1352.0.55.054

Australian Bureau of Statistics

Research Paper

# Strategies for Synthetic Estimation in ABS Business Surveys

www.abs.gov.au

# Research Paper

# Strategies for Synthetic Estimation in ABS Business Surveys

Robert Clark and James Chipperfield

Statistical Services Branch

Methodology Advisory Committee

20 June 2003, Canberra

AUSTRALIAN BUREAU OF STATISTICS

EMBARGO: 11.30 AM (CANBERRA TIME) MON 20 MAY 2002

ABS Catalogue no. 1352.0.55.054
ISBN 0 642 48158 X



Produced by the Australian Bureau of Statistics

## INQUIRIES

The ABS welcomes comments on the research presented in this paper.
For further information, please contact Mr James Chipperfield, Statistical Services Branch on Canberra (02) 6252 7301 or email <james.chipperfield@abs.gov.au>.

# CONTENTS

# STRATEGIES FOR SYNTHETIC ESTIMATION IN ABS BUSINESS SURVEYS

Robert Clark and James Chipperfield
Statistical Services Branch

## 1. SUMMARY

A number of external factors have contributed to a need to better understand the role of synthetic estimation methods in ABS business surveys:

i. ABS currently produces several statistical products for small subpopulations of the population of Australian businesses, in response to growing user demand for these products.

ii. Sample sizes for direct collections by the ABS have decreased, for example the Manufacturing and Agricultural Censuses have become samples in the past five years. This is due to the need to reduce provider load and costs, and to an increased focus on national statistical priorities.

iii. The quality and availability of taxation data have greatly improved. Business Income Tax (BIT) data has been in use for several years in the Economic Activity Survey (EAS). Business Activity Survey (BAS) data has been in limited use for several years for the purposes of sample design, and it has been used as an auxiliary variable for sample design or estimation in many surveys since August 2002.

iv. An ABS small area estimates project is being conducted by MD. The project will produce a manual and a set of guidelines for applying small area estimation techniques in the ABS, by the end of 2003.

As a result, it is a good time to clarify the concepts underlying our current practices in synthetic or data substitution methods, and to suggest directions for future developments.

EAS/Tax is the main economy-wide annual survey conducted by the ABS. As the name suggests, EAS/Tax statistics are currently compiled with extensive use of tax data, including the use of tax data to completely impute the values of interest for businesses smaller than a size cutoff. EAS/State estimates are also produced by combining this data with data from other surveys collecting state information. This paper will concentrate mainly on the EAS/Tax and EAS/State statistics. It is intended that the EAS/Tax framework will in future be the framework for all annual business surveys in the ABS.

It will be argued that the current methodologies used in EAS/Tax and EAS/State are not synthetic estimation methods in the usual sense. Rather they are a combination of direct estimation methods, synthetic estimation, extrapolation and imputation methods. This paper will outline how these processes can operate together and how EAS/Tax estimation processes could be organised in this framework.

Other issues covered by this paper include: measuring the quality of our current approaches; measuring the quality of synthetic methods in general; and dealing with missing auxiliary variables. The paper will discuss disaggregation by industry and state, not geographical small area estimates. The latter are produced on an irregular basis for the Retail industry, but issues are quite different than for industry and state statistics.

## 2. CURRENT METHODOLOGY FOR EAS/TAX STATISTICS

EAS/Tax statistics give annual information on Australian businesses over almost the full range of industries, including both employers and non-employers. A combination of direct collected data and tax data is used, with the frame being partitioned into the following three streams:

- Stream D stands for direct collection only. It contains businesses with a complex structure, such that there is not a one-to-one match between the ABS unit and the Australian Business Number (ABN). Data is collected for a sample of these businesses using an EAS form and no tax data is used. Stream D units tend to be larger businesses. About 18% of businesses and 63% of total income come from stream D.
- Stream B stands for both direct and tax collection. It contains businesses where there is a one-to-one match with the ABN. A sample of stream B units are sent an EAS form, and this data is combined with less detailed Business Income Tax (BIT) data for a much larger sample of businesses. About 24% of businesses and 26% of total turnover come from stream B.
- Stream T stands for tax data only. It contains the smallest businesses where there is a one-to-one match with the ABN. Tax (BIT) data is used exclusively for these businesses – they are not surveyed directly. BIT data is less detailed than required for EAS, so pro-rating factors are calculated using stream B businesses and applied to stream T businesses, for the purpose of estimating detailed financial characteristics. For example, BIT data includes total income, but income from interest earned is not included. The proportion of income coming from interest earned is estimated from stream B and this proportion is multiplied by the BIT total income for stream T. About 58% of businesses and 11% of total turnover come from stream T.

BIT data is missing for 20–25% of stream B and T units. The BIT values are imputed for these businesses. A complicating factor is that businesses with missing BIT data are more likely to be defunct or inactive businesses. To account for this, missing BIT values are imputed using an industry (ANZSIC subdivision) mean, multiplied by a live factor. The live factor is estimated using sample information on the rate of defunct businesses with missing and non-missing BIT data. The live factor implies a model for the probability of defunct businesses conditional on their industry and whether their BIT data is available or not. Investigations are continuing into whether this model can be improved.

After imputation, BIT values are available for the whole population of stream B and T units. In the past a large sample of BIT values was edited and only this sample of

values was used, however from the 2000/2001 reference year all BIT values have been edited and used.

BIT data is assumed to be more reliable than EAS data in streams B and T. However, only a few major data items are available from BIT (e.g. total income from each business) whereas EAS has much more detailed data items (e.g. components of income such as income earned from interest).

A pro-rating method is used for combining BIT data and EAS data in stream B. For example, let $t_{\text{EAS},i}$ be total income collected by EAS, let $t_{\text{TAX},i}$ be total income from BIT data and let $c_{\text{EAS},i}$ be a component of income (e.g. income from interest earned) collected by EAS. Let $k$ be an output category which is within an ANZSIC subdivision $h$. The pro-rated estimate of the total of the component of income for subpopulation $k$ of stream B is:

$$\hat{C}_k^{(\text{stream B})} = \frac{\hat{C}_{\text{EAS},h}^{(\text{stream B})}}{\hat{T}_{\text{EAS},h}^{(\text{stream B})}} T_{\text{TAX},k}^{(\text{stream B})}$$

where $T_{\text{TAX},k}^{(\text{stream B})}$ is the population total of $t_{\text{TAX},i}$ over subpopulation $k$ of stream B. The main output classes $k$ are industry (ANZSIC subdivision and broader). ANZSIC class estimates (finer than ANZSIC subdivision) are published on an experimental basis.

The formula above does a few different things at once. It ensures that the estimates for components of income add up to the total income known from BIT. In addition, it makes use of the fact that $t_{\text{TAX},i}$ is available for the whole population, so that sampling errors are greatly reduced compared to a probability-weighted estimator of $C_k$. For output classes finer than ANZSIC subdivision, the estimator is a small-area estimator because it is assumed that the ratio

$$\frac{\hat{C}_{\text{EAS},h}^{(\text{stream B})}}{\hat{T}_{\text{EAS},h}^{(\text{stream B})}}$$

can be applied to subpopulations of $h$.

A pro-rating method is also used for stream T, where only BIT data is available. Let $k$ be an output category within an ANZSIC subdivision $h$. The pro-rated estimate of the total of the component of income for subpopulation $k$ of stream T is:

$$\hat{C}_k^{(\text{stream T})} = \frac{\hat{C}_{\text{EAS},h}^{(\text{stream B})}}{\hat{T}_{\text{EAS},h}^{(\text{stream B})}} T_{\text{TAX},k}^{(\text{stream T})}$$

where $T_{\text{TAX},k}^{(\text{stream T})}$ is the population total of $t_{\text{TAX},i}$ over subpopulation $k$ of stream T.

The pro-rating factor

$$\frac{\hat{C}_{\text{EAS},h}^{(\text{stream B})}}{\hat{T}_{\text{EAS},h}^{(\text{stream B})}}$$

is calculated from stream B data, and applied to the smaller businesses in stream T. Using stream B pro-rating factors adds both variance and bias to stream T estimates. There is additional variance because the pro-rating factors are estimated using stream B sample data. Estimates are biased because the pro-rating factors for stream B may be different to the actual proportions for stream T. The variance for pro-rated items is being included for the first time in the 2002/2003 sample design. The strategy for controlling the bias is to ensure that no more than 15% of total income comes from stream T.

# 3. A FRAMEWORK: IMPUTATION, EXTRAPOLATION AND SYNTHETIC ESTIMATION

The pro-rating method for EAS/Tax described in Section 2 is a combination of imputation, extrapolation and synthetic estimation. It is suggested that these three processes be performed separately, to allow better validation and documentation of model assumptions. This would also enable standard synthetic and direct estimation techniques to be applied, rather than EAS using an apparently unique estimation methodology.

An imputation process is needed to force the components of income to add to the known BIT total of income. The pro-rating method described in Section 2 translates to

$$\hat{c}_i = \frac{\hat{C}_{\text{EAS},h}^{(\text{stream B})}}{\hat{T}_{\text{EAS},h}^{(\text{stream B})}} T_{\text{TAX},i}$$

for each sampled business $i$. Other imputation schemes could also be used. A difficulty is that there are no businesses for which $c_i$ is directly observed, so that it is not possible to analyse whether ANZSIC subdivision ($h$) is the right level at which to impute. One possibility is to impute at the business level, to remove any arbitrary grouping for imputation:

$$\hat{c}_i = \frac{c_{\text{EAS},i}}{t_{\text{EAS},i}} t_{\text{TAX},i}$$

An estimation process is needed to make use of the BIT data which is available for the whole population, to estimate the totals of $\hat{c}_i$. One approach is ratio estimation, with $t_{\text{TAX},i}$ as the benchmark variable and $\hat{c}_i$ as the variable of interest. If ratio estimation was done within each ANZSIC class $h$, and $\hat{c}_i$ were calculated using the first formula in the previous paragraph, this would give identical ANZSIC subdivision estimates as the current methodology. However, more sophisticated estimation methods could also be used, such as generalized regression estimation.

The current estimation method for ANZSIC class estimates is synthetic, because pro-rating factors calculated at the subdivision level are assumed to apply to each class in the subdivision. A range of synthetic estimation techniques could be applied, using $\hat{c}_i$ as the variable of interest and $t_{\text{TAX},i}$ as the benchmark variable.

The current estimation method for stream T estimates is an extrapolation method, because pro-rating factors calculated from stream B are assumed to apply to the smaller businesses in stream T. Pro-rating for stream T of EAS/Tax is an extrapolation because pro-rating factors are calculated from stream B and applied to the different population of stream T. Estimates are biased to the extent that the correct pro-rating

factors are different between streams B and T. The bias is controlled by making stream T a relatively small proportion of total income, no more than 15%.

The use of a stream T is essentially a trade-off between variance and bias, however the bias implications of this strategy are not well understood. The bias is due to the fact that the appropriate pro-rating factors for stream B may be different to those in stream T. Bias would be expected if

$$\frac{\hat{C}_{\mathrm{EAS},b}^{(\mathrm{stream\,B})}}{\hat{T}_{\mathrm{EAS},b}^{(\mathrm{stream\,B})}} \neq \frac{\hat{C}_{\mathrm{EAS},b}^{(\mathrm{stream\,T})}}{\hat{T}_{\mathrm{EAS},b}^{(\mathrm{stream\,T})}}$$

Some sensitivity analysis could be undertaken on how different these ratios might be, and what effect scenarios would have on the bias of estimates across all streams. An analysis of how $\frac{c_{\mathrm{EAS},i}}{t_{\mathrm{EAS},i}}$ varies with business size within stream B could suggest how differently this ratio might behave for the smaller businesses in stream T. Data from quarterly surveys could also be used to examine how this ratio depends on size. A more radical alternative would be to introduce some selective sampling from stream T for the purpose of model validation. This analysis will give a number of plausible alternative values for the extrapolation bias, rather than a single estimate of bias and mean-squared error.

The assumption that BIT data is correct for broad data items for stream B businesses is also a trade-off between variance and bias. BIT data is probably more accurate than directly collected data where it is available, however this is probably not the case where the BIT data has been imputed. An alternative approach would be to use the EAS data instead of BIT for businesses where BIT data is not available, possibly with an adjustment to allow for systematic differences between EAS and BIT. The mean-squared error of different ways of using the BIT data could be evaluated by assuming that BIT data is correct when it is available and assuming a model relating EAS and BIT data.

# 4. CURRENT METHODOLOGY FOR EAS/STATE STATISTICS

EAS and Tax data items relate to national activity (e.g. total expenses for all Australia). State breakdowns of EAS/Tax data items for all businesses selected in the EAS/Tax survey are derived by one of the following:

A. Editing / Imputing:

   i. matching the business' response in EAS/Tax to its response to a state-based ABS collection. A state-based survey is one that collects measures of economic activity broken down by state. EAS/Tax units are first matched to the Manufacturing Survey, and in the absence of such a match, are then matched to Mining, the Quarterly Business Income Survey (QBIS), the Retail Industry Survey and so on down a list of about 8 surveys in total.

   ii. editing the state-based responses as required, to obtain the correct state breakdowns of the business' national activity. If multiple matches to state-based surveys are obtained, editing involves choosing the most reliable source. It may also involve referencing the ABS profile report of the business, direct telephone contact, or by referencing the yellow pages. Where possible the impute is annualised to represent the annual reference period.

About 56% of the estimate of Total Income at the Australia level is based on state breakdowns being derived in this way.

B. Direct contact via telephone, if a match cannot be found and the business' income is over $100 million. About 11% of the estimate of Total Income at the Australia level is based on state breakdowns being derived in this way;

C. Businesses with less than $100m annual income are assumed to operate only in the state of head office (available on the frame). About 33% of the estimate of Total Income at the Australia level is based on state breakdowns being derived in this way.

D. If none of the above produce reliable results and a business' weighted response is considered significant then the business' state breakdown is either imputed with the subdivision average or with its historical response.

This methodology is subject to 6 main assumptions, of which only the first two are considered to be a significant risk of bias. The assumptions have not been fully tested partly because the EAS/State statistics have only been produced since the 1998/1999 reference period. The statistics are published on an experimental basis. The first two assumptions will now be described.

*Assumption 1: All state breakdowns based on editing/imputing relate to the appropriate annual reference period.*

The broad assumption is that the imputed state breakdowns are representative of the business's activity for current EAS/Tax reference period. This broad assumption can be broken up into two specific assumptions:

(i) imputed breakdowns, from previous reference periods, are the same as the actual breakdowns in the current reference period;

(ii) imputed breakdowns, based on a subset of the current period (e.g. two of the four quarters), are the same as the actual breakdowns over the whole reference period.

The size of the bias due to editing/ imputing depends on two factors:

(i) the contribution of the estimate from businesses subject to editing/ imputing. At the Australian level this contribution is currently about 56% (the corresponding contribution at arbitary levels is readily available) ; and

(ii) the degree to which the imputed state breakdowns are a biased estimate of the true state breakdowns for the current reference period.

Validating assumption 1 and the current imputing/editing approach involves assessing the size and volatility over time of factor (ii). This can be done by using the longitudinal quality of QBIS survey data to follow a sample of business over four quarters and assess how the state breakdowns vary across quarters.

*Assumption 2: Employing businesses that are not edited/imputed and have less than $100 million income operate solely in their state of head office*

As mentioned above, businesses with less than $100 million income are assumed to be single state businesses (except those that are edited/imputed). The bias depends on two factors:

(i) the contribution of the estimate from businesses assumed to be single state. The size of this contribution to the estimate of Total income at the Australia level is 33% (including a small component of non-employers). The corresponding contribution at arbitrary levels is readily available; and

(ii) the level of activity assumed to be single state that is in fact multi-state

For units with employment less than 200 (as measured on the frame for the Sales Survey in 1999), between 5–10% of activity at state level is associated with businesses that have the head office not in that state. Note that for a given state, the corresponding percentages change by up to 2 percentage points in two adjacent

quarters. The size of the bias of finer level estimates may be much larger in percentage terms and more volatile.

To estimate (ii) more accurately would involve estimating the state breakdowns for the population of businesses with income less than $100m, using data from state-based surveys such as QBIS.

*An Alternative Methodology for EAS/State*

An alternative approach would be to multiply: EAS/Tax subdivision estimates at the national level; by the state breakdowns of the subdivision estimate, as estimated from the appropriate state-based survey for that subdivision. For example, if the Manufacturing survey estimates 10% of income in a subdivision is associated with South Australia, and the EAS/Tax estimates the total income for that subdivision is $10 billion, the EAS/Tax state estimate for SA would be $1 billion (10% × $10 billion). In this way the accuracy of the state estimate is dependent upon the accuracy of the EAS/Tax national estimate as well as the accuracy of the state breakdowns which are estimated fom the state based survey.

The main benefits of the new approach would be:

- A larger sample size would be available to calculate the state breakdowns, because all sampled units from the state-based surveys could be used, not just units which match an EAS/Tax response.
- Assumptions 1 and 2 would not be required, resulting in less bias to state estimates.

Future work on EAS/State will involve analysing assumptions 1 and 2 in more depth, to determine whether they cause significant biases. Depending on the outcome of this analysis, the new approach just mentioned may be implemented, or other improvements may be introduced. EAS/State estimates should also be placed in the conceptual framework outlined in Section 3, with a view to standardising concepts, methodologies and possibly estimation systems for ABS annual surveys.

# 5. STRATEGIES FOR PROGRESSING SYNTHETIC ESTIMATION FOR ABS BUSINESS SURVEYS

Imputing because of inconsistent (usually only slightly) data sources, and using auxiliary data to improve efficiency, should be managed as two separate processes. At present, pro-rating accomplishes both steps, with the result that it becomes difficult to use more general models to improve efficiency without causing a statistical impact. Separating imputation and estimation will allow more flexible use of auxiliary data in regression or synthetic estimation. This is expected to be an outcome of the reviews of the EAS/State and EAS/Tax methodologies which will be completed by the end of 2003.

Better quality measures are needed for the extrapolation of pro-rating factors to non-direct collection businesses. The currently used measure of "percentage contribution of non-direct units" should be disseminated more widely. Some measures of bias and mean-squared error should also be developed, although this may have to rely on sensitivity analysis because of the lack of collected data to compare against. Some work on this issue will be undertaken as part of the EAS/Tax review later in 2003.

There is demand for synthetic estimates both for regular, repeated business statistics (e.g. EAS/Tax ANZSIC class estimates), and for one-off studies (e.g. Retail geographic estimates). The latter have the advantage that models can be developed and checked in depth and quality issues can be investigated and documented in a flexible way. The former have the advantage of meeting external users and National Accounts requirements for regular, ongoing statistics. Ongoing synthetic estimates need to be based on simple, robust methods, with easily calculated quality measures. Because of the difficulty of extensively checking the model each time, estimators should be robust – for example, modelling might only be used for the smallest businesses. A review should be conducted sometime in 2004, to identify high priority candidates for each of these approaches.

Analytical Services Branch are currently managing an ABS Small Area Estimation (SAE) Project which will produce a manual of small area techniques, case studies and training materials. The manager of the SAE project, Daniel Elazar, is involved in the EAS work described here in an advisory role. The objective of the EAS work is to develop a simple, robust estimation method with good documentation of the methodology and quality, in order to remove the experimental flags from the EAS/State and EAS ANZSIC class estimates. SAE staff will advise on this work, and the SAE manual may be used in future to further enhance EAS methodology using more sophisticated techniques.

# 6. ISSUES FOR MAC

Comment is sought from MAC members on all aspects of this report particularly the following areas:

i. the current use of pro-rating for EAS/Tax and EAS/State estimates;

ii. the process of extrapolating stream B pro-rating factors to the smaller businesses in stream T in EAS/Tax;

iii. measuring the quality of the non-directly collected estimates, i.e. how to assess the bias from extrapolating a model from medium-sized businesses to small businesses.

iv. The use of synthetic estimation techniques for ongoing regular statistics, so that it is not possible to evaluate the model in detail every cycle.

## FOR MORE INFORMATION . . .

| | |
|---|---|
| *INTERNET* | **www.abs.gov.au** the ABS web site is the best place for data from our publications and information about the ABS. |
| *LIBRARY* | A range of ABS publications are available from public and tertiary libraries Australia wide. Contact your nearest library to determine whether it has the ABS statistics you require, or visit our web site for a list of libraries. |

### INFORMATION AND REFERRAL SERVICE

Our consultants can help you access the full range of information published by the ABS that is available free of charge from our web site, or purchase a hard copy publication. Information tailored to your needs can also be requested as a 'user pays' service. Specialists are on hand to help you with analytical or methodological advice.

| | |
|---|---|
| *PHONE* | 1300 135 070 |
| *EMAIL* | client.services@abs.gov.au |
| *FAX* | 1300 135 211 |
| *POST* | Client Services, ABS, GPO Box 796, Sydney NSW 2001 |

## FREE ACCESS TO STATISTICS

All ABS statistics can be downloaded free of charge from the ABS web site.

| | |
|---|---|
| *WEB ADDRESS* | www.abs.gov.au |

2000001524381
ISBN 0 642 48158 X

**RRP $11.00**