Methodological Issues Associated with Using BAS Data for Improving Sample Design and Estimation for Business Surveys

Glenys Bishop

Statistical Services Branch
Australian Bureau of Statistics
PO Box 10
Belconnen ACT 2616

Email: glenys.bishop@abs.gov.au

Tel: (02) 6252 7084
Fax: (02) 6252 8015

Executive Summary

Under the new taxation system, there will be a quarterly census of businesses in the form of Business Activity Statements (BAS) returned to the Australian Taxation Office. The data collected on these statements will be made available to the Australian Bureau of Statistics, which has made a corporate decision to use the data to effect gains in terms of reduced sample size, reduced provider load and reduced survey costs. The ways in which these gains may be achieved have been identified but only some are relevant to this paper. One is to redesign business surveys by replacing our existing size stratification with a new one based on BAS variables. Another is to improve our estimation procedures by introducing the use of generalised regression estimation with BAS variables as auxiliaries. The third is to use BAS data as a substitute for currently collected survey data.

In this paper I examine each of these ways of achieving gains by considering an approach that will require the minimum amount of change to our current procedures and then a more comprehensive approach. In the case of estimation the minimum approach is to use ratio estimation but with new auxiliary variables while the more comprehensive approach involves implementing generalised regression. Finally, I discuss how the three ways of achieving gains will interact. Some of the issues that have arisen are theoretical and some are practical. They are listed in the Appendix at the end of the paper.

1. Introduction

The aim of this paper is to examine some of the methodological issues that have arisen in the economic surveys of the Australian Bureau of Statistics with the introduction of the new taxation system. The focus is on changes that the Bureau can make to its sample design and estimation procedures as a result of having access to a quarterly census of businesses conducted by the Australian Taxation Office. This census is in the form of a Business Activity Statement completed either monthly or quarterly by all businesses for taxation purposes. Another change introduced with the new system, but of lesser interest in this paper, will be the formation of the Australian Business Register, a public register of all businesses in Australia.

When the Business Activity Statement data become available, a number of extra variables will be provided regularly by the Australian Taxation Office. Some of these are similar to variables already collected by the ABS in its surveys. These are:

Business Activity Statement	ABS survey variable
G1 Total sales & income & other supplies	Turnover/output
G10 Capital acquisitions	Capital formation (expenditure)
G11 Other acquisitions	Current expenses
W1 Total of salary, wages & other payments	Wages & salaries

Other variables will also be available and include exports, estimated total fringe benefits tax payable, goods and services tax payable, wine equalisation tax payable and luxury car tax payable.

It is proposed that these variables, henceforth called BAS variables, may be used for data substitution and supplementation, i.e. replacing variables normally collected in a survey, for developing new survey designs and for new or improved methods of estimation. Ultimately, the goal is to use BAS data to gain the efficiencies of smaller sample sizes and reduced provider load.

The main methodological focus is on estimation methods, in particular generalised regression, imputation of missing or late BAS data and development of new survey designs. As a secondary consideration, using BAS data to impute missing survey variable values will also be investigated. Since the procedures for this will be straightforward, no discussion on imputation of survey variable values has been included in this paper.

To assess the proposed methods we must use BAS data. It is generally thought that the BAS data for the first quarter, i.e. September quarter 2000, will be unreliable but we expect it to be available for ABS use by the end of November 2000. Subsequent quarters' BAS data are expected at the end of February 2001, mid-May 2001 and at the end of August 2001. Some of the methods proposed in this paper require two successive quarters of BAS data for testing and so we have, somewhat arbitrarily, decided that decisions and recommendations about gains to the organisation can be made after receiving the fourth quarter BAS data. This gives us a chance to consider two testing sets, namely second and third quarters and then third and fourth quarters. Allowing time for the final analysis, the date we have set is September 30, 2001. A corporate decision has set the date for the introduction of the new methodologies at September quarter 2002 or July month 2002 for monthly surveys and at June 2002 for annual surveys.

In section 2 of this paper I discuss plans for the use of BAS data to modify stratification for economic surveys. I describe a minimum modification to the existing stratification, and a more comprehensive way of investigating BAS data for substantial changes to stratification methods. In section 3, I address changes to estimation methods, including a minimum modification to the existing ratio estimation and a methodology for incorporating generalised regression into our estimation procedures. In section 4, I describe initial plans for a methodology for imputing values for missing or late BAS data, while in section 5, I consider how these various findings may all be brought together to afford gains to the organisation.

2. Stratification

At present, the organisation uses synchronised sampling to control overlap between surveys and also to control rotation by guaranteeing a maximum time in sample. Some constraints are placed on stratification by the method of selection. Synchronised sampling allows for births and deaths on the frame and different rotation rates in different strata (McKenzie and Gross, 2000). It can be used for different stratifications in different surveys, but diversity in the stratification can lead to higher individual provider load (ibid.). Brewer, Gross and Lee (1999) also point out that synchronised sampling gives very nearly simple random sampling within strata but that it does not automatically control rotation or overlap when sampling units change strata.

The stratification variable used at present for a large number of surveys is employment size as recorded on the Business Register. There are several possible cut-offs that may be used to form strata and the most appropriate ones will be used for any particular survey. However, all surveys use a common cut-off of 20 employees as recorded on the register to enable synchronised sampling to work. Typically an economic survey uses state by industry by employment size stratification. The Economic Activity Survey, the annual economy wide survey, also uses incorporated/unincorporated as a stratification variable.

Employment size is entered on the Business Register at registration of a new business and is not updated thereafter. A comparison of the register value with reported employment size from a survey shows substantial mismatching. The BAS data, on the other hand, will be updated at least quarterly and so will not have this problem. We want to investigate stratification using one or more of the BAS variables in place of employment size. It is desirable that the same variable or combination of variables will be used for most surveys to make synchronised sampling easier to implement.

2.1 Minimum Approach

The minimalist approach would entail preserving current strata as much as possible. Such an approach would imply very few changes to the way that businesses rotate into and out of sample. Thus we would require a stratification variable that approximates employment size. The new Australian Business Register (ABR) will not record employment size. However, the version of the ABR available to the ABS will record number of payees as well as an employment indicator and a labour hire indicator at registration. One possibility is to use number of payees from the ABR. This variable will not be updated either, and so eventually it will suffer from the same problems, caused by lack of updating, as number of employees on the current Business Register. However, initially it will be up-to-date.

Another possibility is to find a BAS variable that represents number of employees reasonably well. The most likely contender is total salary, wages, and other payments, W1. This will have a seasonal component for many industries and so, for stratification purposes, it will be practical to find an annual average or total. Call this AW1. For each industry division, the problem is to find α_{IND} such that stratum changes are minimised when employment size is replaced by α_{IND}*AW1.

In order to use this approach we must decide how to value stratum changes. Possible options are to minimise the number of stratum changes in sample or to minimise the variance of movement estimates. We also need to consider options for sample allocation and rotation. We could preserve the stratum sample sizes or we could preserve the sampling fraction for each stratum, which would have the effect of minimising the rotation of continuing units, or we could minimise the effect of rotation. The last two options would be constrained by the current total sample size. A complicating factor in these decisions is that the industry stratification variable supplied by the Australian Taxation Office will be updated at the same time.

2.2 Comprehensive Approach

We can extend the minimum approach by assessing the suitability of other BAS variables, either singly or in combination, as stratification variables. For a particular survey we could use a single size stratification, or we could cross two or more BAS variables to form a rectangular stratification or we could use some function of two or more BAS variables. A single size stratification would be either a single BAS variable or a linear combination of BAS variables. A function of BAS variables may be obtained by crossing, say, two variables, and then collapsing some classes of one differentially within the classes of the other, to form a non-rectangular stratification.

In the following discussion, all of the possibilities described above will be referred to generally as stratification variables. For a particular survey, we must consider whether to use the same stratification for each industry division. Within an industry, correlations or some other measure of relationship between stratification variables and response variables will enable us to choose the most likely candidate for stratification. An issue is to decide which response variable(s) best represents the survey in the event that different response variables yield different results. The process is further complicated for some surveys by the additional use of state and incorporated/unincorporated (type) for stratification. We could treat state and type in a similar way to industry division.

This approach has the disadvantage that a stratification variable may lead to substantial gains in some surveys but not in others. Our investigations may entail finding a stratification variable that works well for one survey and see how well it works for others. Or we could consider various stratification variables until we find an overall "best" stratification variable for all surveys. Alternatively, we could find stratification variables that work "best" for groups of surveys.

The important issue is how to decide whether a particular stratification works well. One criterion may be to see how the variances of the estimates change from one stratification to another. An alternative criterion may be the size of the reduction in variance achieved by adding an extra BAS variable to the stratification. Both of these criteria require some method of estimation and a method of sample allocation to be in place. The interaction between estimation and stratification using BAS data will be discussed in section 5.

An important part of selecting the stratification variable will be the choice of stratum boundaries. One option is to use a measure of relationship between response and stratification variables that allows for optimal probabilities of selection of units; this will tend to overstate the gains from the new stratification method, because, in practice, we will have sub-optimal allocation. The other option is to use an initial stratification and allocation combination that is thought to be reasonably good; this will tend to understate gains because the stratification and allocation are not fine-tuned to the response variable. The stratum boundaries will determine the completely enumerated sector and the strength of the relationship between response and stratification variables will determine the efficiency of this sector. A possible scenario is that there is a strong relationship for most units but with a non-negligible number of deviants. These deviants would then become the completely enumerated sector but only if they can be identified on the frame.

One more detail requiring attention is how to stratify births to the register when they have not yet submitted a BAS. This is dealt with in section 4.

2.3 Sticky Stratification

Because BAS variable values are updated regularly, a business unit may swap between strata. It is undesirable to have units swapping back and forth quarterly or even annually because of difficulties in managing rotation control. Rotation will occur because of changes in the probability of selection and because of deficiencies in the sample selection method. However, if a business unit has genuinely increased or decreased in size, then it should move to another stratum to decrease the standard error of a level estimate. We need a way of distinguishing short-term swapping from long-term changes.

Seasonality in trading may cause a business to swap strata twice or more a year. A way to counteract this problem would be to determine its stratum using an annual average or by deseasonalising the stratification variable data before determining the strata. We can make the stratification stick by placing, beyond the stratum bound, an extra barrier that must be crossed before a unit can change to that stratum. For instance, if there were two strata 0-19 and 20-50, we may require a unit in the lower stratum to increase its value to 25 before allowing it to change to the next stratum.

A method for determining these extra barriers is required. It must be robust because of the tight time constraints of quarterly surveys. Costs associated with sticky stratification include the incorrect initial allocation of a new unit to a stratum and the subsequent difficulty of moving it when more information is forthcoming, and increased standard errors of level estimates because of mis-stratification. A benefit of stickiness is that standard errors of movement estimates should be kept small because a common sample is maintained between time periods. There may be problems if we ignore industry changes that occur simultaneously and it is probably better not to use stickiness in this case.

3. Estimation

The only two methods of estimation currently supported by the organisation are expansion (Horvitz-Thompson) and ratio. These two methods can be applied to estimates of levels, rates and movements each using single phase and two-phase sampling.

Variables that are commonly used as auxiliaries for ratio estimation are obtainable from the Business Register, which is used as the sampling frame. Employment as recorded on the register is generally used. Employment is not updated and so ratio estimation could be improved by using the regularly updated BAS variables. Another example is the use of turnover for ratio estimation in the Manufacturing survey, where the turnover variable used is based on a census held every three to five years. The BAS variable,G1, total sales and income and other supplies, could be used instead.

The minimum approach to using BAS data in estimation is to use ratio estimation with BAS variables. The advantage of this approach is that no new theory would need to be implemented and generalised facilities for ratio estimation are already in place.

We could also improve our estimates by using generalised regression. This allows the inclusion of an intercept, thus removing a potential bias, and one or more auxiliary variables in the expression for the estimate. The auxiliary variables may enter the estimate in the form of population or class totals called benchmarks. Generalised Regression Estimation (GREG) is a well developed method of using multiple benchmarks and is used in the ABS by household surveys and by other national statistical agencies.

Whatever method of estimation we use, it must be robust. Because of the short time between obtaining survey data and publication of estimates for quarterly and monthly surveys, there is not time for expert intervention in refining the estimation procedure. There can be a standard set of diagnostics with alternative procedures to follow depending on those diagnostics, but there cannot be fine-tuning by a statistical methodologist or econometrician once the method is in production.

Estimation in business surveys must allow for three other sources of variation. The first is the fact that new businesses are forming all the time so that some new businesses may be in operation during part or all of the reference period of the survey but will not have been processed to the Business Register in time for sample selection. The New Business Provisions (NBP) estimate the net numbers of new businesses for each industry subdivision by size class (Gross, 1997). The effect of the new businesses on the estimate is via an adjustment that can be obtained by multiplying the NBP by the mean of the live respondents for each class. However, in practice it is calculated by adjusting the weights of the sampled units (*ibid*). The NBP has implications for estimating variances and is discussed below.

The second source of variation is dealing with outliers, either by winsorisation or by treating them as surprise outliers, while the third is that survey results are sometimes missing for businesses so that values must be imputed for them. The issue for the third source is how to treat imputed values in variance estimation; for instance when we are setting the sample estimate to agree with population benchmarks, should we be using the full sample or the responding sample? The mean of all respondents that are deemed to be live, in the class of interest, is found and that mean is substituted for all missing survey values in the class. There are several other standard procedures.

In the past the organisation has used Taylor series expansions to approximate the variances of estimates and this was later changed to Jackknife estimation. Because of the complexity of the estimation procedures proposed and the extra estimates and sources of variation that must be included, it has been suggested that we might also consider using bootstrap variance estimation (Jon Rao, personal communication).

3.1 Generalised Regression Estimation

The following description is abridged from chapter 6 of Sarndal, Swensson and Wretman (1992). Suppose we have J auxiliary variables, $x_1,...,x_j,...x_J$. For the *k*th population element, define the vector $\mathbf{x_k} = (x_{1k}, ...,x_{jk},...x_{Jk})^T$. The survey or response variable, y, takes the value, y_k, for the *k*th element. Let U represent the population, s the sample and π_k the probability that unit k is selected in the sample. For a simple random sample of n units from a population of N units, π_k would be n/N for each unit. Other probability sampling schemes may be used, with appropriate definitions of π_k.

We observe $(y_k, \mathbf{x_k})$ for $k \in s$ and $\mathbf{x_k}$ is also known for $k \in U$-s. Suppose the population points, $\{(y_k, x_{1k}, ...x_{Jk}): k=1,...,N\}$ look as if they have been generated by the model, ξ, which has the following features:

i. y_1, ..., y_N are assumed to be realised values of independent random variables, $Y_1, ...,Y_N$.

ii. $E_\xi(Y_k) = \sum_{j=1}^{J} \beta_j x_{jk} (k = 1,\ldots N)$, the expected value with respect to ξ

iii. $V_\xi(Y_k) = \sigma_k^2 (k = 1,\ldots N)$, the variance with respect to ξ

where $\beta_1,...,\beta_J$ and $\sigma_1^2,\ldots\sigma_N^2$ are model parameters.

If a census were conducted, the weighted least squares estimator of $\boldsymbol{\beta}= (\beta_1,...,\beta_J)^T$ under the model, ξ, would be $\mathbf{B} = (B_1, ..., B_J)^T$

$$=\left(\sum_U \mathbf{x_k}\mathbf{x_k}'/\sigma_k^2\right)^{-1}\sum_U \mathbf{x_k} y_k/\sigma_k^2 .$$

By expressing $\mathbf{B = T^{-1}t}$**,** where

$$\mathbf{T} = \sum_U \mathbf{x_k}\mathbf{x_k}'/\sigma_k^2 \text{ and } \mathbf{t} = \sum_U \mathbf{x_k} y_k/\sigma_k^2$$

we can use the sample to find π estimators of **T** and **t** and so obtain an estimate of **B**.

$$\hat{\mathbf{B}} = \hat{\mathbf{T}}^{-1}\hat{\mathbf{t}} = \left(\sum_s \mathbf{x_k}\mathbf{x_k}'/\sigma_k^2\pi_k\right)^{-1}\sum_s \mathbf{x_k} y_k/\sigma_k^2\pi_k$$

There are several forms of the regression estimator for the population total of y. The simplest form is

$$\hat{Y}_{GREG} = \hat{Y}_\pi + \sum_{j=1}^{J}\hat{B}_j(X_{jT} - \hat{X}_{j\pi}) \text{ where}$$

$$\hat{Y}_\pi = \sum_s y_k/\pi_k,\ \hat{X}_{j\pi} = \sum_s x_{jk}/\pi_k \text{ and } X_{jT} = \sum_U x_{jk}$$

and with some algebra, this can be expressed as:

$$\hat{Y}_{GREG} = \sum_s g_{ks} y_k/\pi_k \text{ where}$$

$$g_{ks} = 1 + \left(\mathbf{X_T} - \hat{\mathbf{X}}_\pi\right)^T \hat{\mathbf{T}}^{-1}\mathbf{x_k}/\sigma_k^2$$

The total weight given to the observed value y_k is the product of the sampling weight, $1/\pi_k$ and the g weight, g_{ks}. The subscript, s, indicates that the g weights depend on the sample. Inference is design-based.

It has been shown that the above g-weights can be found by minimising a function defining the distance between the sampling weights and the g weights subject to a binding constraint involving auxiliary variable totals. The distance function is:

$$\sum_s \left(g_{ks} - 1/\pi_k\right)^2 / 1/\pi_k$$

and the binding constraint is:

$$\sum_{k\in s} g_{ks} x_{jk} = X_{jT},\ j = 1,\ldots J$$

GREG can be viewed as a method of adjusting initial weights, based on the probability of selection, to ensure that the benchmarks are estimated exactly, while minimising the distance between the initial and adjusted weights.

The simplest distance function for GREG has the risk of generating negative weights or weights with values less than 1 as a consequence of imposing too many constraints on the calibration. A way of dealing with this would be to set any weight that is less than 1 to 1 and then reiterate the estimation procedure; the distance function will no longer be Euclidean. Alternative distance functions are available but they may lead to unrealistic or extreme weights and may be more difficult to minimise (Deville and Sarndal, 1992). A choice needs to be made in the context of the sample designs used in business surveys, the benchmarks used, the level at which they are set and the time and expertise available when they are used in production.

Calibration using generalised regression could be to BAS benchmark totals, to population counts for strata and to publication cell levels. When calibrating it is not necessary to use all strata; marginal totals can be used instead. For instance, we could calibrate at State level and at industry level but not all combinations of State by industry. We could incorporate other variables to improve efficiency, such as the age of the unit on the Register or whether or not the unit is a recent birth, ignoring strata.

3.2 Incorporating New Business Provisions

Consider the population quantities, where Y_T is the total for register businesses and Y_{NBP} is the total for new businesses

$$Y_{all} = Y_T + Y_{NBP}$$

We need a method of estimating Y_{all} and since Y_T can be estimated using generalised regression our task is to estimate Y_{NBP}. No units are surveyed from the new business population but estimates of the numbers of new businesses in G classes or domains of the population are available. The classes may change when we have BAS data but for the time being we can assume that the method of estimating using New Business Provisions is fixed. For each class, g:

$$Y_{all,g} = Y_{T,g} + Y_{NBP,g} \text{ is estimated by } \hat{Y}_{all,g} = \hat{Y}_{GREG,g} + \hat{Y}_{NBP,g}$$

We estimate $Z_{T,g}$, the number of live units in class g on the register and use it to estimate the average value for live register units in that class.

$$\hat{\mu}(Y_g) = \hat{Y}_{GREG,g} \big/ \hat{Z}_{T,g}$$

Assuming that the expected value of new businesses in class g equals the actual average value for live register units in class g, we can obtain a synthetic estimate of the new business population total Y_{NBP} by first finding synthetic estimates of the new business total in each class.

$$\hat{Y}_{NBP,g} = N_{NBP,g}\hat{\mu}(Y_g) \text{ and } \hat{Y}_{NBP} = \sum_g \hat{Y}_{NBP,g} \text{ . Hence } \hat{Y}_{all} = \hat{Y}_{GREG} + \hat{Y}_{NBP} \text{ .}$$

The synthetic estimate of the total of the entire population from both the register and new businesses can be expressed in the form: (Stephen Carlton, Model Assisted Methods course notes)

$$\hat{Y}_{all} = \sum_g \left(\hat{Y}_{GREG,g} + \hat{Y}_{GREG,g} N_{NBP,g} \big/ \hat{Z}_{T,g} \right) = \sum_g \hat{Y}_{GREG,g} \left(1 + N_{NBP,g} \big/ \hat{Z}_{T,g} \right) \qquad (3.2.1)$$

An alternative way of considering this is to let w_i be the weights found when fitting generalised regression calibrated to frame benchmark totals. Then adjust the weights for New Business provisions as follows (Bill Gross, personal communication):

$$w_i^* = \begin{array}{l} w_i \left(1 + \dfrac{N_{NBP,g}}{\sum\limits_{j \in g \cap live} w_j} \right) \text{ for } i \in g \cap live \\ w_i \text{ for } i \notin live \end{array} \qquad (3.2.2)$$

Equations 3.2.1 and 3.2.2 would be equivalent if

$$\hat{Z}_{T,g} = \sum_{i \in live \cap g} w_i$$

and this is not necessarily the best estimate for $Z_{T,g}$ for estimating $Y_{NBP,g}$.

Now consider an estimate of the population mean. Under the method of 3.2.1, we would use the best estimate of the population total divided by the best estimate of the population count, which would have the form:

$$\hat{\bar{Y}} = \frac{\sum w_{1i} y_i}{\sum w_{2i}} \qquad 3.2.3$$

whereas, under 3.2.2, the estimate would have the form:

$$\hat{\bar{Y}} = \frac{\sum w_i y_i}{\sum w_i} \qquad 3.2.4$$

If the y_i's in sample were all constant, the estimate of the mean in 3.2.3 would not equal this constant value but it would using 3.2.4. This property of 3.2.4 gives the solution in 3.2.2 an intrinsic appeal, although ultimately a solution's appeal will depend on its bias and variance.

3.3 Winsorising

Very large sample values above some cut-off value are reduced in size in a process called winsorising. A winsorised estimator is one calculated using the new adjusted values. This estimator is biased but, if the cut-offs are chosen appropriately, it will have a smaller mean-squared error than the equivalent non-winsorised estimator. A bad choice of cutoffs could lead to a larger mean-squared error. Kokic and Bell (1994) developed a method for calculating optimal winsorising cutoffs for the set of strata in a repeated survey that will ensure a reduction in mean-squared error. They use type II winsorisation, namely:

$$
\begin{aligned}
X_{hi} &= X_{hi} && \text{if } X_{hi} < K_h \\
&= f_h X_{hi} + (1\text{-}f_h) K_h && \text{otherwise}
\end{aligned}
$$

where X_{hi} is the response from unit i in stratum h, K_h is the cut-off for stratum h, and f_h is the sample fraction, n_h / N_h, for stratum h

The extension of winsorisation in the presence of generalised regression estimation involves replacing the response, X_{hi}, with the generalised regression model's residual,

$$e_{hi} = X_{hi} - X_{h,GREG}.$$

In other words, outliers will be identified by deviation from the GREG 's model prediction. However, it is now possible to have very large negative as well as positive values and so the method of determining cutoffs would need to be extended to allow for outliers in both directions. Clark (unknown reference) has considered this problem for the case of ratio estimation.

We need to follow up Clark's method, assess its suitability for generalised regression and extend it if necessary.

3.4 Variance Estimation

Currently we use Jackknife variance estimation. With samples as large as 5000, this would mean that 5000 estimates would have to be calculated in order to estimate the variance. Consequently, the programs in the generalised facilities have been written with a view to efficient implementation of this procedure.

With the introduction of a new method of estimation we need to investigate suitable methods for variance estimation. The variance of a straightforward GREG estimator is approximated through Taylor linearisation. To do this we must define some further quantities. The hypothetical population fit of the model, ξ, produces **B**, and thence the fitted values and associated residuals for all population members,

$$y_k^0 = \mathbf{x}_k'\mathbf{B} \qquad E_k = y_k - y_k^0.$$

The fitted values for the sample and the associated residuals are

$$\hat{y}_k = \mathbf{x}_k'\hat{\mathbf{B}} \text{ for } k \in s \text{ and } e_{ks} = y_k - \hat{y}_k.$$

The approximate variance is:

$$AV\left(\hat{Y}_{GREG}\right) = \sum\sum_U (\pi_{kl} - \pi_k\pi_l)\, E_k E_l / \pi_k\pi_l$$

and this can be estimated from the sample by

$$\hat{V}\left(\hat{Y}_{GREG}\right) = \sum\sum_s (1 - \pi_k\pi_l/\pi_{kl})\left(g_{ks}e_{ks}\right)\left(g_{ls}e_{ls}\right)/\pi_k\pi_l$$

However, with extras such as new business provisions and live respondent mean imputation, it is more difficult to derive an analytical solution. Kovar, Rao and Wu (1988) compared several variance estimators in the context of stratified simple random sampling and found that balanced repeated replication and repeated random group methods performed consistently well, although the former tended to over-estimate the variance. Rust and Rao (1996) also compared variance estimation methods for large and complex survey designs and found that confidence intervals calculated using balanced repeated replication and bootstrapping were closer to the nominal coverage rates than for other methods. The economic surveys at which this paper is directed are all stratified with strata either completely enumerated or using simple random sampling.

If we adopt one of these methods of variance estimation, we will only need to calculate 100-200 estimates and so can forego some of the efficiencies of programming in favour of a more structured approach to the generalised facilities. A suggested structure is to have a module for estimation, a module for generating a replicate sample and a module for producing a table of estimates with associated relative standard errors.

One problem that may arise is that when calculating a variance for a small subdomain, most of the sampled units may always be in the replicate and so our variance estimate will be very low. The variance of movement estimates also presents a problem in that the population and the sample replication can change between two consecutive time periods. The method of variance calculation would need to reflect the way we control rotation.

Issues concerning us for variance estimation are how to handle non-response and imputation in replicated sampling, how well replicated sampling will perform for small subdomains, whether replicated variance estimation will work for two-phase sampling, how we should calculate movement variances and how to deal with New Business Provisions.

4. Imputation of Missing BAS Data

With new methods arising from the use of BAS data, we must face the fact that some BAS returns will be too late for our purposes. Timeliness of the BAS data will be most important in data substitution and so we require a model for imputing missing BAS data. We may also need to impute BAS data for use in estimation, stratification and imputation of survey variables. The solution to this second requirement may be to use the same method as for the first, with possible weakening of the success criteria. Certainly an imputation method that works well for substitution should be more than adequate for other purposes.

An example of an imputation method is to calculate the live respondent mean for the BAS variable(s) of interest and deflate this value by a death factor to allow for the fact that some non-respondents are dead. Then aggregate the outcomes for substitution. For estimation purposes, the units with missing BAS data could be put into a separate stratum and an alternative estimation method used, such as calibrating to a population count instead of a BAS variable total. A second example of an imputation method may entail using BAS data from previous quarters.

We will need to examine submission dates on BAS forms and compare with expected dates of transfer of BAS data from the ATO so that we can see how much BAS data comes on time. If responses are missing at random we could use a nested model but this is an unlikely scenario. We could develop a risk profile of BAS providers that will enable us to model values for missing BAS data. Profiles will also allow us to impute values for births to be used in assigning new businesses to strata until they start submitting actual BAS data.

Some constraints on developing and evaluating an imputation method include the fact that seasonal patterns may change over time, that a method of imputation works well for totals and means but may not work so well for movements and rates, that standard errors will not be sampling errors but modelling errors.

There are several issues to consider in devising a method or methods for imputation. First, the problem of late response may be compounded if late respondents simply amalgamate data over two quarters and submit them in a single BAS return. Early indications are that the Australian Taxation Office will allow this. Thus we would have a missing value one quarter followed by a value that is much too large the next quarter. For substitution purposes we need to be able to identify these units, impute the missing value and adjust the amalgamated value.

The imputed BAS data will be a potential source of bias and so we should assess an imputation method by estimating the biases of imputed data items retrospectively, when most of the missing data have arrived. It is worth noting that while the imputation method may introduce a bias for level estimates, the bias may disappear for estimates of movements. The practice of amalgamating over quarters may make calculation of bias impossible for some units.

We require a way to determine an appropriate level of accuracy for making these comparisons. The level of accuracy will depend on the use to which the imputed data are to be put. For instance, a greater error may be tolerable for estimation and stratification purposes than for data substitution. Although we need to impute values at the unit level, we may assess the accuracy at the aggregate level. So the second issue is to devise an evaluation method, including calculating the bias and determining the level of accuracy required.

The third issue is how to deal with births of new units versus continuing units. When a new business registers, it may not start submitting BAS immediately. It may not be active initially and so have nothing to report, it may be tardy at sending in the paperwork initially or the Taxation Office may be slower at processing its first BAS return. If an imputation model were based on the previous quarter's values, it could not be used for new units. In this case a model based on characteristics at registration would be more appropriate.

This issue is of particular importance in stratification. Given that we do not want units to swap strata too often, one or two quarters of missing BAS data can be dealt with by maintaining the status quo. However, a unit that is new to the Australian Business Register, but from which a BAS form has not yet been processed, presents a different problem. Such a unit could be allocated to a stratum on the basis of registration information and then reallocated once BAS data are forthcoming. This strategy could lead to too much volatility among new units and so some "stickiness" as discussed in section 2 would need to be introduced. An alternative could be to keep new units out of sample and allow for them in the New Business Provisions until BAS data are available.

The fourth issue involves distinguishing deaths from true late respondents. This could be addressed by using mixture modelling, consisting of a model for death probability mixed with a regression imputation model. There are two approaches, the first using current BAS data such as is used in ratio imputation, and the second involving previous data. For instance, a set of death parameters could be calculated from previous data and applied to the current count of missing data.

5. Synthesis

If we can improve the accuracy of estimates for a given sample size, we can do the reverse and maintain the current accuracy and thus effect a reduction in sample size. The stratification and estimation procedures described above will not have independent effects on estimate variances and so there will need to be some joint investigation. The best estimation technique will be conditional on the stratification used and vice versa. An ad hoc investigation would be to take some stratification perceived to be good and use that to find the best estimation. Then use this estimation method to experiment with stratification methods until some optimal combination becomes apparent.

With the aid of graphics we can ensure that trends are not being hidden or emphasised by a few influential values. It is quite likely if we smooth BAS data for stratification that there will still be correlation within strata indicating that the variable(s) can be used for estimation purposes as well.

Berger, Tirari and Tille (2000) suggest a way of incorporating strata into the GREG model. They use two types of auxiliary variables, calibration variables and design variables. The latter are used in the sampling design; for example, indicator variables for strata are used in a stratified random sample. They show that it is not necessary to have precise knowledge of inclusion probabilities to make inferences provided that the design variables are known. They use a regression estimator with design variables and calibration variables as regressors. From our point of view the practical disadvantage of this method is that is has only been applied to count data.

One thing we should be aware of in our search for solutions to the problems outlined in this paper is that the best solution from a methodological viewpoint may not afford the organisation enough gains to be worth the extra complexity involved in its implementation. The final issue is whether this paper has covered all the worthwhile options for using BAS data in stratification and estimation and for imputing missing or late BAS data.

References

Berger, Y.G., Tirari, M.E. and Tille, Y. (2000) Optimal generalised regression estimation under complex sampling designs, *unpublished manuscript*.

Brewer, K.R.W., Gross, W.F. and Lee, G.F. (1999) PRN sampling: the Australian experience, International Statistical Institute Proceedings ACTES.155-163.

Deville, J.C. and Sarndal, C.E. (1992) Calibration estimators in survey sampling, *JASA* **87**: 376-382.

Gross, W. (1997) Making provisions for new businesses in ABS surveys, MAC June 1997 Meeting, item 5.

Kokic, P.N. and Bell, P.A. (1994) Optimal winsorizing cutoffs for a stratified finite population estimator, *J. Official Stats*, **10(4)**: 419-435.

Kovar, JG, Rao, JNK and Wu, CFJ (1988) Bootstrap and other methods to measure errors in survey estimates, *The Canadian Journal of Statistics*, **16** (supplement) 25-45.

McKenzie, R. and Gross, W. (2000) Synchronised sampling, *Proceedings of the Second International Conference on Establishment Surveys.*

Rust, KF, and Rao, JNK (1996) Variance estimation for complex surveys using replication techniques, *Stat. Meth. in Med. Res*, **5**: 283-310.

Sarndal, Swensson and Wretman (1992) *Model assisted survey sampling*, Springer Verlag.

Acknowledgments

I would like to thank Dr Bill Gross, Mr Graeme Brown and Mr James Chipperfield for their valuable comments in the writing of this paper.

Appendix--Issues

These are the issues that have arisen in the writing of this paper. While we would appreciate MAC input to all of them, we would particularly like MAC members to comment and provide advice on those marked with an asterisk.

Stratification

1. For the minimum approach we would like advice on how to value stratum changes. We also need to consider whether sample allocation and rotation will be preserved.
2. For the comprehensive approach, we would like recommendations on how to decide whether a particular stratification works well.
3. For sticky stratification, we would like advice on a method for determining extra barriers inside the stratum boundaries, taking into account the completely enumerated sector.
4. *We would like a summary measure for the effectiveness of a variable for stratification, i.e. a measure that does not require us to actually perform the stratification.

Estimation

1. Choice of a suitable distance function or an iterative Ordinary Least Squares for use in generalised regression estimation.
2. Determine a suitable method for winsorising outlier residuals resulting from generalised regression estimation.
3. *How to handle non-response and imputation in replicated sampling for variance estimation?
4. *How well does replicated sampling for variance estimation perform in small subdomains?
5. *Will replicated variance estimation work for two-phase sampling?
6. *How should we calculate variances of movement estimates?
7. *How should we deal with New Business Provisions when estimating variance?

Imputation

1. Identifying units with missing data in one quarter and amalgamated data the next.
2. *Devise an evaluation method, including calculating the bias and determining the level of accuracy required.
3. How to deal with births of new units versus continuing units.
4. Distinguish deaths from true late respondents.

Synthesis

1. *Have we covered all the worthwhile options for using BAS data in stratification and estimation and for imputing missing or late BAS data?
2. Is the greater complexity involved in developing and implementing new stratification and estimation procedures worth the gains in sample efficiency?