benchmark variables $X_{(1)}$ and $X_{(2)}$. The corresponding ratio estimates of the population totals of Y_1 and Y_2 will be denoted $\hat{T}_R(1)$ and $\hat{T}_R(2)$. Suppose now that we also wish to publish an estimate of the population total of Y_1 plus Y_2. One estimator of this total is of course the sum $\hat{T}_R(1) + \hat{T}_R(2)$. However, this is by no means the only estimator we can construct. In fact, it may be more efficient to estimate this total by using a ratio estimator of the form

$$\hat{T}_R(1+2) = \frac{T_s(Y_1) + T_s(Y_2)}{T_s(X_{(1)}) + T_s(X_{(1)})}(T(X_{(1)}) + T(X_{(2)})).$$

That is, we apply the ratio estimation strategy to the derived variable $Y_1 + Y_2$, based on the derived benchmark $X_{(1)} + X_{(2)}$. Clearly $\hat{T}_R(1+2)$ will not equal $\hat{T}_R(1) + \hat{T}_R(2)$ in general. Which approach should we take? If (as will sometimes be the case) $\hat{T}_R(1+2)$ is more efficient than $\hat{T}_R(1) + \hat{T}_R(2)$, then we are essentially in the position of having to trade-off efficiency in estimation against the internal consistency of our estimates.

The problem becomes much worse when we have a large number of survey variables, with complicated interrelationships. It is extremely difficult, if not impossible, to decide, a priori, how to put together a basic set of ratio type survey estimates so that all published survey estimates are obtained as linear combinations of these basic estimates. If the raw survey data are in fact released in some form of public use data file, and users are at liberty to "put these data together" in any way that suits them, then it is quite impossible to achieve consistency. We can never be sure that the analyst who constructs the variable Y_5 by adding together the variables Y_1 and Y_2, will end up with the same estimate of the population total of Y_5 as the analyst who instead (and with equal validity) defines Y_5 by subtracting Y_3 from Y_4.

Of course, if the only output from the survey is a (relatively) restricted set of tabulations, with no subsequent reuse of the sample data for secondary analyses, then one could put in place procedures that ensure that these survey tabulations are always internally consistent. However, the trend in modern surveys is to make the data collected in the survey as widely available as possible, either, as mentioned above, by release of the raw data in a public use data file, or, more generally, by the construction of large scale survey data bases which integrate the data from many related surveys. These data bases are then used for a variety of secondary analyses by a wide range of analysts, most of whom will have had nothing to do with the original survey. In fact, in most cases these analysts are not survey statisticians, but professionals from areas which make heavy use of survey data, e.g., economists, sociologists, etc. For such a scenario, the concept of a fixed set of survey tabulations is meaningless, and internal consistency of survey estimates derived on a continuing basis from the survey data base becomes paramount.

Such consistency is easily achieved by using a method of estimation that allocates a unique weight, say w_i, to each unit or case in the sample, with all survey estimates then being computed as weighted sums based on these case-weights. The survey estimate of the population total $T(Y_k)$ of the variable Y_k is therefore

$$\hat{T}_w(Y_k) = \sum_s w_i Y_{ki}. \tag{4}$$

Given such an approach is taken, a natural question that arises is: How much efficiency (if any) is lost when a case-weighted strategy based on an estimator like (4) is used instead of the more conventional ratio estimation strategy described previously?

In order to answer this question, one needs to specify exactly how the case-weights w_i are computed. For example, a ratio estimation strategy is a special case of a case-weighted strategy provided the same estimation benchmark X is used for all the survey variables. If the number of survey variables is large, however, it is extremely unlikely that one estimation benchmark variable will suffice. Consequently the ratio estimation strategy outlined above can be seen as intrinsically different from a case-weighted estimation strategy. In Section 2 below we briefly describe the standard method of case-weighting for this situation. This approach is based on postulating a linear regression model linking the survey variables and the estimation benchmark variables. Case-weights derived under this approach are automatically calibrated on the benchmark totals for the benchmark variables in the model, but, as shown in Section 3, can suffer from the serious practical problem of sometimes being negative. Positive regression type case-weights can be guaranteed by introducing a ridge modification, as described in Section 4, at the cost of allowing a small amount of slippage in the calibration constraints. However, this approach, like the regression-based weighting procedure from which it is derived, is sensitive to misspecification of the underlying regression model. Robustness to such misspecification can be achieved by introducing a nonparametric bias correction into the ridged case-weights. This modification is described in Section 5. Empirical results on the comparative performances of these different methods of survey weighting are presented in Section 6. These show that a case-weighting approach to survey estimation which includes ridging (to ensure strictly positive weights) and nonparametric bias correction performs extremely creditably in comparison with both the conventional ratio estimation strategy as well as with standard regression-based methods of case-weighting. Finally, we conclude in Section 7 with a short discussion of the closely related issue of variance and confidence interval estimation when a case-weighted approach to survey estimation is adopted.

2. Optimal Calibrated Case-Weights

The classical case-weighted estimator is the Horvitz-Thompson estimator (3). However, this estimator may not be very efficient. Furthermore, it has the major drawback that it is typically not calibrated on the benchmark totals $T(X_1), T(X_2), \ldots, T(X_p)$. That is, there is no guarantee that for the realised sample, the π-weighted survey estimates $\hat{T}_\pi(X_1) . \hat{T}_\pi(X_2), \ldots, \hat{T}_\pi(X_p)$ of these benchmark variables will equal their known population totals. The degree to which a set of survey case-weights is "benchmark calibrated" is widely used as an indication of the likely error of the survey estimates computed using these weights. A large positive (negative) difference between $\hat{T}_w(X_{(k)})$ and $T(X_{(k)})$ is seen as indicating a corresponding positive (negative) value for the estimation error $\hat{T}_w(Y_k) - T(Y_k)$.

As an aside, it should be noted that an almost universally required calibration equation is that the case-weights sum to the population size, N. This constraint is easily incorporated in the above framework by defining X_1 to be identically one.

Case-weights that are calibrated on the benchmark variables can be defined by introducing a global regression model for the survey variables (Huang and Fuller

1978; Bardsley and Chambers 1984; Bethlehem and Keller 1987). That is, we assume that the population values of the kth survey variable can be treated as realisations of a random variable satisfying the regression model

$$E(Y_{ki}|X_i) = X_i'\beta_k \tag{5}$$

where X_i is the p-vector of values of the benchmark variables for the ith population unit. Typically, this model is modified to include industry and geographical effects by including corresponding indicators in the set of benchmark variables. If appropriate, interaction effects between the benchmark variables can also be integrated into the model. The crucial thing to note about this model, compared with the previous ratio estimator model, is that it is overspecified in general, since it is extremely unlikely that every component of β_k will be non-zero.

For the purpose of identifying efficient case-weights, it is necessary to specify the second order moments of the survey variables. Typically, one assumes that

$$\text{var}(Y_{ki}|X_i) = \sigma_k^2 D_i \tag{6}$$

where D_i is the measure of the size of the ith population unit. Following convention, we assume that, conditional on their values of the benchmark variables, different population units are uncorrelated. Again, this assumption can be relaxed, see Royall (1976b).

Given the above working model, there are two distinct approaches to computation of the case-weights. The first is design-based (though model-assisted) in its philosophy, and derives these weights by applying Generalised Regression Estimation (GREG) based on this working model (Särndal, Swensson, and Wretman 1992). The second is completely model-based, and derives these weights by applying Best Linear Unbiased Prediction (BLUP) based on the working model (Royall 1976a). If the underlying population model is such that different population units are uncorrelated, then both approaches can be obtained as special cases of the following constrained optimisation result. Its proof is straightforward.

OPTIMUM (1)

Let $\{\Omega_i; i \in s\}$ and $\{g_i; i \in s\}$ denote two pre-specified sets of positive numbers. The case-weights $\{w_i; i \in s\}$ minimising

$$Q(\Omega, g) = \sum_s \Omega_i\left(\frac{(w_i - g_i)^2}{g_i}\right) \tag{7}$$

subject to the benchmark calibration constraints

$$\hat{T}_w(X_j) = T(X_j); j = 1, 2, \ldots, p \tag{8}$$

are given by

$$w = g + A_s^{-1}X_s(X_s'A_s^{-1}X_s)^{-1}(T - \hat{T}_g) \tag{9}$$

where w denotes the vector with ith element w_i, g denotes the vector with ith element g_i, A_s is the diagonal matrix with ith diagonal element $\Omega_i g_i^{-1}$, X_s denote the $n \times p$ matrix of sample values for the benchmark variables, T is the vector of known population totals for these variables and $\hat{T}_g$ is the corresponding vector of g-weighted sample totals, with jth component

$$\sum_s g_i X_{ji}.$$

Optimal GREG-type case-weights are defined by taking $\Omega_i = D_i$ and $g_i = \pi_i^{-1}$. That is, these weights approximate the Horvitz-Thompson case-weights as closely as possible in terms of the modified chi-square metric defined by $Q(\Omega, \boldsymbol{g})$ above, subject to the calibration constraints (Deville and Särndal 1992). BLUP-type case-weights, on the other hand, are defined by $\Omega_i = D_i$ and $g_i = 1$. Here, $Q(\Omega, \boldsymbol{g})$ corresponds to the prediction variance of the case-weighted estimator under the working model (5), (6). Finally, observe that Särndal and Wright's QR class of estimators (Särndal and Wright 1984) is obtained by setting $\Omega_i = r_i/q_i$ and $g_i = r_i$, where q_i and r_i are constants defined in that reference.

An important point to note in the development above is that in all cases, given the assumed linear regression working model, the calibration constraints are equivalent to imposing a model-unbiasedness condition on the case-weighted estimator.

Which set of case-weights should one use? Advocates of the "design-based/model-assisted" approach to survey inference argue that the GREG case-weights offer the best of both worlds, since the resulting estimator is asymptotically design-unbiased as well as (because of the calibration constraints) exactly model-unbiased. This is claimed to make the GREG estimator both efficient and robust. Unfortunately, the author finds this argument unconvincing. Clearly, if the model is correctly specified, then the GREG estimator, by definition, must be less efficient than the BLUP estimator. On the other hand, if the underlying model is misspecified, there is no logical argument why, for a fixed sample size, the GREG estimator should be more robust, in the sense that the resulting estimate does not deviate markedly from the true population value than the corresponding BLUP. There are alternative, more efficient ways of rendering the BLUP robust to misspecification bias (Chambers, Dorfman, and Wehrly 1993). This issue will be taken up in more detail in Section 5.

3. Negative Case-Weights

Irrespective of whether BLUP or GREG-type case-weights are preferred, there is a very practical problem associated with using case-weights in survey estimation. This is the fact that we have no guarantee that these weights will be greater than or equal to one. Equivalently, if we treat the "representative weight" $u_i = w_i - 1$ as indicating how many non-sample units are "represented" in sample by the ith sample unit, then we have no guarantee that every u_i will be greater than or equal to zero. In fact, there is a good chance that some of these u-weights will be negative.

A number of authors (Huang and Fuller 1978; Bardsley and Chambers 1984; Deville and Särndal 1992; Bankier, Rathwell, and Majkowski 1992; Fuller, Loughin, and Baker 1994) have expressed concern about negative u-weights. Aside from observing that negative weights can lead to negative estimates of intrinsically positive population quantities, and that such weights are also of considerable concern to non-statistical users of the survey data (typically expressed in the form "How can a sample unit possibly represent a negative number of other units in the population?"), we note that such values are in fact symptomatic of deeper problems with the ability of the sample to effectively represent the population.

To see this, consider the optimal case-weights for $p = 2$, with $X_1 = 1$, $X_2 = X$, and $D = X$. That is, our working model assumes that every survey variable Y_k is linked to the benchmark variable X via a simple linear regression model with error variance proportional to X. After some algebra, one can show that the optimal case-weights are given by

$$w_i = \left(\frac{Ng_i}{n\bar{g}_s}\right)\frac{1}{X_i\bar{X}_{sg}^{(-1)}}\left\{1+\frac{(X_i\bar{X}_{sg}^{(-1)}-1)(\bar{X}\bar{X}_{sg}^{(-1)}-1)}{\bar{X}_{sg}\bar{X}_{sg}^{(-1)}-1}\right\}$$

where $\bar{g}_s$ denotes the average of the g-values underlying these case-weights, $\bar{X}$ is the population mean of X,

$$\bar{X}_{sg} = \frac{1}{n\bar{g}_s}\sum_s g_i X_i$$

and

$$\bar{X}_{sg}^{(-1)} = \frac{1}{n\bar{g}_s}\sum_s g_i X_i^{-1.}$$

These optimal case-weights can be negative, particularly if $X_i \gg 1/\bar{X}_{sg}^{(-1)}$ and $\bar{X} \ll 1/\bar{X}_{sg}^{(-1)}$. That is, if our sample is skewed towards large units, there is a good chance that some of the large sample units will have negative case-weights. Conversely, under this working model, samples that are "g-balanced," i.e., ones that satisfy

$$\bar{X} = \frac{1}{\bar{X}_{sg}^{(-1)}}$$

will always generate positive case-weights.

Furthermore, the BLUP and GREG forms of these weights have different propensities for generating negative case-weights, with the GREG tending to be more at risk in this regard, especially in size-biased samples. For example, in the case where the sample inclusion probabilities are proportional to X, it can be shown that, for samples where

$$\bar{X} < \frac{\bar{X}_s^{(-1)}}{\bar{X}_s^{(-2)}}$$

the necessary condition for a GREG case-weight to be positive is

$$X_i < \frac{1-\bar{X}\bar{X}_s^{(-1)}}{\bar{X}_s^{(-1)}-\bar{X}\bar{X}_s^{(-2)}}.$$

Here $\bar{X}_s^{(j)}$ denotes the sample mean of the jth power of X. For the same sample, the condition for the corresponding BLUP case-weight to be positive is the easier to satisfy

$$X_i < \frac{\bar{X}_s-\bar{X}}{1-\bar{X}\bar{X}_s^{(-1)}}.$$

4. Using Ridging to Obtain Positive Case-Weights

Negative case-weights are essentially a symptom of the weighting procedure's attempt to compensate for sample imbalance when meeting the benchmark calibration constraints. We expect to see negative case-weights in unbalanced samples because that is the only way the calibration constraints can be met for these samples. Conversely, as has been noted above, samples that are close to "g-balance" should have little problem with negative case-weights.

The situation gets more complicated as the number of benchmarks (p) increases. In general, the sign of a case-weight computed via (9) depends on the inverse of the matrix $\boldsymbol{X}_s'\boldsymbol{A}_s^{-1}\boldsymbol{X}_s$. If this matrix is close to multicollinear, then a negative case-weight may result. Approximate multicollinearity may be due to the definition of the benchmark variables making up $\boldsymbol{X}_s$, the realised sample distribution of the values of these variables or (most likely) some combination of these two effects. In any case, the end result is the same – the only way the case-weights can be made to satisfy the benchmark calibration constraints (8) is for some of these weights to be negative.

Recognising this, Bardsley and Chambers (1984) proposed that a ridge modification be incorporated into the optimality criterion used to derive the optimal case-weights (9). The effect is to replace OPTIMUM (1) by

OPTIMUM (2)

Let $\{C_j; j = 1, \ldots, p\}$ denote a set of prespecified non-negative constants, where C_j represents the cost of the case-weighted estimator not satisfying the jth calibration constraint, and let λ denote a user-specified scale factor. The case-weights $\{w_i(\lambda); i \in s\}$ minimising the λ-scaled and cost-ridged loss function

$$Q_\lambda(\Omega, \boldsymbol{g}, \boldsymbol{C}) = \sum_s \Omega_i \left(\frac{(w_i - g_i)^2}{g_i} \right) + \frac{1}{\lambda} \sum_{j=1}^{p} C_j (\hat{T}_w(X_j) - T(X_j))^2 \tag{10}$$

are

$$\boldsymbol{w}(\lambda) = \boldsymbol{g} + \boldsymbol{A}_s^{-1}\boldsymbol{X}_s(\lambda\boldsymbol{C}^{-1} + \boldsymbol{X}_s'\boldsymbol{A}_s^{-1}\boldsymbol{X}_s)^{-1}(\boldsymbol{T} - \hat{\boldsymbol{T}}_g). \tag{11}$$

Here $\boldsymbol{C}$ is the diagonal matrix of order p defined by the costs $\{C_j; j = 1, \ldots, p\}$. All other quantities were defined in OPTIMUM (1).

For $\lambda = 0$, these ridged case-weights are identical to the optimal calibrated case-weights (9). For $\lambda > 0$ the ridged case-weights define a set of biased estimators of the population totals of the survey variables. This is because when $\lambda > 0$ the ridged case-weights do not satisfy the p calibration constraints, and hence do not satisfy the necessary conditions for defining an unbiased estimator under the working model (5).

On the other hand, as λ increases away from zero, the variability of these ridged case-weights decreases. An immediate consequence is that the number of negative ridged case-weights decreases, and eventually, for a large enough value of λ, all case-weights defined by (11) are greater than or equal to one. Bardsley and Chambers (1984) argued that this reduction in the variability of the ridged case-weights as λ

increases implied a corresponding reduction in the variance of the survey estimator defined by these weights. Hence, by appropriately choosing λ, one could define a set of ridged case-weights with better mean squared error properties than the (unbiased) optimal calibrated case-weights (9).

To illustrate the nature of the ridged case-weights, consider again the special case where $p = 2$, with $X_1 = 1$, $X_2 = X$, and $D = 1$. Let C_1 denote the cost associated with not meeting the calibration constraint on X_1 (i.e., the sample sum of the case-weights should equal N), and let C_2 denote the cost associated with not meeting the calibration constraint on X_2. Also, assume a model-based approach is taken, so $g_i = 1$. Then, after some algebra, one can show that the ridged case-weight for the ith sample unit is

$$w_i(\lambda) = 1 + \left(\frac{N-n}{n}\right)\left(\frac{S_X^2 + (X_i - \bar{X}_s)(\bar{X}_r - \bar{X}_s) + \frac{\lambda}{n}(C_2^{-1} + C_1^{-1}X_i\bar{X}_r)}{S_X^2 + \frac{\lambda}{n}(C_2^{-1} + C_1^{-1}\bar{X}_s^{(2)} + \lambda C_1^{-1}C_2^{-1})}\right)$$

where $S_X^2 = \bar{X}_s^{(2)} - \bar{X}_s^2$ and $\bar{X}_r$ denotes the non-sample mean of X. In practice, the calibration constraint on X_1 must be met, which corresponds to putting an infinite cost on not meeting it, or equivalently, setting $C_1^{-1} = 0$. In this case

$$w_i(\lambda) = 1 + \left(\frac{N-n}{n}\right)\left(1 + \frac{(X_i - \bar{X}_s)(\bar{X}_r - \bar{X}_s)}{S_X^2 + \lambda n^{-1}C_2^{-1}}\right).$$

Observe that for any value of λ these weights sum to N. Furthermore as λ increases, $w_i(\lambda)$ approaches the sample expansion factor N/n. Thus for large values of λ, the ridged case-weighted estimator behaves like the simple expansion estimator.

Once the cost matrix C has been specified, two diagnostic plots can be used to decide on a value for λ. The first is a plot showing the change in the individual ridged case-weights $w_i(\lambda)$ as the natural logarithm of λ, $\ln(\lambda)$, changes. Figure 1 shows such a plot for one of the simple random samples ($n = 100$) used in the simulation study reported in Section 6. Here $2\ln(\lambda) + 21$ ranges between 1 and 51. The second is a plot of the change in the calibration errors $\hat{T}_{w(\lambda)}(X_j) - T(X_j); j = 1, 2, \ldots, p$ as $\ln(\lambda)$ changes. Figure 2 shows this plot for the same sample as in Figure 1. Both plots were generated using the BLUP version of the ridged case-weights (i.e., $g_i = 1$ in (11)).

Inspection of Figure 1 shows that at $\ln(\lambda) = -10.5$ (i.e., where the $w_i(\lambda)$ are essentially the BLUP weights) ten of the case-weights are negative. As $\ln(\lambda)$ increases, the spread of these case-weights decreases, and eventually, these weights are all greater than or equal to one. Conversely, as $\ln(\lambda)$ increases, Figure 2 shows that the corresponding calibration errors tend to increase. The value of λ recommended by Bardsley and Chambers (1984) is the one where all ridged case-weights are greater than or equal to one, but where Figure 2 shows an acceptable level of calibration error. In the case of the sample underlying these figures, all case-weights are greater than or equal to one when $2\ln(\lambda) + 21 = 33$, corresponding to relative calibration errors of approximately -10% for the benchmarks *Wheat area*, *Sheep number*, *Beef number* and *Dairy number* shown in Figure 2 (see Table 1 for definitions of these variables).

Using a ridged case-weight like (11) is not necessarily the best or only way to avoid negative case-weights. Since the incidence of such weights increases with the size or

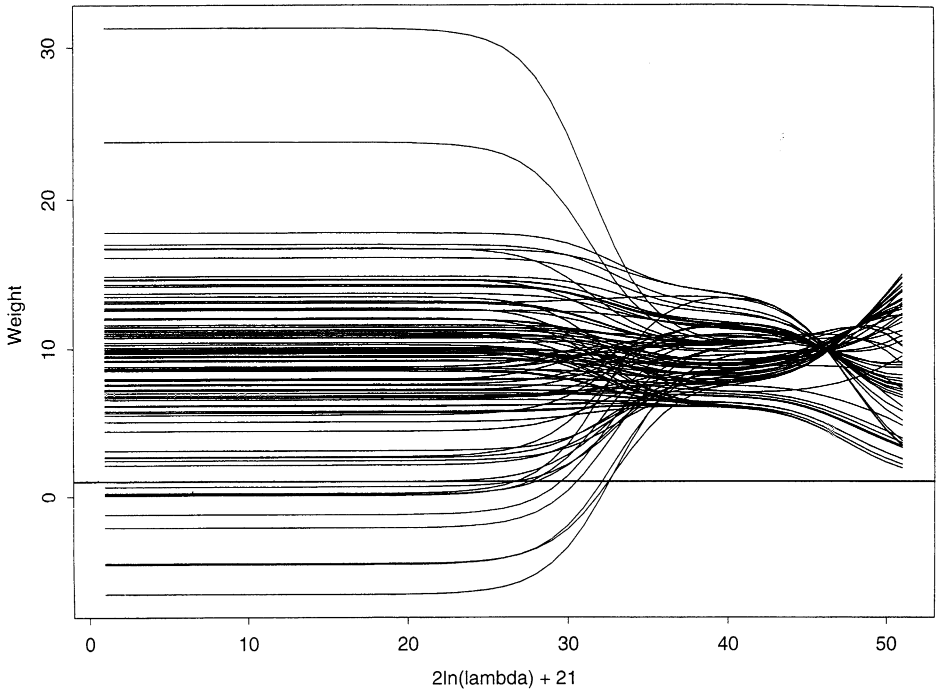

Fig. 1. Ridge trace plot of case-weights generated by the weighting method (11) with $g_i = 1$. Each line in the plot shows how the weight of an individual sample unit changes as the ridge parameter λ changes (in units of $2\,ln(\lambda) + 21$). The sample is of size 100 and is one of the simple random samples used in the simulation exercise reported in Section 6. The underlying model is the model "L" referred to in that exercise, and the cost matrix C is defined in Table 3

complexity (p) of the working model, another approach (Bankier, Rathwell, and Majkowski 1992) is to reduce this complexity by dropping some of the calibration constraints. Since each constraint is equivalent to a parameter in the working model, this approach is equivalent to computing the case-weights via OPTIMUM (1), but under a smaller working model. Note that the ridged case-weights can be made to emulate this strategy by defining the cost matrix $\boldsymbol{C}$ so that the costs associated with the dropped benchmarks are considerably less than those associated with the benchmarks retained in the smaller working model. In the limit, one can make the costs associated with the retained benchmarks effectively infinite, in which case the ridged case-weights recover the values of (9) under the smaller model. As Bardsley and Chambers (1984) point out, the ridged case-weights can then be seen as interpolating from a big working model (all benchmark constraints satisfied) to a small working model (subset of constraints satisfied).

Another, more ad hoc, method of dealing with negative case-weights is to arbitrarily set extreme case-weights to unity and remove the corresponding sample units from the weighting process. Such extreme units are typically those whose case-weights are considerably less than zero, being far removed from the case-weights of the remaining sample units. Standard case-weights are then computed for the remaining sample units, using appropriately adjusted benchmark constraints. This

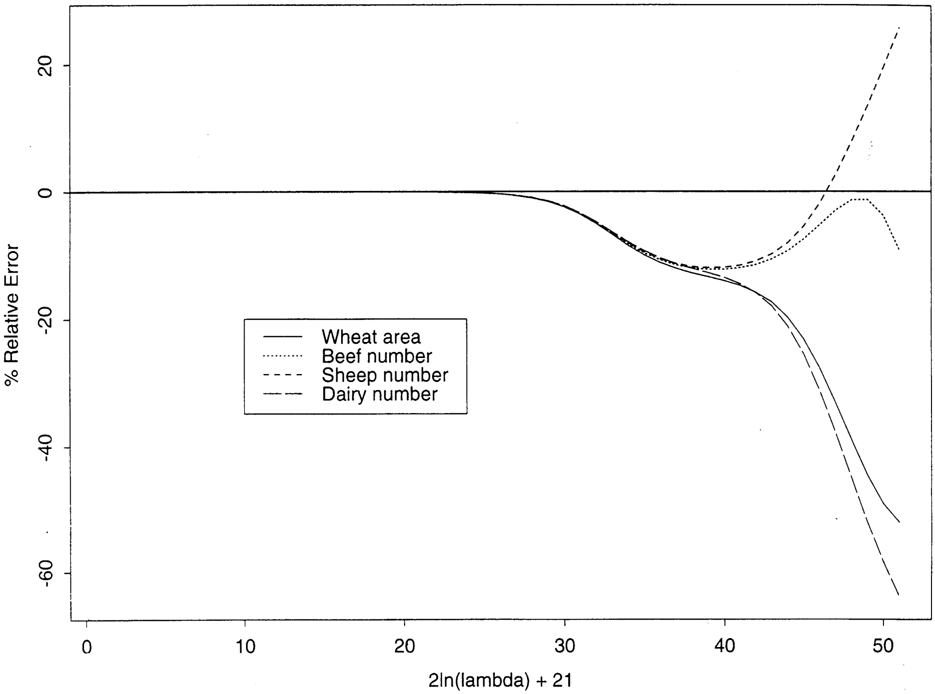

Fig. 2. Ridge trace plot showing how the relative benchmark errors $(T_k - \hat{T}_{gk})/T_k \times 100$ *change as the ridge parameter* λ *changes (in units of* $2\,ln(\lambda) + 21$*). This plot was generated by the same sample and case-weights as in Figure 1. The benchmarks shown (*Wheat area, Beef number, Sheep number *and* Dairy number*) are a subset of those used in the model "L" defined in Section 6*

sample modification approach works reasonably well provided the underlying sample imbalance that is causing the negative case-weights can be effectively corrected by the removal of one or more units from the sample. However, this is not always possible since, as one extreme sample unit is excluded, another unit moves in to take its place. We do not consider this approach further here.

Huang and Fuller (1978), see also Fuller, Loughin, and Baker (1994), take a somewhat different approach to dealing with such extreme weights. They suggest calculating GREG-type weights based on the model (5) but with (9) essentially replaced by iterating the following sequence

$$\boldsymbol{w}^{(k)} = \boldsymbol{g} + \mathrm{diag}(\boldsymbol{g})\boldsymbol{F}_s^{(k)}\boldsymbol{X}_s(\boldsymbol{X}_s'\mathrm{diag}(\boldsymbol{g})\boldsymbol{F}_s^{(0)}\boldsymbol{F}_s^{(1)}\cdots\boldsymbol{F}_s^{(k)}\boldsymbol{X}_s)^{-1}(\boldsymbol{T} - \boldsymbol{T}_{sg}).$$

Here $\boldsymbol{g}$ is the vector of inverses of the sample inclusion probabilities and $\boldsymbol{F}_s^{(0)}, \boldsymbol{F}_s^{(1)}, \boldsymbol{F}_s^{(2)}, \cdots$ is a sequence of diagonal matrices chosen in such a way that $\boldsymbol{F}_s^{(k)}$ equals the identity matrix only when all case-weights lie within pre-specified bounds (e.g., they are all positive) and the benchmark constraints are met. This procedure is not guaranteed to converge. Furthermore, even if convergence is achieved, the resulting weights do not appear to have any optimality properties beyond the unbiasedness under (5) that is a consequence of the benchmark constraints being met.

Finally, one can replace the modified chi-square metric $Q(\Omega, \boldsymbol{g})$ that is minimised in OPTIMUM (1) by another metric with the property that the resulting case-weights (if they exist) are always positive. This approach was investigated by Deville and Särndal (1992). This option is not attractive to the author, mainly because these alternative metrics have no interpretation from a model-based perspective. Furthermore, three of the four alternative metrics suggested by Deville and Särndal (1992, table 1) are such that a solution to the constrained optimisation problem is not guaranteed, while the sole alternative metric (their case 2) where an explicit solution exists was observed by them to be susceptible to extreme positive weights.

5. Robust Ridged Case-Weighting

The ridged case-weights (11) based on the BLUP (i.e., $g_i = 1$) depend for their validity on the linear model (5) being an accurate representation of the relationship between the survey variables and the benchmark variables. This may be reasonable if the survey is restricted to a relatively homogeneous group of units, but becomes problematical if the target population of the survey is heterogeneous. In such cases, it is unlikely that (5) will be adequate for characterising the relationship between the survey variables and the benchmark variables, and estimation methods based on (5), like the BLUP and its ridged alternative, will be biased.

A method of compensating for potential bias in linear regression weighting is described in Chambers, Dorfman, and Wehrly (1993). This method adds a bias correction term to the linear regression model-based estimate, where the bias correction is computed by nonparametrically smoothing the linear model residuals against frame variables (i.e., variables whose values are known for all units in the population) which are either wrongly excluded from the model, or included, but with an incorrect functional specification, in the model. Application of this idea to ridged case-weighting under the model (5) is described below.

Let $Z_1, Z_2, \ldots, Z_Q$ denote frame variables which are potential smoothers for this process. Note that some of these variables could already by included in the set of benchmark variables used in the ridge weighting process. The bias corrected version of the ridge weighted estimate $\hat{T}_{\text{ridge}}$ of the total of a survey variable Y is

$$\hat{T}_{\text{bias corrected}} = \hat{T}_{\text{ridge}} + \sum_s m_i(Y_i - X_i'\hat{\boldsymbol{\beta}}(\lambda)) \tag{12}$$

where $\hat{\boldsymbol{\beta}}(\lambda)$ denotes the implied ridge weighted estimate of the regression parameter β associated with Y in the linear model (5), and the m_i are nonparametric prediction weights obtained by summing the contributions of the ith sample farm to a nonparametric prediction of the residual associated with the fit of this linear model at each of the $N - n$ nonsample farms. In the case where the Z's are all interval scaled, and a product kernel Nadaraya-Watson nonparametric smoother is used, the m_i are given by

$$m_i = \sum_{k \in r} \frac{\prod_{q=1}^{Q} K(b_q^{-1}(Z_{iq} - Z_{kq}))}{\sum_{j \in s} \prod_{q=1}^{Q} K(b_q^{-1}(Z_{jq} - Z_{kq}))}$$

Table 1. Study variables

Name	Description
	Framework variables
ASIC	Unique industry classification (Australian Standard Industry Classification) for each farm, with values 181 Wheat growing 182 Wheat growing + Sheep production 183 Wheat growing + Beef cattle production 184 Sheep + Beef cattle production 185 Sheep production 186 Beef cattle production 187 Dairy farm
State	State/Territory in which farm is located NSW New South Wales VIC Victoria QLD Queensland SA South Australia WA Western Australia TAS Tasmania NT Northern Territory
Region	Identifier for 39 geographically defined regions (these are nested within State)
DSE	Size measure (Dry Sheep Equivalent) for a farm. Defined as a linear combination of the outputs from the farm
	Benchmark variables
Wheat area	Area (hectares) sown to wheat during the year
Beef number	Number of beef cattle on the farm at the end of the year
Sheep number	Number of sheep on the farm at the end of the year
Dairy number	Number of dairy cattle on the farm at the end of the year
	Survey variables
Wheat income	Annual income from sale of wheat
Beef income	Annual income from sale of beef cattle
Sheep income	Annual income from sale of wool and sheep
Dairy income	Annual income from sale of milk products
Total income	Annual income from all four activities above

where r denotes the set of $N-n$ nonsample units, K denotes the kernel function of the smoother, and the $b_1, b_2, \ldots, b_Q$ denote the bandwidths of the component smooths.

For categorical Z variables (e.g., industry and geographic indicators) kernel-based smoothing can be replaced by weighting according to closeness defined by an appropriate metric (e.g., a "counting" metric) for data of this type. In general, smoothing against a combination of Q_1 interval scaled Z's and Q_2 categorical Z's is easily accommodated by multiplying the weights associated with the smooth (against the

Q_1 interval scaled Z's) by the closeness weights (associated with the Q_2 categorical Z's) and then renormalising so that the final m_i weights sum to $N - n$.

Given a vector $\boldsymbol{m}$ of nonparametric prediction weights defined in this way, the corresponding set of nonparametrically bias corrected ridge weights is then given by

$$w_s(\lambda, \boldsymbol{m}) = \mathbf{1}_s + \boldsymbol{m} + A_s^{-1}X_s(\lambda C^{-1} + X_s' A_s^{-1} X_s)^{-1}(\boldsymbol{T} - X_s'\mathbf{1}_s - X_s'\boldsymbol{m}). \tag{13}$$

Here $\mathbf{1}_s$ is a n-vector of ones, $\boldsymbol{T}$ is the vector of population totals for the variables defining X_s and A_s is the diagonal matrix with ith diagonal element D_i. Note that these weights are a special case of (11) with

$$\boldsymbol{g} = \mathbf{1}_s + \boldsymbol{m}.$$

The weights (13) depend on choice of an appropriate ridge parameter λ as well as an appropriate set of bandwidths for computing the nonparametric prediction weights $\boldsymbol{m}$. Fortunately, the choices are quite separate. For interval scaled Z's, good empirical results have been obtained by setting $c = 3.0$ in the simple formula

$$b_q = \frac{c}{n^{1/5}}(\text{sample range of } Z_q). \tag{14}$$

Although not investigated in this article, it is quite likely that further gains in efficiency can be obtained using the bandwidth selection procedure described in Chambers, Dorfman, and Wehrly (1993). Once these bandwidth values have been determined, choice of λ can then be carried out by choosing it as the smallest positive value such that all components of $w_s(\lambda, \boldsymbol{m})$ are greater than or equal to one.

6. An Empirical Evaluation

A comparison of the ratio estimation strategy with the various case-weighted strategies described in the previous sections was carried out using economic and production

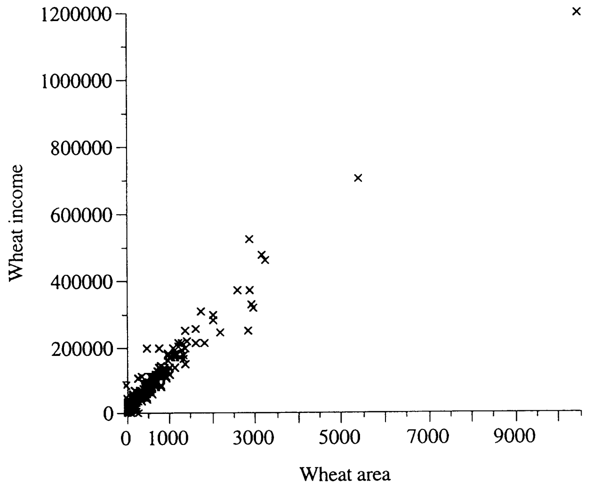

Fig. 3. Scatterplot of Wheat income *vs* Wheat area *for the study population*

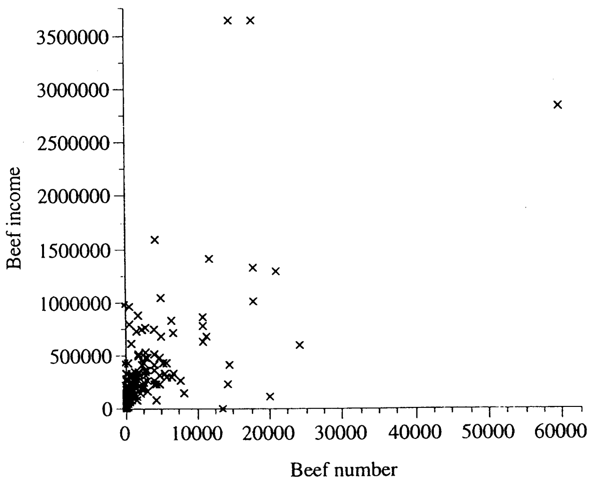

Fig. 4. Scatterplot of Beef income *vs* Beef number *for the study population*

data based on that collected from $N = 904$ broadacre and dairy farms that participated in the annual Australian Agricultural and Grazing Industries Survey (AAGIS) and the annual Australian Dairy Industry Survey (ADIS) in the late 1980s. Both surveys are carried out by the Australian Bureau of Agricultural and Resource Economics. In the simulation study these 904 farms were taken as defining the population of interest.

Table 1 lists the variables that were assumed known for all of these farms (the frame variables), the variables whose population totals were assumed known, but whose nonsample values were assumed unknown (the benchmark variables) and the survey variables (values only known for the sampled farms). Figures 3 to 7 are scatterplots showing the relationships between each survey variable and its corresponding benchmark variable for these 904 farms. Note the large heteroskedasticity for *Beef income*

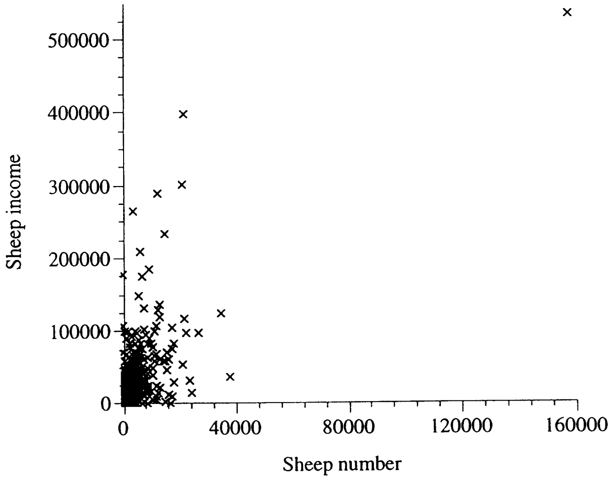

Fig. 5. Scatterplot of Sheep income *vs* Sheep number *for the study population*

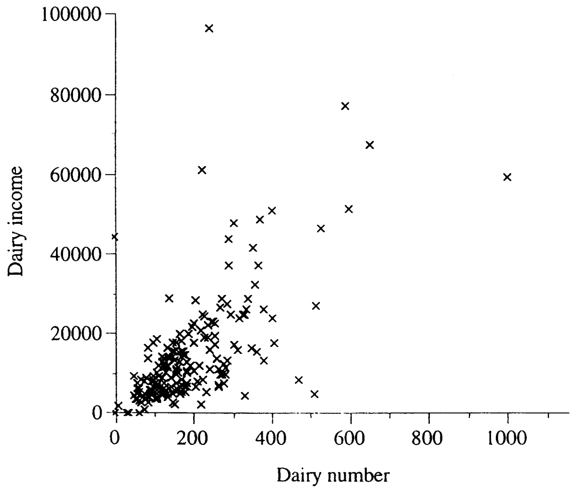

Fig. 6. Scatterplot of Dairy income *vs* Dairy number *for the study population*

and *Sheep income*, and the many outliers. Clearly, for this population the model (5), (6) is (at best) a rough approximation. Three different sample designs were investigated, each based on a total sample of $n = 100$ farms. Details of these designs are set out in Table 2. Each design was independently replicated 500 times, and, for each sample selected, a variety of weighting methods were used to estimate the population means for each of the survey variables. Details of these weighting methods (numbered from 1 to 11) are set out in Table 3. Note that the value of the ridge parameter λ used in the ridged weighting methods (RIDGE, NWD3 and NWDAR3, or methods 6–11) varied from sample to sample and was set just large enough to ensure that all sample weights were at least unity.

Two model specifications were used in case-weighting. The first, denoted by an "S" prefix (for Small model), included only the four production variables (*Wheat area*, *Beef number*, *Sheep number* and *Dairy number*) in the vector $\mathbf{X}_i$ in (5), together with

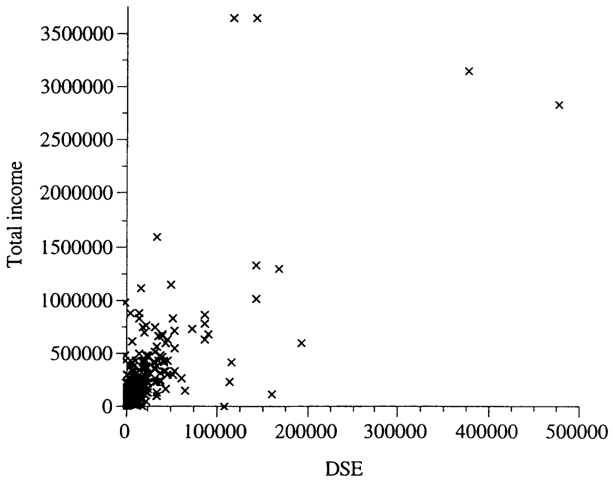

Fig. 7. Scatterplot of Total income *vs* DSE *for the study population*

Table 2. Sampling methods

<table>
<tr><th>Method</th><th>Description</th></tr>
<tr><td>Simple Random Sampling</td><td>Random sample of size $n = 100$ taken without replacement from $N = 904$. Sample rejected if missing one or more farms from each of the seven ASIC industries, or without production in one of the four farm outputs (wheat, sheep, beef or dairy).</td></tr>
<tr><td>Size Stratification with "Compromise" Allocation</td><td>Independent random samples taken from four size strata, defined by values of the size variable DSE. Stratum boundaries defined so that total DSE is approximately the same in each stratum. Stratum allocations defined by averaging proportional and Neyman allocation (based on DSE), resulting in the design:
<table>
<tr><th>Stratum</th><th>DSE Range</th><th>N_h</th><th>n_h</th></tr>
<tr><td>1</td><td>200–9499</td><td>665</td><td>50</td></tr>
<tr><td>2</td><td>9500–24999</td><td>166</td><td>25</td></tr>
<tr><td>3</td><td>25000–99999</td><td>52</td><td>18</td></tr>
<tr><td>4</td><td>100000+</td><td>12</td><td>7</td></tr>
</table>
Under this design, 9 farms with DSE < 200 were excluded from selection. In addition, a sample was rejected if missing one or more farms from each of the seven ASIC industries, or without production in one of the four farm outputs (wheat, sheep, beef or dairy).</td></tr>
<tr><td>Size Stratification with "Optimal" Allocation</td><td>Same stratification and sample rejection rule as for size stratification with "compromise" allocation, but with Neyman allocation based on DSE and with the "top" stratum completely enumerated.
<table>
<tr><th>Stratum</th><th>DSE Range</th><th>N_h</th><th>n_h</th></tr>
<tr><td>1</td><td>200–9499</td><td>665</td><td>30</td></tr>
<tr><td>2</td><td>9500–24999</td><td>166</td><td>29</td></tr>
<tr><td>3</td><td>25000–99999</td><td>52</td><td>29</td></tr>
<tr><td>4</td><td>100000+</td><td>12</td><td>12</td></tr>
</table></td></tr>
</table>

an overall intercept. The second, denoted by an "L" prefix (for Large model), also included these four production variables in X_i but replaced the overall intercept by industry specific intercepts defined by seven zero-one indicators for the seven ASIC industries represented in the population. In both models, DSE was used as the heteroskedasticity factor D_i in (6).

Two of the weighting methods investigated (GREG and BLUP weighting, methods 2–5 in Table 3) allow negative u-weights (i.e., case-weights less than one). Table 4 sets out the percentages of samples that recorded negative u-weights for these procedures, as well as the average number of sample units with negative u-weights in samples with at least one negative u-weight. These results confirm the statement made earlier that

Table 3. Weighting methods

Name	Number	Description
RATIO	1	π-weighted ratio estimator (2) with estimation benchmarks as follows $Y =$ Wheat income $X =$ Wheat area $Y =$ Beef income $X =$ Beef number $Y =$ Sheep income $X =$ Sheep number $Y =$ Dairy income $X =$ Dairy number $Y =$ Total income $X =$ DSE
S/GREG	2	GREG case-weights (9, $g_i = 1/\pi_i$) based on model "S"
S/BLUP	4	BLUP case-weights (9, $g_i = 1$) based on model "S"
S/RIDGE	6	Ridged BLUP case-weights (11, $g_i = 1$) based on model "S" with $C_k = 1000$ for each of the four production benchmarks in the model. Weights normalised to sum to N
S/NWD3	8	Nonparametrically corrected and ridged BLUP weights (13) based on model "S" with Nadaraya-Watson smoothing ($c = 3$ in (14)) against $Z =$ DSE. Same C-values as S/RIDGE. Weights normalised to sum to N
S/NWDAR3	10	Nonparametrically corrected and ridged BLUP weights (13) based on model "S" with Nadaraya-Watson smoothing ($c = 3$ in (14)) against $Z_1 =$ DSE, $Z_2 =$ ASIC and $Z_3 =$ Region. Same C-values as S/RIDGE. Weights normalised to sum to N
L/GREG	3	GREG case-weights (9, $g_i = 1/\pi_i$) based on model "L"
L/BLUP	5	BLUP case-weights (9, $g_i = 1$) based on model "L"
L/RIDGE	7	Ridged BLUP case-weights (11, $g_i = 1$) based on model "L" with $C_k = 1000$ for each of the four production benchmarks in the model, and $C_k = 100000$ for each of the seven industry benchmarks in the model. Weights normalised to sum to N
L/NWD3	9	Nonparametrically corrected and ridged BLUP weights (13) based on model "L" with Nadaraya-Watson smoothing ($c = 3$ in (14)) against $Z =$ DSE. Same C-values as L/RIDGE. Weights normalised to sum to N
L/NWDAR3	11	Nonparametrically corrected and ridged BLUP weights (13) based on model "L" with Nadaraya-Watson smoothing ($c = 3$ in (14)) against $Z_1 =$ DSE, $Z_2 =$ ASIC and $Z_3 =$ Region. Same C-values as L/RIDGE. Weights normalised to sum to N

Table 4. Percentages of samples that generate negative weights under various weighting systems/sample design combinations. Numbers in parentheses are the average number of sample units with a negative weight in samples containing at least one negative weight

	Simple Random Sampling	Size Stratification and "Compromise" Allocation	Size Stratification and "Optimal" Allocation
S/GREG	44 (4.58)	11 (1.38)	82 (7.59)
L/GREG	77 (4.27)	53 (1.83)	94 (9.73)
S/BLUP	44 (4.58)	6 (1.33)	48 (1.81)
L/BLUP	77 (4.27)	20 (1.83)	93 (5.87)

GREG weighting is more susceptible to negative weights, especially in samples that are heavily "size biased" (such as those generated under size stratification with optimal allocation). The large number of simple random samples that recorded negative u-weights reflects the high variability in this sample design. Note that GREG and BLUP weighting coincide for simple random sampling. The compromise allocation design is least affected by negative weights, especially when the S(mall) model is used to generate the weights. This reflects the fact that this model imposes fewer calibration constraints on the weighting process, and consequently results in fewer negative u-weights.

For each sample, survey estimates were computed based on the weighting methods defined in Table 3 and the Root Mean Squared Error (RMSE) of each method, expressed as a percentage of the true population value, calculated over the 500 independent samples selected under each design. These results are displayed in Tables 5 to 7. Table 5 displays the RMSE values for population estimates of the mean values of the survey variables, while Tables 6 and 7 show the RMSE values generated when domain estimates of mean *Total income* (State estimates in Table 6 and ASIC Industry estimates in Table 7) are computed by appropriately summing case-weighted survey data over the domains of interest. In the case of the RATIO strategy, these domain estimates were computed by first converting the ratio estimator (2) to case-weighted form and then summing over the domain of interest. In addition, Figures 8–10 show the RMSE values for each weighting method superimposed on a "skeletal" boxplot of the distribution of these values (a skeletal boxplot is one where a central box shows the range of the data between the lower and upper quartiles, with the median drawn as a line in this central box, and the 10 and 90 percentiles of the data are shown as lines or "fences" below and above this central box).

Inspection of the results in Table 5 (and the corresponding display in Figure 8) shows that the methods (8–11) based on the bias corrected ridge weighting procedure (13) have much to recommend them. In all cases this procedure results in estimates whose RMSE is acceptably close to the best RMSE values observed. In no cases does this procedure result in a worst RMSE. By and large, the RMSEs observed under model S are somewhat smaller than those observed under model L, reflecting the extra cost in attempting to meet (or at least minimise deviations from) the extra constraints implied by the extra terms in this model. In contrast, the standard RIDGE

Table 5. Root Mean Squared Errors (expressed as a percentage of the corresponding population value) of estimates of the population means of the survey variables

	Wheat income	Beef income	Sheep income	Dairy income	Total income
Simple Random Sampling					
1. RATIO	14.7	28.9	19.1	14.4	16.7
2. S/GREG	14.0	27.4	17.2	15.6	17.8
3. L/GREG	13.6	26.1	17.0	15.0	17.3
4. S/BLUP	14.0	27.4	17.2	15.6	17.8
5. L/BLUP	13.6	26.1	17.0	15.0	17.3
6. S/RIDGE	15.8	24.2	16.3	20.4	15.8
7. L/RIDGE	15.7	23.6	16.0	17.1	15.7
8. S/NWD3	15.1	22.2	16.1	18.1	14.5
9. L/NWD3	15.0	22.1	15.9	17.5	14.6
10. S/NWDAR3	14.4	22.6	15.9	17.3	14.7
11. L/NWDAR3	14.5	22.4	15.6	17.0	14.7
Size Stratification with "Compromise" Allocation					
1. RATIO	10.0	11.6	15.5	19.2	8.3
2. S/GREG	10.0	11.4	14.7	19.3	7.9
3. L/GREG	9.9	11.9	14.8	20.3	8.4
4. S/BLUP	10.8	14.5	14.6	25.0	10.1
5. L/BLUP	10.8	12.8	14.3	20.5	8.9
6. S/RIDGE	11.4	14.5	14.8	25.2	10.2
7. L/RIDGE	13.2	13.0	15.6	23.1	9.8
8. S/NWD3	10.1	11.8	13.9	19.6	8.1
9. L/NWD3	10.5	11.5	14.1	19.8	8.1
10. S/NWDAR3	9.9	12.1	13.8	19.9	8.2
11. L/NWDAR3	10.5	11.6	14.1	19.7	8.1
Size Stratification with "Optimal" Allocation					
1. RATIO	10.1	10.1	15.9	25.7	7.9
2. S/GREG	10.2	10.3	15.6	26.8	7.4
3. L/GREG	11.6	11.6	17.4	32.3	8.4
4. S/BLUP	9.1	11.1	14.8	34.2	8.3
5. L/BLUP	11.9	11.1	16.4	32.1	8.0
6. S/RIDGE	12.6	10.7	15.7	37.2	8.7
7. L/RIDGE	23.5	9.6	21.3	47.8	11.9
8. S/NWD3	11.5	9.8	14.3	29.2	7.4
9. L/NWD3	12.5	9.1	15.6	30.7	7.3
10. S/NWDAR3	11.5	9.6	14.4	29.7	7.2
11. L/NWDAR3	12.9	8.9	15.7	31.5	7.3

procedure (methods 6–7) performs poorly, especially in size-biased samples, due to biases incurred in forcing the BLUP case-weights to be all positive. The BLUP and GREG weighting methods (both of which can result in negative weights) and the RATIO method (which is not a case-weighting approach) all seem to perform on a par, with the S/GREG weights (method 2) marginally the best of this group.

The results displayed in Tables 6 and 7 (and the accompanying boxplots in Figures 9 and 10) provide a perspective on how case-weighting methods cope with domain estimation, a standard form of secondary analysis carried out on survey data bases.

Table 6. Root Mean Squared Errors (expressed as a percentage of the corresponding population value) of estimates of the mean of Total income within each State

	NSW	VIC	QLD	SA	WA	TAS	NT
Simple Random Sampling							
1. RATIO	48.1	55.3	44.1	43.3	50.7	65.7	77.1
2. S/GREG	49.1	56.0	50.9	41.7	42.2	70.8	68.2
3. L/GREG	49.1	55.5	51.5	42.3	40.6	72.2	67.9
4. S/BLUP	49.1	56.0	50.9	41.7	42.2	70.8	68.2
5. L/BLUP	49.1	55.5	51.5	42.3	40.6	72.2	67.9
6. S/RIDGE	45.4	51.6	45.7	39.2	44.2	67.2	67.0
7. L/RIDGE	45.3	50.8	45.0	39.1	45.1	66.7	66.6
8. S/NWD3	42.9	54.6	42.5	37.8	39.9	66.7	64.2
9. L/NWD3	43.5	54.2	42.5	38.0	40.1	66.2	64.5
10. S/NWDAR3	43.0	50.7	43.4	37.8	39.0	63.6	65.6
11. L/NWDAR3	43.7	50.8	43.1	38.0	39.3	63.3	65.6
Size Stratification with "Compromise" Allocation							
1. RATIO	25.5	42.0	23.2	31.1	26.6	48.8	31.4
2. S/GREG	25.5	42.6	22.5	31.5	25.1	48.9	28.8
3. L/GREG	27.6	45.4	24.3	32.1	25.5	52.3	28.4
4. S/BLUP	25.6	35.7	26.3	29.9	35.8	44.1	32.0
5. L/BLUP	26.9	38.7	25.8	30.2	33.3	52.7	29.6
6. S/RIDGE	25.8	35.7	26.3	29.9	36.1	44.2	32.1
7. L/RIDGE	27.6	38.2	25.5	30.0	39.1	51.8	32.2
8. S/NWD3	24.4	36.8	23.0	29.2	28.1	47.0	29.8
9. L/NWD3	26.5	39.1	22.9	29.7	28.6	51.3	29.5
10. S/NWDAR3	24.3	35.8	23.2	29.2	28.3	45.3	30.4
11. L/NWDAR3	26.5	38.6	23.0	29.7	28.6	50.2	29.8
Size Stratification with "Optimal" Allocation							
1. RATIO	24.3	49.6	22.4	36.0	21.6	50.1	14.2
2. S/GREG	24.4	49.8	20.6	35.9	21.7	50.9	14.8
3. L/GREG	31.5	54.8	22.4	39.8	24.0	60.4	16.7
4. S/BLUP	18.3	41.7	19.8	30.2	23.7	47.4	13.8
5. L/BLUP	26.0	47.1	22.2	34.6	28.3	63.1	19.9
6. S/RIDGE	18.4	41.1	19.5	30.4	34.0	47.7	14.8
7. L/RIDGE	21.1	42.0	17.3	31.9	58.5	53.1	22.3
8. S/NWD3	19.5	42.5	19.4	31.6	24.8	53.5	14.3
9. L/NWD3	21.3	43.3	18.6	33.1	26.9	55.3	15.5
10. S/NWDAR3	18.7	41.5	19.2	31.6	25.6	51.5	14.7
11. L/NWDAR3	20.9	42.6	18.3	33.1	28.2	54.1	15.9

In Table 6 the domains are defined by splitting the farms making up the study population according to the State in which they are located, while in Table 7 these domains are defined according to the ASIC industry classification (see Table 1) of these farms.

Again, we see that the nonparametrically adjusted ridge weights (8–11) perform well, with the weights based on the S model (methods 8 and 10) providing the more stable and consistent estimation performance across the different domains. In the case of ASIC domains (Table 7/Figure 10) the weights based on the *L* model perform well, which is not surprising since this model includes calibration on ASIC counts. Note that here, unlike the case with the population level results in Table 5,

Table 7. Root Mean Squared Errors (expressed as a percentage of the corresponding population value) of estimates of the mean of Total income within each ASIC industry

	181	182	183	184	185	186	187
Simple Random Sampling							
1. RATIO	41.6	47.5	84.2	55.5	42.6	35.2	48.3
2. S/GREG	36.2	34.2	97.6	60.5	34.8	35.3	33.1
3. L/GREG	33.9	31.6	89.8	52.3	29.2	33.8	34.5
4. S/BLUP	36.2	34.2	97.6	60.5	34.8	35.3	33.1
5. L/BLUP	33.9	31.6	89.8	52.3	29.2	33.8	34.5
6. S/RIDGE	34.3	35.9	89.7	54.0	32.6	32.0	31.4
7. L/RIDGE	32.4	35.8	86.3	49.1	29.1	31.0	30.9
8. S/NWD3	34.3	32.9	94.2	53.3	31.2	30.0	32.9
9. L/NWD3	33.2	32.1	92.6	50.2	28.5	29.3	32.6
10. S/NWDAR3	33.7	31.4	84.0	49.1	30.9	30.6	30.4
11. L/NWDAR3	32.5	30.8	84.9	46.5	28.3	29.9	30.3
Size Stratification with "Compromise" Allocation							
1. RATIO	38.1	26.1	67.2	38.5	31.0	16.4	42.5
2. S/GREG	34.4	22.5	67.9	37.5	29.0	13.7	35.4
3. L/GREG	34.3	19.6	79.2	33.0	24.1	13.3	35.9
4. S/BLUP	32.4	32.6	54.2	34.2	31.5	16.1	27.0
5. L/BLUP	36.3	29.3	74.8	33.1	27.0	15.0	27.9
6. S/RIDGE	32.4	32.9	54.3	34.6	31.5	16.1	27.2
7. L/RIDGE	33.8	34.9	71.4	34.0	28.4	15.4	29.1
8. S/NWD3	31.0	25.9	56.7	34.5	29.3	14.1	28.3
9. L/NWD3	28.8	23.8	69.7	31.2	25.7	13.7	29.1
10. S/NWDAR3	31.1	26.5	54.2	33.3	29.3	13.9	27.4
11. L/NWDAR3	28.9	24.1	68.6	30.4	25.7	13.7	28.3
Size Stratification with "Optimal" Allocation							
1. RATIO	44.3	20.3	72.5	37.4	33.8	13.1	55.7
2. S/GREG	40.6	18.5	72.0	37.3	31.1	10.9	46.6
3. L/GREG	37.9	19.9	106.5	40.6	30.0	11.1	52.0
4. S/BLUP	34.4	22.4	54.7	29.2	30.6	11.5	31.5
5. L/BLUP	41.5	30.6	99.8	46.5	33.4	12.2	39.3
6. S/RIDGE	33.7	31.7	54.4	30.5	30.8	11.0	33.5
7. L/RIDGE	32.5	56.7	61.6	28.6	34.9	10.3	43.5
8. S/NWD3	34.8	23.0	59.4	34.6	30.1	9.8	33.4
9. L/NWD3	30.3	24.0	59.9	31.5	30.2	8.9	34.7
10. S/NWDAR3	34.7	24.9	56.1	32.3	30.2	9.3	32.7
11. L/NWDAR3	30.2	26.2	58.1	29.5	30.2	8.6	34.7

the RIDGE case-weights (methods 6 and 7) seem to perform reasonably well in domain estimation. The remaining methods (1–5) do not admit of a clear winner as far as domain estimation is concerned. The L/GREG procedure (method 3) works well with ASIC domains, but is unremarkable with State domains. On balance, the S/BLUP procedure (method 4) would appear to be the best of this group of conventional case-weighting methods.

Finally, in Figures 11 to 16 we illustrate the conditional behaviour of the different case-weighting methods investigated in the study. Figures 11–13 are scatterplot smooths showing how the average number of negative u-weights (under GREG

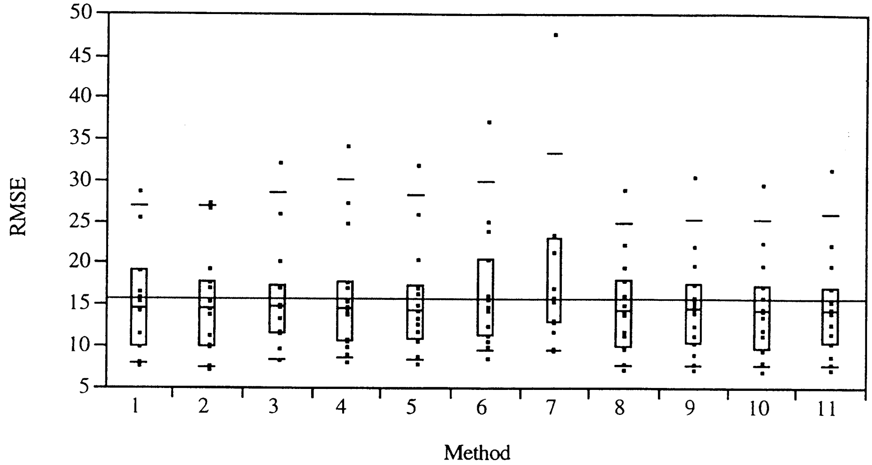

Fig. 8. Boxplots showing the distribution of RMSE's associated with estimation of Wheat income, Beef income, Sheep income, Dairy income *and* Total income *for the three sample designs (simple random sampling, "compromise" stratified sampling and "optimal" stratified sampling) shown in Table 5. A separate boxplot is shown for each of the weighting methods 1 to 11 considered in the study*

and BLUP weighting) in a sample changes as the sample configuration changes. For this analysis, sample configuration has been defined in terms of the sample DSE-rank. That is, the rank, over the 500 samples drawn for each design, of the sum of the population ranks of the sample DSE values. The sample DSE-rank provides a robust measure of the overall size of the sample, since it is relatively unaffected by a few sample units with very large DSE values. Similarly, Figures 14–16 are scatterplot smooths that show how the average estimation error (expressed as a percentage

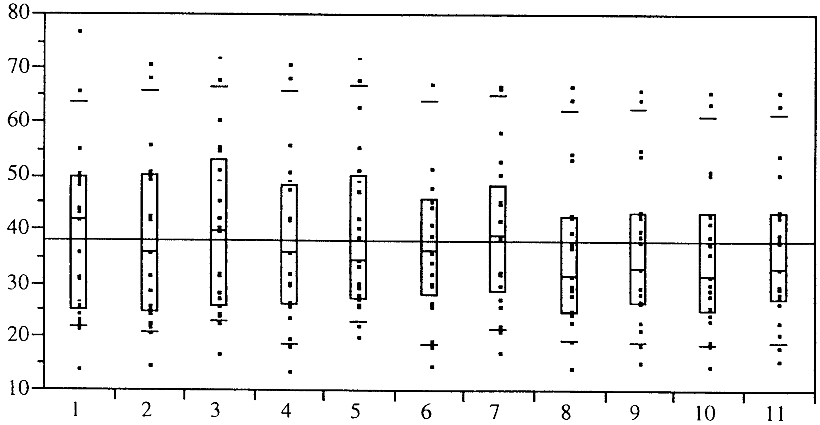

Fig. 9. Boxplots showing the distribution of RMSE's associated with estimation of Total income *in each State/Territory for the three sample designs (simple random sampling, "compromise" stratified sampling and "optimal" stratified sampling) shown in Table 6. A separate boxplot is shown for each of the weighting methods 1 to 11 considered in the study*

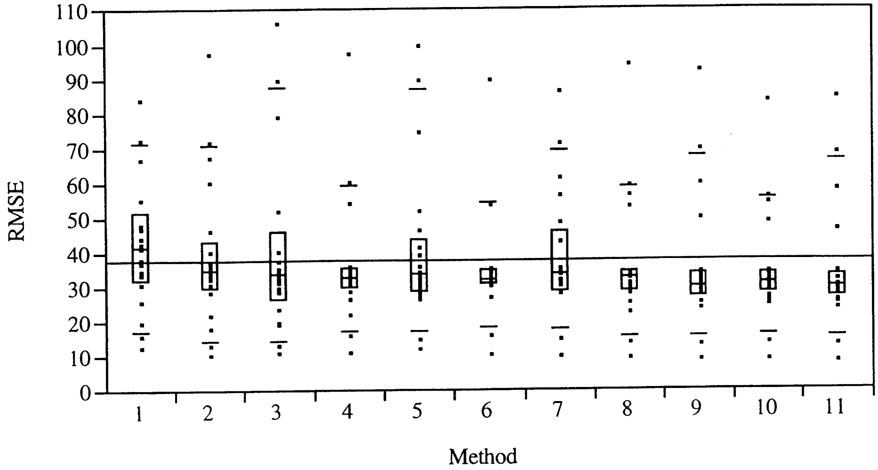

Fig. 10. Boxplots showing the distribution of RMSE's associated with estimation of Total income *in each industry group (ASIC) for the three sample designs (simple random sampling, "compromise" stratified sampling and "optimal" stratified sampling) shown in Table 7. A separate boxplot is shown for each of the weighting methods 1 to 11 considered in the study*

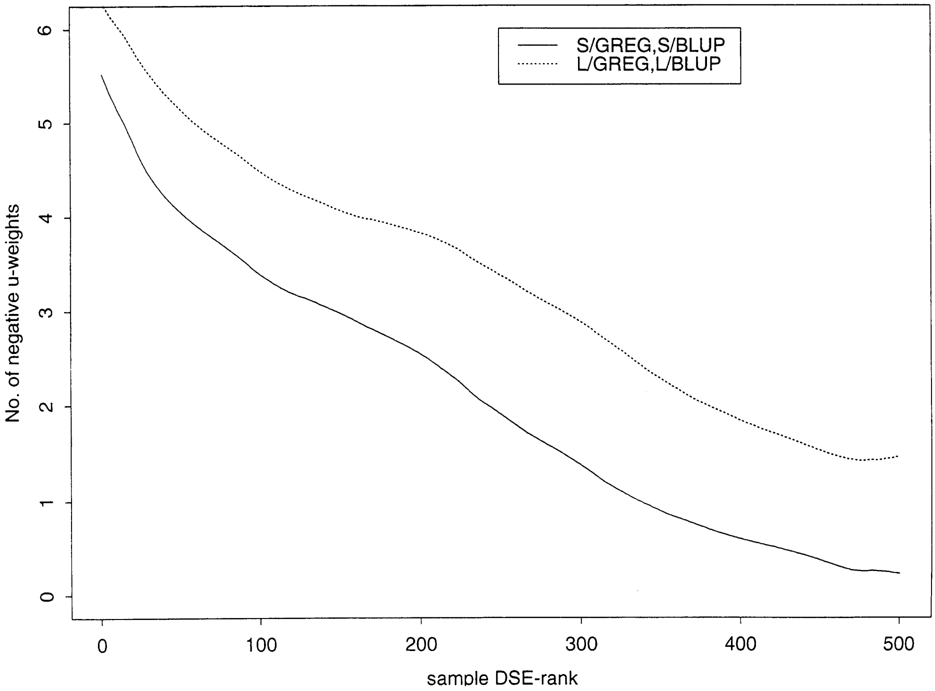

Fig. 11. Scatterplot smooths showing how the average number of negative u-weights *changes as the sample DSE-rank increases for GREG and BLUP-type case-weighting under simple random sampling*

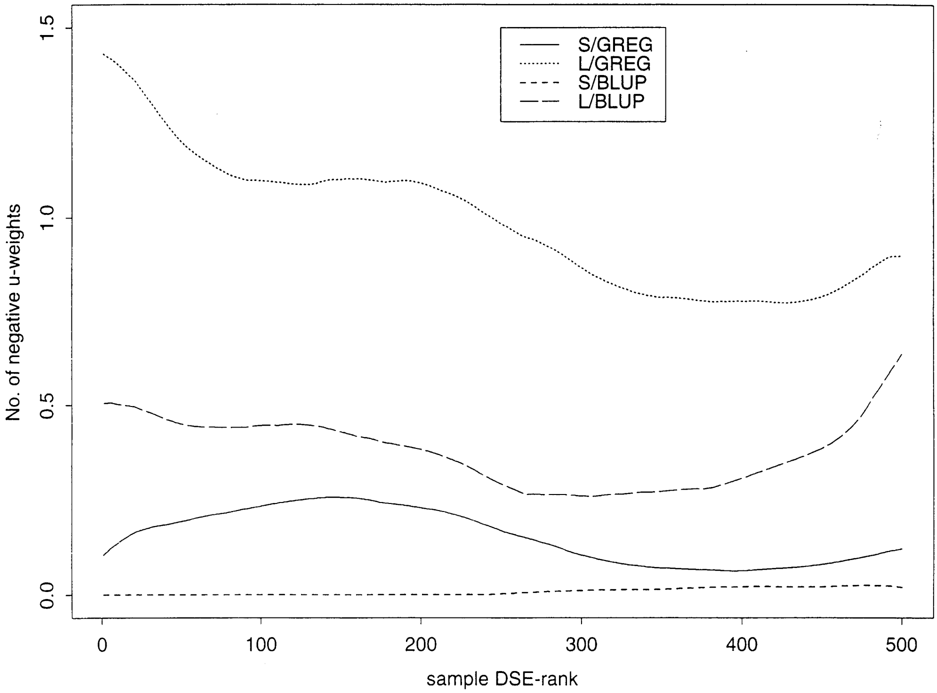

Fig. 12. Scatterplot smooths showing how the average number of negative u-*weights changes as the sample DSE-rank increases for GREG and BLUP-type case-weighting under "compromise" stratified sampling*

of the population total) for the variable Total income changes with sample DSE-rank.

Clearly, sample design is the main factor dictating the number of negative u-weights that occur in a particular sample, with the actual method (either GREG or BLUP) of case-weighting being of secondary concern. From Figure 11 we see that, for simple random sampling, most negative u-weights occur in samples with small DSE-rank, with the propensity for negative u-weights rapidly decreasing as sample DSE-rank increases. In contrast, with the stratified sampling/optimal allocation design, the reverse occurs, with the propensity for negative u-weights increasing with increased sample DSE-rank (Figure 13). Finally, we see that with the stratified sampling/compromise allocation design, this propensity seems relatively unaffected by sample DSE-rank (Figure 12). These plots also confirm the point made earlier – that GREG-weighting tends to produce more negative u-weights than BLUP-weighting, with the L(arge) model resulting in more negative u-weights than the S(mall) model.

Figures 14–16 show how the behaviour of the estimation error of the variable *Total income* under different case-weighting methods (all based on the S(mall) model) changes depending on the sample DSE-rank. In the case of simple random sampling (Figure 14) we see that both the S/RIDGE and S/NWD3 methods perform well provided sample DSE-rank is greater than about 200, but in samples with small DSE-rank these methods are biased low (S/RIDGE more than S/NWD3). On the other hand, both the RATIO and S/GREG (which is the same as S/BLUP for this

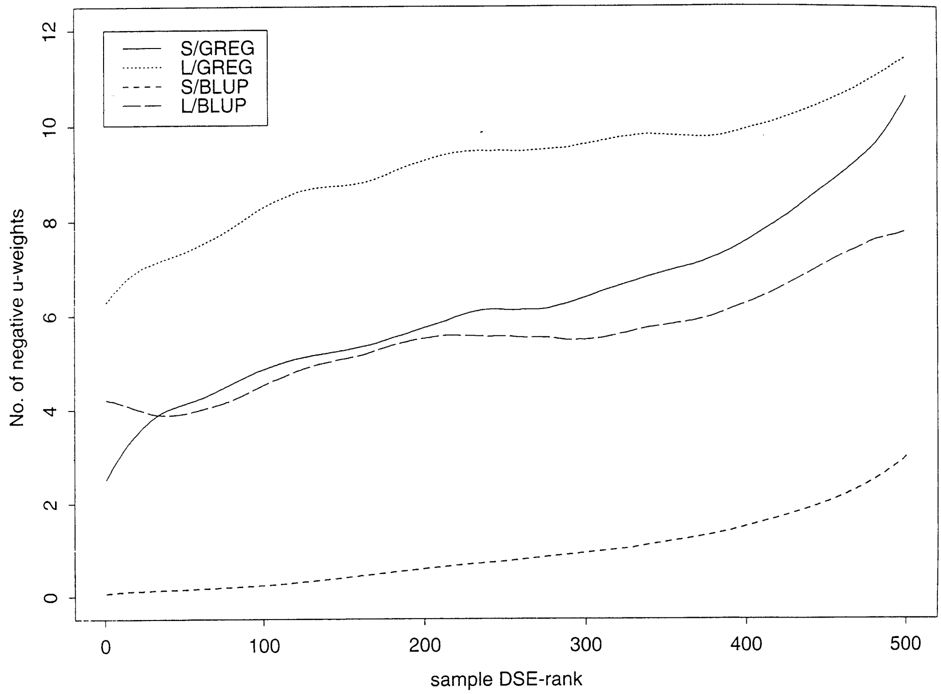

Fig. 13. Scatterplot smooths showing how the average number of negative u*-weights changes as the sample DSE-rank increases for GREG and BLUP-type case-weighting under "optimal" stratified sampling*

design) methods show a downward trend in estimation error, from being biased high in small DSE-rank samples to being biased low in high DSE-rank samples (RIDGE) or with negligible bias in these samples (GREG/BLUP). Overall, for samples that are reasonably DSE-balanced (sample DSE-rank between 200 and 300) we see that the S/RIDGE and S/NWD3 weights are preferable. It is worth pointing out here that Figure 14 shows that in fact for samples of the size taken ($n = 100$) and for the population being studied here (*Total income*) use of either S/GREG or RATIO does not lead to an average bias (i.e., a design bias) of zero. In fact, the average bias of S/GREG in Figure 14 is 3.3%, while that for RATIO is 1.1%. The corresponding average biases for S/RIDGE and S/NWD3 are –0.9% and –0.7%, respectively.

The situation changes when we consider the stratified sampling designs. For the compromise allocation design (Figure 15) we see that both S/BLUP and S/RIDGE tend to be biased low across the entire range of samples selected. The nonparametrically adjusted ridge weights S/NWD3 are also biased low, but to a much lesser extent. The RATIO and S/GREG weights have negligible conditional bias. All weighting methods exhibit a slight upward trend in conditional bias as sample DSE-rank increases. For the optimal allocation design (Figure 16) the S/BLUP is still biased low, but now we see a strong upward trend in the conditional bias of S/RIDGE, while the nonparametrically adjusted ridge weights S/NWD3 are now essentially unbiased, as are the RATIO and S/GREG weights.

Before concluding this section, it should be pointed out that the preceding conditional analysis illustrates the behaviour of the conditional *bias* of the various

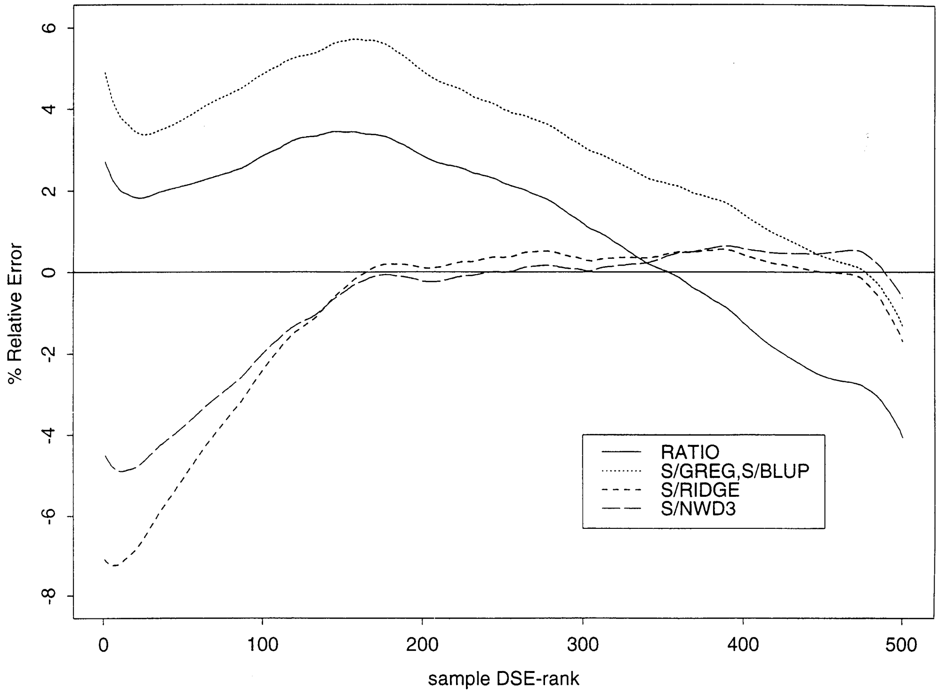

Fig. 14. Scatterplot smooths showing how the percentage relative error in estimating Total income *changes as the sample DSE-rank increases for various methods of weighting under simple random sampling*

weighting methods considered. It does not show how the conditional *variance* of the estimation error associated with these methods changes with sample DSE-rank. The change in variability of the estimation error as sample DSE-rank changes can be assessed along similar lines to that used above to assess the change in bias. Although not presented here, this analysis shows that in the stratified samples the S/BLUP weights generally have the lowest conditional variability, compensating for their conditional bias, while the S/RIDGE weights have low conditional variability in the compromise allocation design, but not in the optimal allocation design. The nonparametrically adjusted ridge weights S/NWD3 have low conditional variability across all three designs considered in the study, outperforming RATIO and S/GREG in this regard in both the simple random sampling design and the optimal allocation design, and with similar performance to S/GREG (and outperforming RATIO) in the compromise allocation design.

7. Conclusion

In this article two different ideas applicable in model-based survey estimation have been combined to produce a procedure that seems to offer the best qualities of both. The first idea is that of ridging to avoid negative case-weights when calibrating on a set of benchmark variables. This is effective but is also model-dependent. The second idea is that of nonparametrically adjusting survey weights to correct for model misspecification. Again, this is effective, but is subject to the occurrence of negative

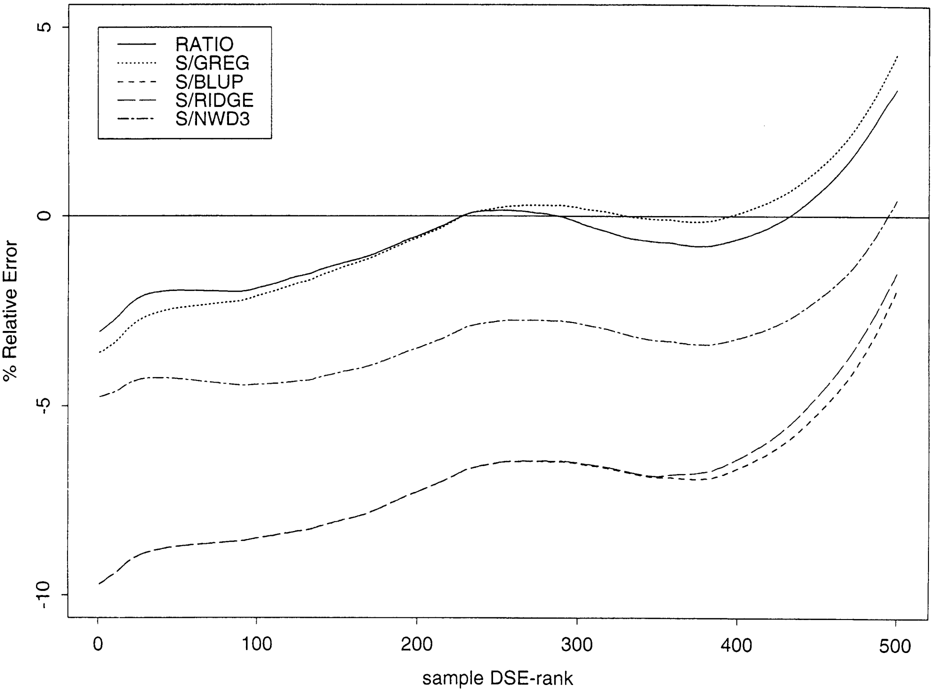

Fig. 15. Scatterplot smooths showing how the percentage relative error in estimating Total income *changes as the sample DSE-rank increases for various methods of weighting under "compromise" stratified sampling*

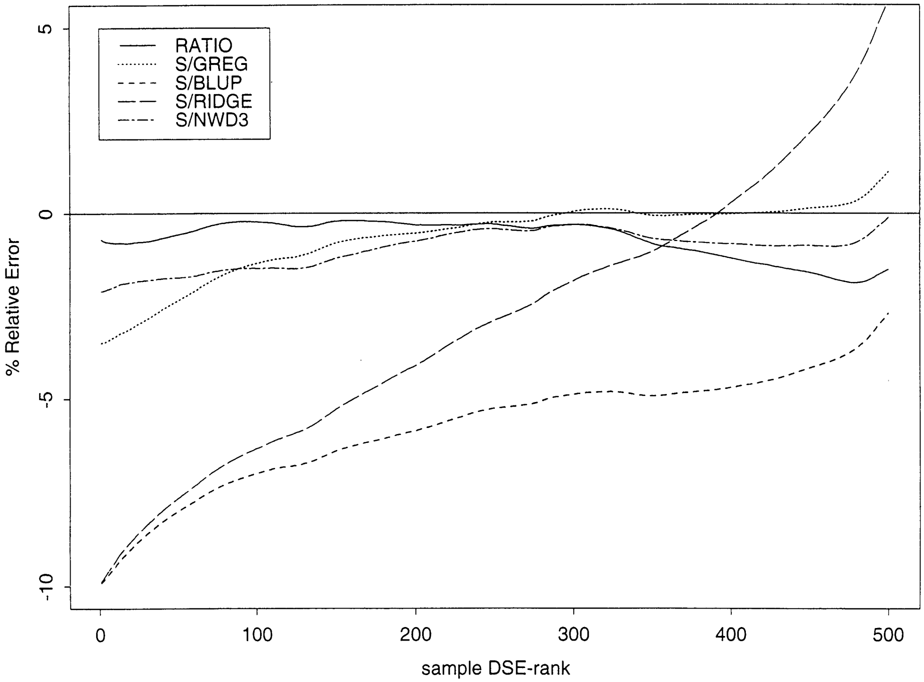

Fig. 16. Scatterplot smooths showing how the percentage relative error in estimating Total income *changes as the sample DSE-rank increases for various methods of weighting under "optimal" stratified sampling*

weights. Putting these two ideas together in (13) offers the promise of a method of case-weighting that should work well in a wide variety of situations. Certainly, the empirical results demonstrated in the previous section show that for those establishment type surveys where economic data are collected, and where positive case-weights that are (at least approximately) calibrated on a key set of benchmark variables are required, a method of weighting based on (13) should work rather well.

An issue that has not been addressed at all in this article is that of confidence interval estimation. In the context of ridge-type weighting, this problem has been considered by Dunstan and Chambers (1986). The basic idea is that the prediction variance of a case-weighted estimator like (4) can be decomposed into a term which depends on the squares of the weights and the underlying population variance function and another term corresponding to a squared bias. Both terms can be estimated (using, for example, the robust variance estimation procedures described in Royall and Cumberland, 1978) and standard 2-sigma type confidence intervals constructed. Alternatively, modern bootstrap ideas (Chambers and Dorfman 1994) may be applied to construct these confidence intervals. Research in this area is continuing.

8. References

Bankier, M.D., Rathwell, S., and Majkowski, M. (1992). Two Step Generalized Least Squares Estimation in the 1991 Canadian Census. Proceedings of the Workshop on Uses of Auxiliary Information in Surveys, Statistics Sweden, Örebro, October 5–7.

Bardsley, P. and Chambers, R.L. (1984). Multipurpose Estimation from Unbalanced Samples. Applied Statistics 33, 290–299.

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

Chambers, R.L., Dorfman, A.H., and Wehrly, T.E. (1993). Bias Robust Estimation in Finite Populations Using Nonparametric Calibration. Journal of the American Statistical Association 88, 260–269.

Chambers, R.L. and Dorfman, A.H. (1994). Robust Sample Survey Inference via Bootstrapping and Bias Correction: The Case of the Ratio Estimator. Invited Paper, Joint Meetings of the American Statistical Association, Toronto, August 14–18.

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

Dunstan, R. and Chambers, R.L. (1986). Model-Based Confidence Intervals in Multipurpose Surveys. Applied Statistics 35, 276–280.

Fuller, W.A., Loughin, W.W., and Baker, H.D. (1994). Regression Weighting in the Presence of Nonresponse with Application to the 1987–88 Nationwide Food Consumption Survey. Survey Methodology 20, 75–85.

Huang, E.T. and Fuller, W.A. (1978). Nonnegative Regression Estimation for Survey Data. Proceedings of the Social Statistics Section, American Statistical Association, 300–303.

Royall, R.M. (1976a). Current Advances in Sampling Theory: Implications for Human Observational Studies. American Journal of Epidemiology 104, 463–474.

Royall, R.M. (1976b). The Linear Least Squares Prediction Approach to Two-Stage Sampling. Journal of the American Statistical Association 71, 657–664.

Royall, R.M. and Cumberland, W.G. (1978). Variance Estimation in Finite Population Sampling. Journal of the American Statistical Association 73, 351–358.

Särndal, C.E. and Wright, R.L. (1984). Cosmetic Form of Estimators in Survey Sampling. Scandinavian Journal of Statistics 11, 146–156.

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

Received September 1994
Revised September 1995