1352.0.55.070

Australian Bureau of Statistics

Research Paper

Estimation Methodologies Using Taxation Data for ABS Business Surveys

Research Paper

Estimation Methodologies Using Taxation Data for ABS Business Surveys

Melissa Gare, John Preston and Edward Szoldra

Statistical Services Branch

Methodology Advisory Committee

17 June 2005, Canberra

AUSTRALIAN BUREAU OF STATISTICS

EMBARGO: 11.30 AM (CANBERRA TIME) THURS 16 FEB 2006

ABS Catalogue no. 1352.0.55.070
ISBN 0 642 48174 1

Views expressed in this paper are those of the author(s), and do not necessarily represent those of the Australian Bureau of Statistics.
Where quoted, they should be attributed clearly to the author(s).

Produced by the Australian Bureau of Statistics

INQUIRIES

The ABS welcomes comments on the research presented in this paper.
For further information, please contact Ms Melissa Gare, Statistical Services Branch on Canberra (02) 6252 7147 or email <m.gare@abs.gov.au>.

ESTIMATION METHODOLOGIES USING TAXATION DATA FOR ABS BUSINESS SURVEYS

Melissa Gare, John Preston and Edward Szoldra
Statistical Services

EXECUTIVE SUMMARY

The Economic Activity Survey (EAS) is an annual business survey conducted by the ABS to provide information about the operation and financial performance of all businesses in the public trading and private employing sectors of the Australian economy for use in Input-Output tables and the National Accounts. The availability of identifiable unit level Business Income Tax (BIT) data provided by the ATO has enabled the ABS to produce estimates at a finer industry level with the benefits of improved coverage and data quality but with no increase in provider load, based on directly collected EAS data combined with BIT data, referred to as EAS/Tax estimates.

A paper presented to the June 2003 MAC “Strategies for Synthetic Estimation in ABS Business Surveys”, by Robert Clark and James Chipperfield detailed the estimation methodology being used at that time for the EAS/Tax estimates. The current MAC paper is a follow on from this previous paper. The paper identifies the five key issues that need to be considered in terms of using the BIT data in the annual business surveys: population coverage, incompleteness with respect to survey population frame, data item coverage, data item quality and timing. The assumptions underlying the current estimation methodology for the EAS/Tax estimates are examined in relation to the incompleteness of the BIT data with respect to survey population frame and limited data item coverage of the BIT data. The current estimation methodology is compared with an alternative Generalised Regression (GREG) estimation methodology that is currently being considered.

QUESTIONS FOR MAC MEMBERS

The questions for MAC members in relation to developing an estimation methodology using BIT data are:

1. Are the assumptions underlying the current EAS/Tax estimation methodology to account for the different types of missing data justifiable and sufficient? Should further investigations be undertaken into the impacts of the missing data on the quality of the estimates? How should these impacts on the quality of the estimates be relayed to users?

2. On what basis should a decision be made on whether to use the broad level BIT data items or broad level EAS data items as the variables of interest?

3. What is the best approach to choosing between the options for producing estimates using a GREG estimator?

4. In a survey where there are multiple related variables of interest, on what basis should a decision be made on whether to have a single weight for all of the variables of interest or multiple weights (i.e. a separate weight for each of the variables of interest)?

5. If it is not possible to accurately impute the broad level BIT data items for the non-matched units in the population, is it better to use the GREG estimation methodology (Options 2, 3, 4 or 5) or use the current EAS/Tax "data substitution" methodology?

6. If it is not possible to include all the broad level BIT data items into the GREG model, is it better to use the current EAS/Tax "data substitution" methodology or use the GREG estimation methodology?

7. If estimates are required for small domains, is it better to use the current EAS/Tax "data substitution" methodology or use the GREG estimation methodology?

CONTENTS

The role of the Methodology Advisory Committee (MAC) is to review and direct research into the collection, estimation, dissemination and analytical methodologies associated with ABS statistics. Papers presented to the MAC are often in the early stages of development, and therefore do not represent the considered views of the Australian Bureau of Statistics or the members of the Committee. Readers interested in the subsequent development of a research topic are encouraged to contact either the author or the Australian Bureau of Statistics.

Estimation Methodologies Using Taxation Data for ABS Surveys

Melissa Gare, John Preston and Edward Szoldra
Statistical Services

1. Introduction

1. The Economic Activity Survey (EAS) was originally designed to underpin the Economic Statistics Strategy developed in the late 1980's, with the first survey being conducted in respect of the 1990-91 financial year. The general purpose of EAS is to provide information about the operation and financial performance of all businesses in the public trading and private employing sectors of the Australian economy for use in Input-Output tables and the National Accounts. It also aims to measure changes in the operation, structure and performance of all private and public trading enterprises, and is an economic indicator used to monitor the economy in regard to the business cycle.

2. The availability of identifiable unit level Business Income Tax (BIT) data provided by the ATO in the mid 1990's has enabled the ABS to produce estimates at a finer industry level with the benefits of improved coverage and data quality but with no increase in provider load, based on directly collected EAS data combined with BIT data, referred to as EAS/Tax estimates. A paper presented to the June 2003 MAC 'Strategies for Synthetic Estimation in ABS Business Surveys', by Robert Clark and James Chipperfield detailed the estimation methodology being used at that time for the EAS/Tax estimates. The current MAC paper is a follow on from this previous MAC paper. It aims to summarise the current estimation methodology for the EAS/Tax estimates and compare this methodology with an alternative Generalised Regression (GREG) estimation methodology that is currently being considered for implemented.

2. Annual Business Surveys and the ABS Business Register

3. The ABS currently conducts a suite of annual business surveys which aim to provide information about the Australian economy. The samples of businesses for these annual business surveys are selected from a register of businesses, known as the ABS businesses register. Historically, the ABS business register has covered only employing businesses as there was no definitive source to identify non-employing businesses in a timely fashion. However, a New Taxation System was introduced in Australia in July 2000, where registered businesses received a unique Australian Business Number (ABN) and were included on an Australian Business Register (ABR) maintained by the Australian Taxation Office (ATO). The ABR covers both employing and non-employing businesses. Consequently, the ABS business register (from June 2002) moved to a 'two population model':

1. ATO Maintained Population (ATOMP) which consists of simple businesses where the ATO information and reporting model (i.e. ABN) is sufficient for most purposes; and

2. ABS Maintained Population (ABSMP) which consists of more complex businesses where the ABS statistical units model is applied and maintained by the ABS.

4. Under this 'two population model', the boundary between the populations is defined using ABNs, where the majority of units have a direct one-to-one link through the ABN to the information collected by the ATO. For the purpose of the annual business surveys, active businesses are defined as active units in the ABSMP plus those units in the ATOMP with an active ABN and at least one active Income Tax Instalment (ITI), Income Tax Withholding (ITW) or Goods and Services Tax (GST) role.

3. Business Income Tax data

5. Business Income Tax (BIT) data consists of financial information reported to the ATO to determine a businesses financial year income tax obligation. The key financial data items on the BIT returns include business income and expense items, capital expenditure, opening and closing stocks, profit, assets and liabilities. These data items are conceptually similar to data items collected on the EAS survey forms and can readily be linked to units in the ATOMP population via the ABN. The ABS receives the BIT data from the ATO in respect of Company, Trust, Partnership and Individual income tax returns for which business income and expenses had been declared in a particular financial year. The BIT data is not as timely as the directly collected EAS data, with the initial BIT data for a particular financial year received twelve months after the end of the reference period, compared with three to six months for the directly collected EAS data.

6. In terms of using the BIT data in the annual business surveys, there are five key issues that need to be considered: population coverage, incompleteness with respect to survey population frame, data item coverage, data item quality and timing. These five issues are discussed below.

3.1 Population Coverage

7. Unfortunately, not all units on the population frame for the annual business surveys can be linked to the BIT data. Units in the ABSMP component of the ABS business register have complex structures which do not enable direct linking to BIT data. Furthermore, some businesses in the population are income tax exempt, such as not-for-profit institutions (NPIs), and hence not all businesses in the population are required to complete a BIT return, while for some other businesses in the population, such as non-trading trusts, the income tax data is reported to the ATO but is not provided to the ABS. These problems are currently dealt with by partitioning the population frame for the Economic Activity Survey (EAS) into three distinct streams:

1. Stream D (**D**irectly collected EAS data only) contains businesses maintained by the ABS (i.e. units on the ABSMP) that have a different units structure than the BIT data, plus units that are not covered by BIT data, such as NPIs and non-trading trusts. The data is directly collected for a sample of these businesses using the EAS survey forms. BIT data is not used for these businesses.
2. Stream B (**B**oth Tax and directly collected EAS data) contains the larger businesses where there is a one-to-one link via the ABN with the BIT data. The data is directly collected for a sample of these businesses using the EAS survey form and this data is combined with BIT data.
3. Stream T (**T**ax data only) contains the smaller businesses where there is a one-to-one link via the ABN with the BIT data. The data is soured exclusively from BIT data. No directly collected EAS data are used for these businesses.

8. The estimation methods presented in the remainder of this paper are in respect to the production of estimates for the Stream B component of the population.

3.2 Incompleteness with Respect to the Survey Population Frame

9. All units in the Stream B population are expected to complete a BIT return. However, in practice there are a significant number of units in the Stream B population which are unable to be linked (or matched) with BIT data, leading to incompleteness of BIT data with respect to the population frame. There are several possible reasons for this incompleteness. Firstly, businesses may not have completed their BIT return by the time the BIT data is provided to the ABS. Secondly, the population frame may contain units that are no longer operational and hence no longer obliged to report the BIT data. Therefore, careful consideration needs to be given to the treatment of these units in the imputation and estimation processes.

3.3 Data Item Coverage

10. The BIT data is limited to a small number of broad level data items, as compared to the directly collected EAS data, which contains the broad level data items as well as a large number of fine level breakdowns of these broad level data items. Hence the BIT data does not contain all the data items required for ABS purposes and careful consideration needs to be given to the treatment of these data items in the imputation and estimation processes.

3.4 Data Item Quality

11. Since the broad level data items are available from the BIT data as well as being directly collected in EAS, a choice needs to be made on which of these two sources to base the estimates. This raises questions on the quality of the data items: Is the BIT data reported by an individual businesses to the ATO of higher or lower quality than the EAS data reported to the ABS? The annual business surveys collect the data not long after the end of the financial year and hence respondents may not have completed their final accounts. In this situation respondents might report "estimated" financial information to

the ABS, but higher quality final accounts to the ATO. However, the ABS is an experienced survey organisation and has many practices in place to minimise non-sampling error, including the ability to follow-up respondents where miss-reporting is believed to have occurred. On the other hand, the ABS has limited control over non-sampling errors that may occur in the BIT data provided to the ATO and cannot directly follow-up the respondents.

3.5 Timeliness of the Data

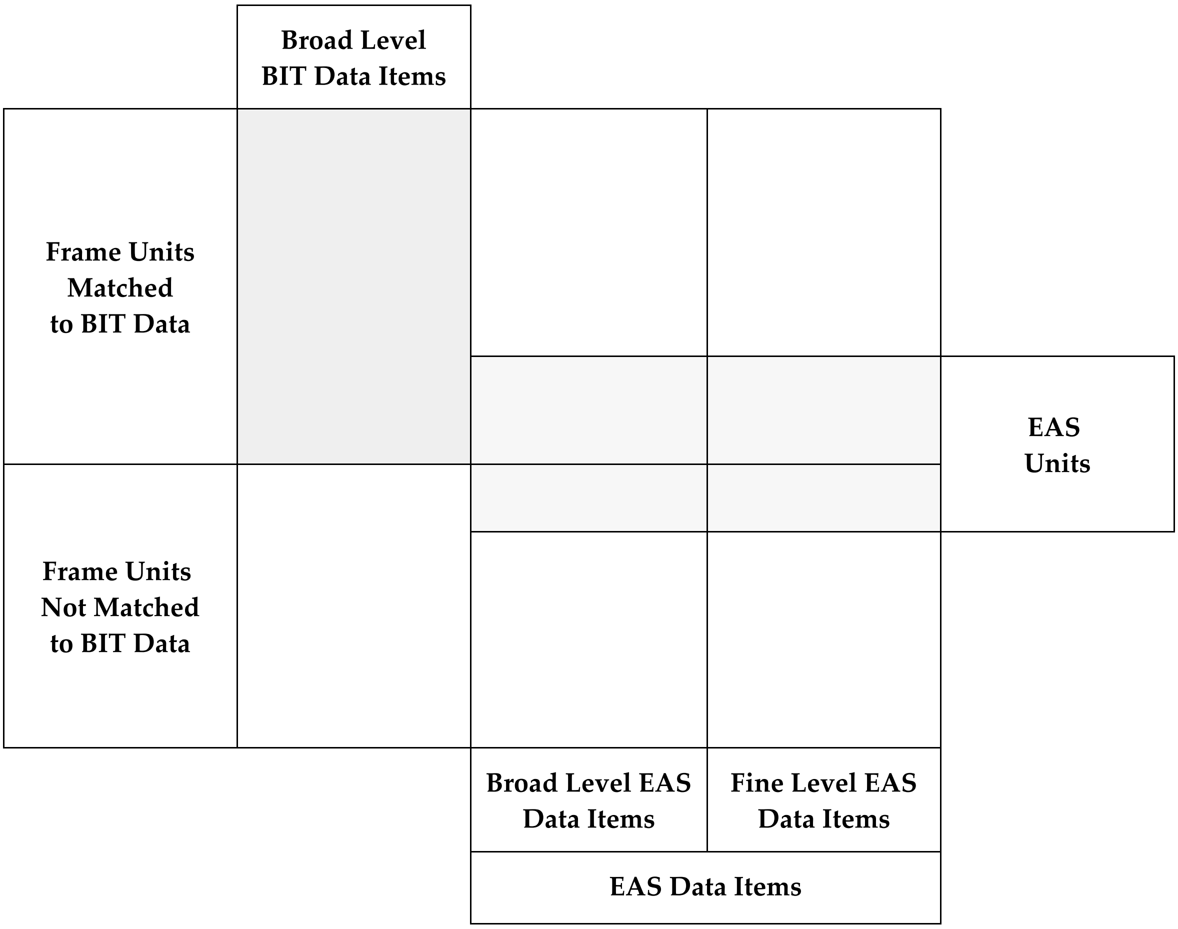
12. The BIT data is not as timely as the directly collected EAS data. The ABS currently receives three snap-shots of BIT data for a financial year taken approximately 12, 15 and 18 months after the end of the reference year. The latest snap-shot has approximately an additional 5 to 10% of businesses with completed BIT returns. However, if the BIT data is to be combined with the directly collected EAS data, then a decision needs to be made in terms of acceptable delays in the release of the estimates. If the key aim is to produce estimates within one year of the end of the reference period, then only BIT data from previous financial years would be available. If the estimates are produced using BIT data from the current financial year, then it will only be possible to produce estimates approximately 18 months after the end of the reference period.

4. Stream B EAS/Tax Framework

13. The goal of annual business surveys is to produce estimates for all units on the population frame, with the aim of using BIT data is to improve upon the accuracy of these estimates and to enable the release of finer level estimates. Diagram 1 provides a pictorial representation of the Stream B population, together with the information available from the BIT data and the directly collected EAS data.

14. The data items are presented from left to right across the diagram, from those available from BIT, to those available from EAS. The EAS data items can be broken into two groups which are referred to as the broad and fine level data items. The broad level data items available from BIT are denoted by $\underset{\sim}{x}_i^{(T)} = (x_{1i}^{(T)}, ...x_{ki}^{(T)}, ...x_{Ki}^{(T)})'$, while the broad level data items collected by EAS are denoted by $\underset{\sim}{x}_i^{(E)} = (x_{1i}^{(E)}, ...x_{ki}^{(E)}, ...x_{Ki}^{(E)})'$. While the broad level data items are those variables in common between the BIT and EAS, the fine level data items are those variables only available from EAS, denoted by $\underset{\sim}{z}_i^{(E)} = (z_{1i}^{(E)}, ...z_{ji}^{(E)}, ...z_{Ji}^{(E)})'$. The broad level data items comprise 24 data items covering income, expense, capital expenditure and key balance sheet items. The fine level data items comprise a large number (differs form industry to industry) of finer level breakdowns of the broad level data items, such as income and expense, as well as other variables such as reported industry and employment.

Diagram 1: Stream B EAS/Tax Framework


Broad Level BIT Data Items
Frame Units Matched to BIT Data
EAS Units
Frame Units Not Matched to BIT Data
Broad Level EAS Data Items
Fine Level EAS Data Items
EAS Data Items

15. The units are presented from top to bottom down the diagram, from those units matched to BIT data denoted by U_m to those units not matched to BIT data denoted by $U_{\overline{m}}$. The EAS selects a random sample of units denoted by s from the population $U = U_m \cup U_{\overline{m}}$ with selection probabilities $\pi_i^{(E)} = \Pr(i \in s)$, resulting in some EAS units matched to BIT data denoted by $s_m$and some units not matched to BIT data denoted by $s_{\overline{m}}$

16. Under this Stream B EAS/Tax framework, there are three types of missing data to be considered when producing estimates:

1. The "**first type of missing data**" is that the broad level BIT data items are not available for the non-matched units in the Stream B population (i.e. $\underset{\sim}{x}_i^{(T)}$ for all units in $U_{\overline{m}}$).
2. The "**second type of missing data**" is that the fine level EAS data items are not available for many units in the Stream B population, since the EAS only collects data from a sample of units in the Stream B population (i.e. $\underset{\sim}{x}_i^{(E)}$ and $\underset{\sim}{z}_i^{(E)}$ for units in U/s). There is also a small amount of unit and item non-response to the EAS which is imputed using standard imputation techniques.
3. The "**third type of missing data**" is that the fine level BIT data items are not available for any units in the Stream B population (i.e.$\underset{\sim}{z}_i^{(T)}$ for units in U), since these data items are not reported to ATO.

5. The Estimation Problem

17. The key issues that need to be considered when using the BIT data in estimation are described in Section 3. Some of these issues translate into the 'types of missing data' in the data described in Section 4. In summary, the key issues that need to be resolved when combining BIT data with the directly collected EAS data are:

1. What are the key set of output data items (BIT or EAS)?
2. Can an imputation or estimation methodology be devised to accurately account for the incompleteness of the BIT data with respect to the survey frame population ("first type of missing data")?
3. Can an estimation methodology be devised to produce population estimates for the broad and fine EAS data items from the EAS sample ("second type of missing data")?
4. If the BIT data items are the key output data items, can an estimation methodology be devised to produce estimates of the fine level BIT data items not covered by the BIT returns ("third type of missing data")?
5. Can an estimation methodology be devised to produce estimates in a timely manner (i.e. within twelve months of the end of the reference period) without significant impacts on the accuracy of the estimates?

18. Two alternative estimation methodologies, together with the assumptions underpinning these estimation methodologies, are presented in this paper. The current EAS/Tax estimation strategy is introduced in Section 6, while the Generalised Regression (GREG) Estimator which is currently being implemented is introduced in Section 7. Section 8 provides some results of several studies conducted to investigate options for replacing the current EAS/Tax estimation methodology with a GREG estimation methodology.

6. Current EAS/Tax Estimation Methodology

19. The current EAS/Tax estimation methodology uses the BIT data and the directly collected EAS data to calculate estimates for a wide variety of broad level and fine level data items. The methodologies used to produce these estimates have been described in June 2003 MAC paper 'Strategies for Synthetic Estimation in ABS Business Surveys', by Robert Clark and James Chipperfield.

6.1 Current Estimation Methodology for Broad Level Data Items

20. The first objective is to estimate the population totals for the broad level data items. There is some anecdotal evidence to indicate that:

Assumption 1: The broad level BIT data items are better quality than the broad level EAS data items.

21. Under Assumption 1, the problem is to estimate $\underset{\sim}{Y}=\sum_{i\in U}\underset{\sim}{x}_i^{(T)}$. Since $Y=\underset{\sim}{Y}_m+\underset{\sim}{Y}_{\bar{m}}$, where $\underset{\sim}{Y}_m=\sum_{i\in U_m}\underset{\sim}{x}_i^{(T)}$ which can be observed by matching the units in the Stream B population to the BIT data via ABN, and extracting the broad level BIT data items for the matched units, and $\underset{\sim}{Y}_{\bar{m}}=\sum_{i\in U_{\bar{m}}}\underset{\sim}{x}_i^{(T)}$ needs to be estimated since the broad level BIT data items are not available for the non-matched units in the Stream B population (i.e. $\underset{\sim}{x}_i^{(T)}$ for all units in $U_{\bar{m}}$).

22. The simplest solution to estimating $\underset{\sim}{Y}_m$ is to stratify the population and predict the stratum mean value of the broad level BIT data items for the non-matched units by the stratum mean value of the broad level BIT data items for the matched units (i.e. $\hat{\underset{\sim}{Y}}_{\bar{m}}=\sum_{h=1}^{H}N_{h\bar{m}}\,\bar{\underset{\sim}{x}}_{hm}^{(T)}$ where $N_{h\bar{m}}$ is the number of non-matched units in the Stream B population in stratum h and $\bar{\underset{\sim}{x}}_{hm}^{(T)}$ is the mean value of the broad level BIT data items for the matched units in stratum h). This solution is based on the assumption that $E\left(\underset{\sim}{x}_i^{(T)}/i\in U_{\bar{m}}\right)=E\left(\underset{\sim}{x}_i^{(T)}/i\in U_m\right)$. However, in practice the ABS has found that a higher proportion of non-matched units are dead compared to matched units. Units that are no longer operational (i.e. dead) are no longer obliged to report BIT data, and hence most dead units in the population will not match to the BIT data. In order to take account of the different proportions of live units in U_m and $U_{\bar{m}}$, a solution to estimating $\underset{\sim}{Y}_{\bar{m}}$ is sought which is based on the assumption:

Assumption 2: $E\left(\underset{\sim}{x}_i^{(T)}/i\in U_{l\bar{m}}\right)=E\left(\underset{\sim}{x}_i^{(T)}/i\in U_{lm}\right)$.

where U_{lm} are those live units matched to BIT data and $U_{l\bar{m}}$, are those live units not matched to the BIT data.

23. Under Assumption 2, $\underset{\sim}{Y}_{\bar{m}}$ is estimated by:

$$\hat{\underset{\sim}{Y}}_{\bar{m}}=\sum_{h=1}^{H}N_{hl\bar{m}}\,\bar{\underset{\sim}{x}}_{hlm}^{(T)}$$

where $N_{hl\bar{m}}$ is the number of live non-matched units in the Stream B population in stratum h and $\bar{\underset{\sim}{x}}_{hlm}^{(T)}$ is the mean value of the broad level BIT data items for the live matched units in stratum h which can be predicted by:

$$\bar{\underset{\sim}{x}}_{hlm}^{(T)}=\frac{N_{hm}}{N_{hlm}}\bar{\underset{\sim}{x}}_{hm}^{(T)}$$

and hence $\underset{\sim}{Y}_{\bar{m}}$ is estimated by:

$$\hat{\underset{\sim}{Y}}_{\bar{m}}=\sum_{h=1}^{H}N_{hl\bar{m}}\frac{N_{hm}}{N_{hlm}}\bar{\underset{\sim}{x}}_{hm}^{(T)}=\sum_{h=1}^{H}\frac{N_{hl\bar{m}}}{N_{hm}}\frac{N_{hm}}{N_{hlm}}\sum_{i\in U_{hm}}\underset{\sim}{x}_i^{(T)}=\sum_{h=1}^{H}f_h\left(\frac{N_{h\bar{m}}}{N_{hm}}\right)\sum_{i\in U_{hm}}\underset{\sim}{x}_i^{(T)}$$

where $f_h=\frac{N_{hl\bar{m}}}{N_{h\bar{m}}}\frac{N_{hm}}{N_{hlm}}$ is referred to as the live factors.

24. To enable a more stable estimator of the live factors, these live factors are estimated at a broad industry by size level f_s (i.e. broader than the stratification used for estimating the mean values), and hence it is assumed that:

Assumption 3: $f_s = f_h$ for all strata h within broad strata s.

25. Under Assumption 3, the broad level data items $\underset{\sim}{Y}$ are estimated by:

$$\underset{\sim}{\hat{Y}} = \sum_{h=1}^{H} \left(1 + f_h\left(\frac{N_{h\bar{m}}}{N_{hm}}\right)\right) \sum_{i\in U_{hm}} \underset{\sim}{x}_i^{(T)}$$

26. Historically, the components of the live factor have been estimated using the BIT data and the EAS data. The number of matched and non-matched units in the Stream B population were estimated from the BIT data, while the number of live matched and live non-matched units in the Stream B population were estimated from the EAS data. However, the introduction of the New Tax System in July 2000 has provided the ABS with the opportunity to use Business Activity Statement (BAS) for statistical purposes. Businesses are obliged to report their GST and other key tax obligations to the ATO via a BAS return, on either a monthly, quarterly or annual basis, depending on the size of the business. The BAS data can be used to help identify the operating status of non-matched businesses. This information is currently used to estimate the number of live matched and live non-matched units in the Stream B population used in the live factor.

27. In order to reduce the possible biases associated with Assumption 3, the live factors could be calculated at finer levels. However, this will lead to greater variability in the estimates between time periods. The biases associated with Assumption 3 can also be reduced by using the most complete set of BIT data returns available (i.e. reducing the number of non-matched units). Unfortunately, this would lead to lengthy delays in the release of estimates. The ABS currently uses the BIT data available twelve months after the end of the reference period, resulting in EAS/Tax estimates being released approximately 18 months after the end of the reference period. In practice, there may be other important factors that influence the non-matching of business to the BIT data, which are unrelated to their live-dead status of the business. If any of these assumptions fail to hold this will lead to biases in the estimates.

28. The current EAS/Tax variance estimator for the EAS/Tax estimator of the broad level BIT data items is given by:

$$Var\left(\underset{\sim}{\hat{Y}}\right) = \sum_{h=1}^{H} N_{hm}\left(1 - \frac{N_{hm}}{N_h}\right)\left(1 + f_s\left(\frac{N_{h\bar{m}}}{N_{hm}}\right)\right)^2 S_h^2$$

where $S_h^2 = \frac{1}{n_{hm}-1} \sum_{i=1}^{n_{hm}} \left(\underset{\sim}{x}_i^{(T)} - \underset{\sim}{\bar{x}}_h^{(T)}\right)^2$.

29. This variance formula ignores the variances associated with the live factor, and hence underestimates the true variance.

6.2 Current Estimation Methodology for Fine Level Data Items

30. The second objective is to estimate the population totals for the fine level data items. Under Assumption 1, the problem is to estimate $\underset{\sim}{Y}=\sum_{i\in U}\underset{\sim}{z}_i^{(T)}$. The fine level data items are estimated in a similar fashion to the broad level data items, except that the fine level BIT data items need to be imputed for all units, since these data items are not available for any units in the Stream B population (i.e.$\underset{\sim}{z}_i^{(T)}$ for units in U). The fine level BIT data items can be imputed by multiplying the broad level BIT data items by the ratio of the estimates for the fine level EAS data items over the estimates for the broad level EAS data items:

$$\underset{\sim}{z}_i^{(T)} = \frac{\underset{\sim}{\hat{Z}}_r^{(E)}}{\underset{\sim}{\hat{X}}_r^{(E)}}\underset{\sim}{x}_i^{(T)}$$

where $\underset{\sim}{\hat{X}}_r^{(E)}$ and $\underset{\sim}{\hat{Z}}_r^{(E)}$ are estimates for the broad and fine level EAS data items using the Horvitz-Thompson estimator at a broad industry by size level r:

$$\underset{\sim}{\hat{Z}}_r^{(E)} = \sum_{i\in s_r} w_i^{(E)}\underset{\sim}{z}_i^{(E)}$$

$$\underset{\sim}{\hat{X}}_r^{(E)} = \sum_{i\in s_r} w_i^{(E)}\underset{\sim}{x}_i^{(E)}$$

where $w_i^{(E)} = 1/\pi_i$ are the EAS sampling weights.

31. This imputation method makes the assumption that:

Assumption 4: The ratio $\frac{\underset{\sim}{\hat{Z}}_r^{(E)}}{\underset{\sim}{\hat{X}}_r^{(E)}}$ is a good approximation of the ratio $\frac{\underset{\sim}{z}_i^{(T)}}{\underset{\sim}{x}_i^{(T)}}$.

32. Under Assumption 4, the fine level data items $\underset{\sim}{Y}$ are estimated by:

$$\underset{\sim}{\hat{Y}}=\sum_{h=1}^{H}\left(1+f_h\left(\frac{N_{h\overline{m}}}{N_{hm}}\right)\right)\left(\frac{\underset{\sim}{\hat{Z}}_r^{(E)}}{\underset{\sim}{\hat{X}}_r^{(E)}}\right)\sum_{i\in U_{hm}}\underset{\sim}{x}_i^{(T)}$$

33. The quality of the current EAS/Tax estimator for the fine level BIT data items will depend on the quality of the current EAS/Tax estimator for the broad level BIT data items, and the ability of the proration factor to accurately impute the fine level EAS data items for the non-sampled units in the Stream B population (i.e. "**the second type of missing data**") and then to accurately impute the unavailable fine level BIT data items in the Stream B population (i.e. "**the third type of missing data**").

34. This current proration method is an implicit form of modelling, where the assumptions underlying the model are not explicitly defined as they would be in a formal model-based approach. The fine level data items for the units not in sample can be estimated by taking an aggregate (industry by size) level proration factor, and then multiplying the broad level BIT data items by this proration factor. If Assumption 4 fails to hold this will lead to biases in the estimates. In order to reduce these biases, the proration factors could be calculated at finer levels, but this leads to greater variability in the proration factor and hence greater variability in the estimates. Furthermore, it is possible that one or two generally large and aberrant units will dominate the proration factor. This leads to a need for outliering to be performed, which in turn leads to unquantifiable biases in the estimates.

35. Another deficiency of the current EAS/Tax estimation methodology is that due to the complexity of the current EAS/Tax estimator for the fine level BIT data items, no variances are calculated for these estimates. A tailorisation approach could be used to obtain a variance estimator:

$$
\begin{aligned}
Var\left(\underset{\sim}{\hat{Y}}\right) &= \sum_r Var\left(\underset{\sim}{\hat{Y}}_r^{(T)} \frac{\underset{\sim}{\hat{Z}}_r^{(E)}}{\underset{\sim}{\hat{X}}_r^{(E)}} \right) \\
&= \sum_r \left(\frac{\underset{\sim}{\hat{Z}}_r^{(E)}}{\underset{\sim}{\hat{X}}_r^{(E)}} \right)^2 Var\left(\underset{\sim}{\hat{Y}}_r^{(T)} \right) \\
&\quad + \sum_r \left[\left(\underset{\sim}{\hat{Y}}_r^{(T)} \right)^2 + Var\left(\underset{\sim}{\hat{Y}}_r^{(T)} \right) \right] \left[\frac{\left(\underset{\sim}{\hat{Z}}_r^{(E)} \right)}{\left(\underset{\sim}{\hat{X}}_r^{(E)} \right)} \right]^2 \left[\frac{Var\left(\underset{\sim}{\hat{Z}}_r^{(E)} \right)}{\left(\underset{\sim}{\hat{Z}}_r^{(E)} \right)^2} + \frac{Var\left(\underset{\sim}{\hat{X}}_r^{(E)} \right)}{\left(\underset{\sim}{\hat{X}}_r^{(E)} \right)^2} - \frac{2Cov\left(\underset{\sim}{\hat{Z}}_r^{(E)}, \underset{\sim}{\hat{X}}_r^{(E)} \right)}{\left(\underset{\sim}{\hat{Z}}_r^{(E)} \right)\left(\underset{\sim}{\hat{X}}_r^{(E)} \right)} \right]
\end{aligned}
$$

where $\underset{\sim}{\hat{Y}}_r^{(T)}$ is the current EAS/Tax estimator for the broad level BIT data items at a broad industry by size level r.

QUESTION 1: Are the assumptions underlying the current EAS/Tax estimation methodology to account for the different types of missing data justifiable and sufficient? Should further investigations be undertaken into the impacts of the missing data on the quality of the estimates? How should these impacts on the quality of the estimates be relayed to users?

6.3 Current Estimation Methodology for Small Domains

36. There is interest from users to produce estimates at fine industry levels. The current EAS/Tax estimator for the broad level BIT data items for (fine industry) domain d is given by:

$$\underset{\sim}{\hat{Y}}_d = \sum_{h=1}^{H} \left(1 + f_h\left(\frac{N_{h\bar{m}}}{N_{hm}}\right)\right) \sum_{i \in U_{hmd}} \underset{\sim}{x}_i^{(T)}$$

while the current EAS/Tax estimator for the fine level BIT data items for (fine industry) domain d is given by:

$$\underset{\sim}{\hat{Y}}_d = \sum_{h=1}^{H} \left(1 + f_h\left(\frac{N_{h\bar{m}}}{N_{hm}}\right)\right)\left(\frac{\underset{\sim}{\hat{Z}}_r^{(E)}}{\underset{\sim}{\hat{X}}_r^{(E)}}\right) \sum_{i \in U_{hmd}} \underset{\sim}{x}_i^{(T)}$$

37. The assumptions underpinning the current EAS/Tax estimator for small domains are similar to those assumptions made in Sections 6.1 and 6.2. In particular, Assumption 3 implies that the live factors are constant across the (fine industry) domains of interest within the (broad industry) aggregate levels, and Assumption 4 implies that the ratios are constant across the (fine industry) domains of interest within the (broad industry) aggregate levels.

6.4 Summary of Current EAS/Tax Estimation Methodology

38. In summary, the current EAS/Tax estimation methodology can be used to:

1. Produce estimates for the broad level and fine level data items;
2. Produce approximate variances for the broad level data items (but not the fine level data items);
3. Produce estimates (and variances for the broad level data items) for small domains;
4. Produce estimates based on the broad level BIT data items (but not broad level EAS data items) as the variable of interest; and
5. Produce untimely estimates, since BIT data is available twelve months after the reference period.

39. The current EAS/Tax estimation methodology is based on many assumptions, that may not hold in practice, which could lead to biases to the estimates.

7. Generalised Regression Estimator

40. The generalised regression (GREG) estimator is an alternative estimation methodology which uses the BIT data and the directly collected EAS data to produce estimates for the broad level and fine level data items. The GREG estimation methodology involves firstly imputing the broad level BIT data items for the non-matched units in the Stream B population, and secondly producing estimates using a GREG estimator with the broad level BIT data items as the auxiliary variables.

41. Consider a finite population $U = \{1, ..., i, ..., N\}$, from which a probability sample $s(s \subseteq U)$ is drawn according to a sample design with selection probabilities $\pi_i = \Pr(i \in s)$. The sampling weights $w_i = 1/\pi_i$ are those used in the Horvitz-Thompson estimator, $\hat{Y}_\pi = \sum_{i \in s} w_i y_i$, for variable of interest y. The objective is to estimate the population total $Y = \sum_{i \in U} y_i$, where y_i is the value of the variable of interest y for unit i. Assume there exists a set of auxiliary variables $\underset{\sim}{x}_i = (x_{1i}, ..., x_{ki}, ...x_{Ki})'$ for which the population totals $X = \sum_{i \in U} x_i$ are known. The generalised regression estimator is given by (Sarndal, Swensson and Wretman, 1992):

$$\hat{Y}_{GREG} = \sum_{i \in s} w_i y_i + \left(\underset{\sim}{X} - \sum_{i \in s} w_i x_i \right)' \underset{\sim}{\hat{\beta}}$$

where $\underset{\sim}{\hat{\beta}}$ is the vector of the linear regression model parameters given by:

$$\underset{\sim}{\hat{\beta}} = \left(\sum_{i \in s} \frac{w_i \underset{\sim}{x}_i \underset{\sim}{x}_i'}{c_i} \right)^{-1} \left(\sum_{i \in s} \frac{w_i \underset{\sim}{x}_i y_i}{c_i} \right)$$

and c_i are specified positive factors that relate to the variance structure of the linear regression model associated with the GREG estimator:

$$y_i = \underset{\sim}{x}_i' \underset{\sim}{\hat{\beta}} + \varepsilon_i$$

where $E(\varepsilon_i) = 0$, $Var(\varepsilon_i) = c_i \sigma^2$ and $Cov(\varepsilon_i, \varepsilon_j) = 0$, for all $i \neq j$.

42. The generalised regression estimator is often written as:

$$\hat{Y}_{GREG} = \sum_{i \in s} w_i g_i y_i$$

where g_i is the g-weight for unit i, defined as:

$$g_i = \left(1 + \left(\underset{\sim}{X} - \sum_{i \in s} w_i x_i \right) \left(\sum_{i \in s} \frac{w_i \underset{\sim}{x}_i \underset{\sim}{x}_i'}{c_i} \right)^{-1} \frac{\underset{\sim}{x}_i}{c_i} \right)$$

43. The variance of the generalised regression estimator is given by:

$$Var(\hat{Y}_{GREG}) = \sum_{i \in s} \sum_{j \in s} \left(\frac{\pi_{ij} - \pi_i \pi_j}{\pi_{ij}} \right) \left(\frac{g_i e_i}{\pi_i} \right) \left(\frac{g_j e_j}{\pi_j} \right)$$

where $e_i = y_i - \underset{\sim}{x}_i' \underset{\sim}{\hat{\beta}}$ are the residuals for the linear regression model associated with the GREG estimator.

44. Under the GREG estimation methodology, the broad level BIT data items would be used as the auxiliary variables for the purpose of estimation of EAS data items. Since the GREG estimation methodology requires auxiliary data to be available for units in the population, the auxiliary variables would need to be imputed for the non-matched units in the Stream B population (i.e. "**the first type of missing data**"). There are a number of imputation methods which could be used to impute these non-matched units. Four possible imputation methods are:

> **Mean imputation**: where the non-matched units are imputed using the mean of broad level BIT data items for the matched units:
>
> $$\hat{\underset{\sim}{x}}_i^{(T)} = \bar{\underset{\sim}{x}}_m^{(T)} \qquad \text{, if } i \in U_{\bar{m}}$$
>
> **Live mean imputation**: where the non-matched units are imputed using the live mean of the broad level BIT data items for the matched units if the unit is identified as live using the BAS data ($\delta_i = 1$) and zero if the unit is identified as dead using the BAS data ($\delta_i = 0$):
>
> $$\hat{\underset{\sim}{x}}_i^{(T)} = \delta_i \, \bar{\underset{\sim}{x}}_{lm}^{(T)} \qquad \text{, if } i \in U_{\bar{m}}$$
>
> **Ratio imputation**: where the non-matched units are imputed using the ratio of the mean of the broad level BIT data items over the mean of an correlated auxiliary variable for the matched units $\bar{b}_m$ multiplied by the auxiliary variable for the non-matched unit b_i:
>
> $$\hat{\underset{\sim}{x}}_i^{(T)} = \frac{\bar{\underset{\sim}{x}}_m^{(T)}}{\bar{b}_m} b_i \qquad \text{, if } i \in U_{\bar{m}}$$
>
> **Live ratio imputation**: where the non-matched units are imputed using the ratio of the live mean of the broad level BIT data items over the live mean of an correlated auxiliary variable for the matched units $\bar{b}_{lm}$ multiplied by the auxiliary variable for the non-matched unit if the unit is identified as live using BAS data and zero if the unit is identified as dead using the BAS data:
>
> $$\hat{\underset{\sim}{x}}_i^{(T)} = \delta_i \frac{\bar{\underset{\sim}{x}}_{lm}^{(T)}}{\bar{b}_{lm}} b_i \qquad \text{, if } i \in U_{\bar{m}}$$

45. The assumptions underpinning the imputation methods for the broad level BIT data items under the GREG estimator are likely to be similar to Assumptions 2 and 3. However, if these assumptions fail to hold this will lead to greater variability in the estimates for the EAS data items, which could still be unbiased asymtotically, unlike the current EAS/Tax estimation methodology which leads to biases in the estimates. Nonetheless, the better the methodology for imputing the broad level BIT data items for non-matched units, the lower the variability in the estimates.

7.1 GREG Estimator for Broad Level Data Items

46. Under the EAS/TAX framework, the broad level data items of interest are available from the BIT data as well as the directly collected EAS data. The current EAS/Tax estimation methodology relies on Assumption 1, since it uses the broad level BIT data items as the variable of interest. However, the GREG estimation methodology is less reliant on Assumption 1, as it can also use the broad level EAS data items as the variable of interest. Therefore there are a number of options for producing estimates for these broad level data items using a GREG estimator. The first three options are:

Option 1: Use the broad level BIT data items as the variables of interest.

$$\hat{\underset{\sim}{Y}}^{(1)} = \sum_{i\in s} w_i g_i \, \hat{\underset{\sim}{x}}_i^{(T)}$$

Option 2: Use the broad level EAS data items as the variables of interest.

$$\hat{\underset{\sim}{Y}}^{(2)} = \sum_{i\in s} w_i g_i \, \underset{\sim}{x}_i^{(E)}$$

Option 3: Use the broad level BIT data items as the variables of interest for the matched units and use the broad level EAS data items as the variables of interest for the non-matched units.

$$\hat{\underset{\sim}{Y}}^{(3)} = \sum_{i\in s_m} w_i g_i \, \hat{\underset{\sim}{x}}_i^{(T)} + \sum_{i\in s_{\overline{m}}} w_i g_i \, \underset{\sim}{x}_i^{(E)}$$

where $g_i = \left(1 + \left(\underset{\sim}{X}^{(T)} - \sum_{i\in s} w_i \, \hat{\underset{\sim}{x}}_i^{(T)}\right)\left(\sum_{i\in s} \frac{w_i \, \hat{\underset{\sim}{x}}_i^{(T)} \, \hat{\underset{\sim}{x}}_i^{(T)'}}{c_i}\right)^{-1} \frac{\hat{\underset{\sim}{x}}_i^{(T)}}{c_i}\right)$ and $\hat{\underset{\sim}{X}}^{(T)} = \sum_{i\in U} \hat{\underset{\sim}{x}}_i^{(T)}$.

47. The advantage under Option 1 and 3 is that for the matched units the contribution to the variance for these matched units will be zero (i.e. $\hat{\underset{\sim}{Y}}_m = \underset{\sim}{Y}_m$). On the other hand, there will not always be a perfect relationship between the variable of interest and the auxiliary variables under Option 2, because of the reporting differences between the BIT data and the EAS data. The advantage of Option 2 over Option 1 is that the variables of interest do not need to be imputed. If the imputation assumptions fail to hold then Option 1 could lead to substantial biases to the estimates. Option 1 and 3 would be justified if there are valid reasons to accept Assumption 1, that the broad level BIT data items are better quality than the broad level EAS data items; whereas Option 2 would be justified if Assumption 1 fails to hold. Option 3 is harder to interpret than that from Options 1 or 2 as the estimator is based on different sources for the matched and non matched units of the Stream B population.

QUESTION 2: On what basis should a decision be made on whether to use the broad level BIT data items or broad level EAS data items as the variables of interest?

48. Options 1 and 2 can be amended by relying on a post-stratified estimator, rather than a GREG estimator, for the non-matched units in the Stream B population. These two options are:

> **Option 4**: Use the GREG estimator with the broad level BIT data items as the variables of interest (i.e. Option 1) for the matched units and use a post-stratified estimator for the non-matched units.
>
> $$\hat{\underset{\sim}{Y}}^{(4)} = \sum_{i \in s_m} w_i g_{mi} \underset{\sim}{x}_i^{(T)} + \sum_{i \in s_{\overline{m}}} w_i g_{\overline{m}i} \underset{\sim}{x}_i^{(E)}$$
>
> **Option 5**: Use the GREG estimator with the broad level EAS data items as the variables of interest (i.e. Option 2) for the matched units and use a post-stratified estimator for the non-matched units.
>
> $$\hat{\underset{\sim}{Y}}^{(5)} = \sum_{i \in s_m} w_i g_{mi} \underset{\sim}{x}_i^{(E)} + \sum_{i \in s_{\overline{m}}} w_i g_{\overline{m}i} \underset{\sim}{x}_i^{(E)}$$

where $g_{mi} = \left(1 + \left(\underset{\sim}{X}_m^{(T)} - \sum_{i \in s_m} w_i \underset{\sim}{x}_i^{(T)} \right) \left(\sum_{i \in s_m} \frac{w_i \underset{\sim}{x}_i^{(T)} \underset{\sim}{x}_i^{(T)'}}{c_i} \right)^{-1} \frac{\underset{\sim}{x}_i^{(T)}}{c_i} \right)$, $g_{\overline{m}i} = \frac{n_h}{N_h} \frac{N_{h\overline{m}}}{n_{h\overline{m}}}$ and $\hat{\underset{\sim}{X}}^{(T)} = \sum_{i \in U_m} \underset{\sim}{x}_i^{(T)}$.

49. The advantage of Option 4 over Option 1 is that the broad level BIT data items do not need to be imputed for the non-matched units in the Stream B population. Option 4 would be justified if Assumption 2 fails to hold, while Option 5 would be justified if Assumption 1 fails to hold.

> QUESTION 3: What is the best approach to choosing between the options for producing estimates using a GREG estimator?

50. Under Option 2 and Option 5 the broad level EAS data items are the variables of interest, and hence these options are open to using alternative auxiliary variables. For example, it is possible to improve the timeliness of the release of the estimates by using the previous years broad level BIT data items as the auxiliary variables. This improvement in the timeliness of the estimates will come at the expense of greater variability in the estimates, as the relationship between the variable of interest in one year and the auxiliary variables in the previous year will be weaker. Under Option 1, Option 3 and Option 4 the broad level BIT data items are the variable of interest and hence the current years broad level BIT data items need to be used.

51. Under the GREG estimation methodology, a number of choices need to be made about the exact implementation of the methodology. Firstly, a decision needs to be made on whether to have a single weight for all of the variables of interest or multiple weights (i.e. a separate weight for each of the variables of interest). While the multiple weight option will provide more accurate estimates for the individual variables of interest, the single weight option will provide more consistent estimates across the entire set of variables of interest as it preserves the relationships reported within each individual business - an important requirement for National Accounts.

52. Secondly, under the single weight option, a decision needs to be made on which of the broad level BIT data items to include in the model, since in practice it is not always possible to include all the broad level BIT data items in the model and still guarantee the weights are greater than or equal to one. In this situation the estimates will be more accurate for the chosen broad variables and less accurate for the other variables.

53. Thirdly, under both the single and multiple weight options, a decision needs to be made on which level the model is fitted (i.e. calibration level). The finer the level at which the model is fitted, the greater the variability in the estimates at the broader levels, while the broader the level at which the model is fitted, the greater the biases in the estimates at the finer levels.

QUESTION 4: In a survey where there are multiple related variables of interest, on what basis should a decision be made on whether to have a single weight for all of the variables of interest or multiple weights (i.e. a separate weight for each of the variables of interest)?

7.2 GREG Estimator for Fine Level Data Items

54. Under the EAS/TAX framework, the fine level data items of interest are only available from the EAS data. The method of producing estimates for these fine level data items using a GREG estimator, will depend on the method of producing estimates for the broad level data items.

55. When using the broad level EAS data items are used as the variables of interest (i.e. Options 2 and 5, and the non-matched units of Options 3 and 4), the GREG estimator would simply use the fine level EAS data items as the variables of interest. However, when using the broad level BIT data items are the variables of interest (i.e. Option 1, and the matched units of Options 3 and 4), the fine level BIT data items would need to be imputed. The fine level BIT data items can be imputed by multiplying the broad level BIT data items by the ratio of the fine level EAS data items over the broad level EAS data items:

$$\hat{\underset{\sim}{z}}_i^{(T)} = \frac{\underset{\sim}{z}_i^{(E)}}{\underset{\sim}{x}_i^{(E)}}\,\hat{\underset{\sim}{x}}_i^{(T)}$$

and hence the GREG estimator is given by:

$$\underset{\sim}{\hat{Y}}_{GREG} = \sum_{i \in s} w_i g_i \, \underset{\sim}{\hat{z}}_i^{(T)}$$

56. This imputation method makes the assumption that:

> **Assumption 4A:** The ratio $\dfrac{\underset{\sim}{z}_i^{(E)}}{\underset{\sim}{x}_i^{(E)}}$ is a good approximation of the ratio $\dfrac{\underset{\sim}{z}_i^{(T)}}{\underset{\sim}{x}_i^{(T)}}$.

57. If Assumption 4A fails to hold this will lead to biases in the estimates.

7.3 GREG Estimator for Small Domains

58. While it is possible to produce estimates for small domains from the current GREG estimator (using a uni-weight estimator):

$$\underset{\sim}{\hat{Y}}_{UNIWGT,d} = \sum_{i \in s_d} w_i g_i \, \underset{\sim}{x}_i^{(*)}$$

there are several alternative estimators which are likely to be more suitable for producing estimates for small domains, such as the linear prediction generalised regression estimator (Estevao and Sarndal, 1999):

$$\underset{\sim}{\hat{Y}}_{LINPRED,d} = \sum_{i \in s_d} w_i \, \underset{\sim}{x}_i^{(*)} + \left(\underset{\sim}{X}^{(T)} - \sum_{i \in s} w_i \, \underset{\sim}{x}_i^{(T)} \right) \underset{\sim}{\hat{\beta}}$$

and the generalised regression synthetic estimator:

$$\underset{\sim}{\hat{Y}}_{SYN,d} = \underset{\sim}{X}^{(T)} \, \underset{\sim}{\hat{\beta}}$$

where $\underset{\sim}{x}_i^{(*)}$ is the chosen variable of interest.

59. Initial empirical studies have shown that linear prediction generalised regression estimator and the generalised regression synthetic estimator provide significant reductions in variances. While the generalised regression synthetic estimator is known to contain a large bias component, when this was taken into account in the empirical studies this estimator still provided significant improvements over the current GREG estimator (using a uni-weight estimator). The assumptions underlying the linear prediction generalised regression estimator and the generalised regression synthetic estimator are:

> **Assumption 5:** The relationships between the broad level and fine level data items are consistent across the small domains.

60. If assumption 5 fails to hold this will lead to biases in the estimates.

7.4 Summary of GREG Estimation Methodology

61. In summary, the GREG estimation methodology can be used to:

1. Produce estimates for the broad level and fine level data items;
2. Produce variances for the broad level and fine level data items;
3. Produce estimates and variances for small domains;
4. Produce estimates based on either the broad level BIT or EAS data items as the variable of interest; and
5. Produce more timely but less accurate estimates (under Option 2 or Option 5) based on using historical broad level BIT data items as auxiliary variables.

62. The GREG estimation methodology is based on many assumptions, that may not hold in practice, which could lead to either biases to the estimates or larger variances depending on the nature of the chosen GREG estimation methodology.

8. Results

63. Over the last couple of years, two separate studies, one of which is still underway, have been conducted to investigate options for replacing the current EAS/Tax estimation methodology with a GREG estimation methodology. The first study was conducted on data from the 2000/01 financial year, while the latest study was conducted on data from the 2001/02 and 2002/03 financial years.

8.1 Investigations on 2000/01 EAS/Tax Estimates

64. The ABS Business Register used for the selection and estimation of the 2000/01 EAS/Tax estimates predates the 'two population model' introduced in Section 2. However, it was still possible to identify units with a simple structure that could be matched to BIT data. The methodology adopted for these units was similar to the current EAS/Tax estimation methodology used to produce estimates for the current Stream B component of the population. Unfortunately Option 2 for the GREG estimator was not considered at the time of this study.

65. The estimates and relative standard errors for broad level data item "Total Income" for several selected "example" industries are presented in Table 1. The selected "example" industries were chosen as representative of the three general outcomes observed for Total Income over a large number of industries examined in the study:

1. The accuracy of the estimates produced by the GREG estimation methodology were significantly better than the current EAS estimates (Example A).
2. The accuracy of the estimates produced by the GREG estimation methodology were significantly worse than the current EAS estimates (Example B).
3. The accuracy of the estimates produced by the GREG estimation methodology were similar to the current EAS estimates (Example C).

66. To provide an overall comparison of the estimation methodologies, the total estimates across all industries included in the study is also presented.

Table 1: Estimates and Relative Standard Errors of Total Income for Selected Key Industries, 2000/01

Selected "Example" Industry	Current EAS	Current EAS/Tax	GREG Option 1	GREG Option 2	GREG Option 3	GREG Option 4	GREG Option 5
	Estimates						
A	43,464.4	50,690.0	51,120.3	N.A.	49,700.1	48,159.5	47,616.0
B	10,927.9	21,520.6	16,468.0		16,462.0	20,087.3	19,258.3
C	6,061.4	5,773.5	7,772.8		7,495.9	7,342.3	6,969.5
Total	**334,222.3**	**410,224.6**	**409,060.8**		**397,007.3**	**400,152.8**	**383,336.7**
	Relative Standard Errors						
A	6.8	0.6	2.5	N.A.	2.4	1.6	3.2
B	10.3	4.8	26.0		26.0	21.0	20.9
C	10.7	2.7	10.3		10.0	9.4	9.9
Total	**3.1**	**0.5**	**2.2**		**2.1**	**2.0**	**2.2**

67. The relative standard errors under the current EAS/Tax estimation methodology are significantly lower than those under the GREG estimation methodology. However, it is important to remember that the current EAS/Tax estimates are biased, so ideally a comparison should be made of the relative mean squared error. Option 3 and Option 4 of the GREG estimator were designed to minimise some of the biases, and hence these estimates could be used to provide a rough indication of the likely biases in the current EAS/TAX estimates. For example, suppose Option 4 of the GREG estimator is very close to an unbiased estimate of the true population value, then the relative mean squared error of the current EAS/Tax estimates across all industries would be 2.5%, which is slightly larger than the relative standard errors for the GREG estimators. These results illustrate that although the bias in the current EAS/Tax estimates are not measured, they are likely to have significant non-ignorable biases and the current relative standard errors provide a poor indication of the quality of the current EAS/Tax estimates.

8.2 Investigations on 2001/02 EAS/Tax Estimates

68. The ABS Business Register used for the selection and estimation of the 2001/02 EAS/Tax estimates were based on the 'two population model' introduced in Section 2. Unfortunately not all of the options for the GREG estimator were considered in this investigation. The estimates and relative standard errors for several data items and several selected "example" industries are presented in Table 2. The selected data items were "Total Income" (a broad level data item included in GREG as an auxiliary variable), "Total Assets" (a broad level data item not included in GREG as an auxiliary variable), "Sales of Goods and Services" (a fine level data item) and "Industry Value Added" (a data item derived from fine level data items). The selected "example" industries were chosen as representative of the three general outcomes observed for Total Income over a large number of industries examined in the study:

1. The accuracy of the estimates produced by the GREG estimation methodology were significantly better than the current EAS estimates (Example D).
2. The accuracy of the estimates produced by the GREG estimation methodology were significantly worse than the current EAS estimates (Example E).
3. The accuracy of the estimates produced by the GREG estimation methodology were similar to the current EAS estimates (Example F).

Table 2: Estimates and Relative Standard Errors of Selected Data Items for Selected Key Industries, 2001/02

Data Item	Selected "Example" Industry	Current EAS		Current EAS/Tax		GREG Option 1		GREG Option 2	
		Estimates	RSE%	Estimates	RSE%	Estimates	RSE%	Estimates	RSE%
Total Income	D	37,603.8	10.3	42,177.2	0.6	42,766.8	0.0	44,646.9	4.9
	E	1,350.5	12.0	1,437.9	8.4	1,603.2	0.0	1,472.1	15.0
	F	15,992.4	9.1	22,701.4	2.4	21,539.8	0.0	19,100.0	9.7
Total Assets	D	17,422.3	15.0	17,543.1	1.4	16,405.1	6.4	20,148.2	12.5
	E	1,124.1	16.5	1,011.4	4.4	944.4	9.4	1,756.6	26.4
	F	7,081.3	12.3	10,293.2	2.2	10,686.1	7.5	8,718.6	18.0
Sales of Goods & Services	D	30,853.6	10.5	41,504.7	0.6	Results not available		43,846.7	4.8
	E	1,351.5	11.8	1,416.2	2.5			1,459.6	15.3
	F	15,890.6	9.2	22,529.7	2.4			27,359.8	7.0
Industry Value Added	D	7,147.1	10.3	7,463.7	0.6			7,538.1	10.0
	E	540.7	12.5	495.1	3.0			579.3	26.5
	F	3,187.3	11.0	3,357.5	1.7			3,718.9	10.1

69. In summary, these results indicate that for some industries and some data items there are significant gains to be made by using the BIT data under either the current estimation methodology or the proposed GREG estimation methodology. However, this may not be the case for all data items and for all industries. For data items not included as an auxiliary variable in the GREG estimator, but which are well correlated with one of the auxiliary variables (e.g. sales of goods and services is highly correlated with total income in most industries) there are considerable gains. For other data items not included as an auxiliary variable in the GREG estimator, which are not correlated with the auxiliary variables, there are no gains.

9 Conclusion

70. In summary, many of the assumptions under the GREG estimation methodology are similar to those under the current EAS/Tax "data substitution" methodology. Indeed, both methodologies require the imputation of the broad level BIT data items; to be used as variable of interest under the current EAS/Tax "data substitution" methodology and as auxiliary variables (Options 1, 2, 3, 4 and 5) and as variables of interest (Option 1) under the GREG estimation methodology. Generally speaking, if the assumptions fail to hold under the current EAS/Tax "data substitution" methodology then this will lead to biases in the estimates. However, if they fail to hold under the GREG estimation methodology this may lead to biases in the estimates (Option 1) or to larger variances (Options 2, 3, 4 or 5).

QUESTION 5: If it is not possible to accurately impute the broad level BIT data items for the non-matched units in the population, it is better to use the GREG estimation methodology (Options 2, 3, 4 or 5) or use the current EAS/Tax "data substitution" methodology?

71. Under the current EAS/Tax "data substitution" methodology there is more flexibility to optimise the estimates for each of the broad level BIT data items, while still maintaining the relationships reported within each individual business. The multiple weight option under the GREG estimation methodology will also optimise the estimates for each of the broad level BIT data items, but will not preserve the relationships reported within each individual business. On the other hand, the single weight option under the GREG estimation methodology will provide more consistent estimates across the entire set of variables of interest, but will provide less accurate estimates for each of the broad level BIT data items. Furthermore, the single weight option, it is not always possible to include all the broad level BIT data items in the model and still guarantee positive weights, and hence positive estimates. In this situation the estimates will be more accurate for the chosen broad variables and less accurate for the other variables.

QUESTION 6: If it is not possible to include all the broad level BIT data items into the GREG model, is it better to use the current EAS/Tax "data substitution" methodology or use the GREG estimation methodology?

72. It is possible to produce estimates for small domain estimates under both the current EAS/Tax "data substitution" methodology and the GREG estimation methodology, using either the linear prediction generalised regression estimator or the generalised regression synthetic estimator. The assumptions underlying the two methodologies are similar (i.e. the relationships between the broad level and fine level data items are consistent across the small domains).

QUESTION 7: If estimates are required for small domains, is it better to use the current EAS/Tax "data substitution" methodology or use the GREG estimation methodology?

References

Estevao, V.M. and Sarndal, C.E. (1999) The use of auxiliary information in design-based estimation for domains. *Survey Methodology*, 25, 213-221.

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

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

2000001523919
ISBN 0 642 48174 1

RRP $11.00

Produced by the Australian Bureau of Statistics

1352.0.55.070 • RESEARCH PAPER: ESTIMATION METHODOLOGIES USING TAXATION DATA FOR ABS BUSINESS SURVEYS (METHODOLOGY ADVISORY COMMITTEE) • June 2005