USE OF ADMINISTRATIVE DATA AS SUBSTITUTES FOR SURVEY DATA FOR SMALL ENTERPRISES IN THE SWEDISH ANNUAL STRUCTURAL BUSINESS STATISTICS

Johan Erikson and Lennart Nordberg, Statistics Sweden
Johan Erikson, Dept. of Economic Statistics, Statistics Sweden, S-701 89 Örebro, Sweden
johan.erikson@scb.se

ABSTRACT

This paper presents some Swedish experiences from the use of administrative data as substitute for survey data for statistical purposes. Two advantages with using administrative data is that more detailed presentation is possible and that better precision of estimates of characteristics that were weakly correlated to the sampling indicator is gained. Also, by using administrative data, several problems with the frame for Structural Business Statistics have been discovered and possibilities of correcting the frame for coverage problems have been improved. On the minus side, not all requested indicators are included in the administrative data. Therefore, a combination of data from questionnaires, administrative registers and other statistical surveys must be used to estimate the missing items.

Key words: Presentation, Precision, Frame errors, Estimation methods

1. INTRODUCTION

Adminstrative data have been used by Statistics Sweden in business surveys for a long time. The VAT register for information on turnover and the PAYE register for wages and salaries are the most commonly used sources. Traditional uses of these data have been in imputation for non-response and for compensation of estimates for bias due to cut-off sampling.

The present Structural Business Statistics (SBS), following the Eurostat Regulation on SBS was introduced in 1997. Previous to 1997, Statistics Sweden conducted two annual surveys on business structure: the Annual Survey of Manufacturing (ASM) and the Financial Accounts Survey (FAS) which covered the whole corporate sector. The ASM surveyed on a take-all basis the local units belonging to manufacturing enterprises with 10 employees or more and all manufacturing local units with 10 employees or more of non-manufacturing enterprises. Adminstrative data were used for imputation and to cover (in a fairly rough way) the manufacturing businesses with less than 10 employees. This is a typical example of traditional uses of adminstrative data in surveys.

For the FAS, which by contrast to the ASM, used enterprises as observational units it was possible to make more extensive use of adminstrative data. While enterprises with 50 employees or more were required to fill in a questionnaire, the data used for the smaller enterprises (0 - 49 employees) were taken from public annual company reports collected from every enterprise in the sample. Although this procedure kept the response burden low for the businesses, it caused a lot of work for Statistics Sweden in reading and interpretation of the company reports.

Since the early 1990s the National Tax Board (National Revenue Service) requires, on an annual basis, from all Swedish enterprises so called SRUs ('Standardiserade RäkenskapsUtdrag' in Swedish, in English Standardised Company Reports). Although the main purpose of the SRUs is to assist the National Tax Board in its work, the material has also been available to Statistics Sweden for statistical purposes since 1993.

The SRU material is considerably richer than traditional adminstrative records such as PAYE and VAT. It is also, by contrast to the ordinary public annual company reports, available in computer readable form.

The SRUs contain more than 200 items (variables). In 1994/95 the SRU material was tested at Statistics Sweden for its potential as data source. As a result of these tests, the material is now used on a large scale for the SBS.

The quality of the SRU material is considered less reliable for larger enterprises than for smaller ones. One reason for this is that the National Tax Board does not use the SRUs for larger enterprises to the same degree as for the smaller ones. For this and a few additional reasons the SRUs are used as main data source in the SBS only for enterprises with less than 50 employees (about 280 000 enterprises) , while the larger enterprises (about 5 000) are still required to fill in a conventional postal questionnaire.

Table 1. The number of enterprises, employment and turnover in the part covered by SRU in proportion to the total population.

Industry sector	Number of Enterprises	Number of Employees	Turnover
	%	%	%
Mining (10-14)	95.6	20.9	22.1
Manufacturing (15-37):	94.1	23.7	16.1
Building (45):	99.3	58.1	53.7
Wholesale (51):	98.6	62.1	52.3
Retail trade (52):	99.2	56.5	50.7
All industries (01-93):	98.5	41.6	38.6

Remark: The numeration of industrial sector refers to the Swedish standard industrial classification SNI 92, equal to Nace Rev.1.

As seen from the table, the SRUs cover a very large proportion of the service sector but also an important part of manufacturing. Hence by introducing the SRU material as a substitute for direct data collection it has been possible to obtain a very large reduction of the response burden for businesses, and at the same time a considerable cost reduction for Statistics Sweden.

There are several additional advantages to the use of SRU:

Since the SBS is now a total enumeration - by contrast to its predecessor FAS which covered the smaller businesses by probability sampling - it is now possible to present the material in more detail. Regional breakdown and key ratios for small enterprises are two areas where this is important. Furthermore, estimates of characteristics like investments which in the old system often showed large variability within strata and for that reason were affected by a large variance can now be presented in more detail. This subject will be discussed in Section 2.

Several problems with the frame for the SBS have been discovered from the use of SRU, and it has been possible to correct the frame for some important covarage problems as a result of this. Section 3 contains more details on this matter.

Since the SRU does not include all items important to the survey, a combination of data from questionnaires, administrative registers and other statistical surveys must be used to estimate the missing items. Some Swedish experience on this matter is presented in Section 4.

2. ESTIMATION IN SMALL DOMAINS

There is no sampling error in the SBS since sampling is not used. However, there is about 15% unit non response among the SRUs and some item non response among the larger enterprises. Hence there is some random variation (and possibly some bias) stemming from the non response. Nevertheless, SBS-97 shows a substantial improvement in precision compared to the older FAS survey. The following table illustrates the difference.

Table 2 Estimated coefficient of variation (relative standard error) for estimates of number of employees and net investments in the FAS-89 and the SBS-97

Industry sector	Number of people employed		Net investments	
	FAS-89	SBS-97	FAS-89	SBS-97
Mining (10-14):	1.7 %	0.1 %	18.0 %	0.5 %
Manufacturing (15-37):	0.9 %	0.02 %	2.8 %	0.2 %
Building (45):	1.1 %	0.06 %	9.1 %	3.5 %
Wholesale (51):	1.5 %	0.05 %	5.9 %	1.2 %
Retail (52):	1.7 %	0.05 %	4.8 %	0.7 %
All industries (01-93):	0.5 %	0.01 %	2.6 %	0.7 %

The gain in precision from FAS to SBS would be even higher than indicated by table 2 if only the small businesses, say 0 - 19 employees, were considered. Unfortunately, it has not been possible to obtain the variances from the FAS for that part of the population. However, as an indication, we observe that the total sample in this group was about 8000 enterprises out of about 200 000 enterprises in the frame. With these 8000 units broken down to, say, 3 digit NACE groups, it should be obvious that estimates for 'difficult' characteristics like investments (with large variability within strata) would be quite poor in many cases.

There is currently a great demand for analysis of the small businesses in Sweden. The gain in precision of estimates for this group from SBS compared to FAS has made it easier to meet this demand.

Table 3 Estimated coefficient of variation (relative standard error) for estimates of number of employees, turnover and net investments in the SBS-97 for small enterprises (0 -19 employees).

Industry sector	Number of people employed	Turnover	Net investments
Mining (10-14):	0.5 %	2.0 %	2.5 %
Manufacturing (15-37):	0.06 %	0.5 %	1.5 %
Building (45):	0.07 %	4.6 %	4.2 %
Wholesale (51):	0.07 %	1.1 %	2.6 %
Retail (52):	0.08 %	0.3 %	1.1 %
All industries (01-93):	0.03 %	0.7 %	1.6 %

Precision of estimates from the SBS for small businesses is quite good as indicated in table 3, even for a ‘difficult’ characteristic such as investments. When the material is broken down further into detailed industries, such as 3- or 4-digit NACE groups, the precision may however in some cases become considerably lower. Nevertheless, the SRU material is a valuable source for data collection from and analysis of the small business community.

So far we have only mentioned random error. Other error sources may remain, such as frame error (see discussion ahead), response error or processing error made at either the National Tax Board or Statistics Sweden (processing errors are tackled by a fairly detailed editing procedure). There may also be systematic error (bias) due to the non response.The possibility of non response bias remains a problem which needs further attention in the future. The non responding units should be investigated in more depth, e.g by the help of other adminstrative data files such as VAT and PAYE. A further error source, model assumption error, occurs in the SRU material. This subject is discussed in Section 4.

3. QUANTIFYING FRAME ERRORS USING ADMINISTRATIVE DATA

The frame for the SBS survey is the Business Register at Statistics Sweden, the central register of enterprises and establishments (CFAR). The survey covers all enterprises in CFAR that are classified as active (the criteria for being classified as active are being registered as paying VAT, being registered as employer or being registered as paying corporate tax) and that belong to SNI 01-93 of the Swedish standard industrial classification (divisions A-O of Nace Rev. 1). The use of SRU as a source for SBS means that there is a new opportunity to check the quality of the frame, and also to quantify the frame errors regarding under coverage and surplus coverage.

For the corporate sector, the under coverage consists of enterprises that have sent in an SRU but that are not included in the CFAR frame. These enterprises can be divided into two categories; enterprises that are not included in the CFAR at all, and enterprises that are not classifed as active. The latter category is by far the largest. For sole proprietors the problem of under coverage is of a different kind. The most common problem concerns enterprises that have not received an industrial classification. These are not included in the frame, but might very well send in an SRU.

There is also a possibility of surplus coverage, consisting of enterprises that are classified as active in the CFAR but that do not send in an SRU. Since it is unknown whether these enterprises should have sent in an SRU, and that it is simply missing, or whether these enterprises no longer are active, it is not possible to say for certain that all these enterprises constitute surplus coverage, but the possibility exists.

In 1997 there were 88 460 enterprises that sent in an SRU but that were not included in the CFAR frame. Of these 1 903 were not included in the CFAR at all, and 86 557 were classified as not being active. The first category consists mainly of newly started enterprises that were not included in the CFAR when the frame was created, but enter into the register at a later date. This category is relatively small, the total turnover for these enterprises amounted to 251 million SEK, the balance sheet total to 16 billion SEK and the number of employees to 607. This means that including them in the survey would affect the total turnover of the corporate sector by 0.007 per cent, the balance sheet total by 0.35 per cent and the number of employees by 0.03 per cent. The enterprises classified as not being active, on the other hand, are a large group, especially when it comes to balance sheet data. This category had a total turnover of 44.2 billion SEK, a balance sheet total of 705.1 billion SEK and 24 085 employees. Including this category in the survey would affect the turnover by 1.23 per cent, the balance sheet total by 15.1 per cent and the number of employees by 1.20 per cent for the total corporate sector.

Even if the number of enterprises above is large, a small number of enterprises account for most of the figures. 634 enterprises with a balance sheet total of 100 million SEK or more accounted for 83 per cent of the balance sheet total above, and 644 enterprises with a turnover of 10 million SEK or more accounted for 68 per cent of the total turnover.

One problem in including these enterprises in the survey is that a large number of them have not received an industrial classification. Of the 88 460 enterprises above, 45 832 had no industrial classification. This problem is noticeable also among the largest enterprises outside the frame. Of 1024 ”large” enterprises outside the frame (enterprises with at least 10 employees, 50 million SEK turnover or 100 million SEK balance sheet total), 380 had no industrial classification.

The problem with enterprises that have no industrial classification is also very large for sole proprietors. The SBS covers sole proprietors within divisions C-O of Nace Rev.1 (10-93 of SNI 92). In 1997, Statistics Sweden received 377 550 SRUs for sole proprietors within these industries. Of these, 183 448 did not have an industrial classification, and thus were excluded from the frame.This means that almost 49 per cent of the SRUs received for sole proprietors did not have an industrial classification. The total turnover of these unclassified SRUs amounted to 23.7 billion SEK. The total turnover of sole proprietors in the SRU amounted to 105.4 billion SEK, which means that including sole proprietors without industrial classification would raise the total turnover for this sector by 22.5 per cent. Compared to the total turnover of the corporate sector, 23.7 billion would only have a small effect on the total turnover of the whole of SBS, but it is an important problem if you want to study sole proprietors by themselves.

Even if the problem of under coverage is the most serious problem with the frame for the SBS, the possibility of surplus coverage may also affect the statistics. In 1997, there were 22 593 enterprises that were included in the frame but that did not send in an SRU at all. Figures for these enterprises had to be estimated, the method used being average figures for the appropriate industry and size class. These estimates amounted to a total turnover of 67 billion SEK, a balance sheet total of 96 billion SEK and 36 256 employees for these 22 593 enterprises, which means that they accounted for 1.9 per cent of the turnover, 2.0 per cent of the balance sheet total and 1.8 per cent of the number of employees for the total corporate sector in the SBS.

The issue of possible under coverage and surplus coverage is a matter to be taken seriously. Especially for balance sheet data, using today´s criteria for including an enterprise in the frame means that a substantial part of the balance sheet total is not included. For other variables, the impact is significantly smaller, but might vary between industries and size classes. Therefore, this matter needs to be studied thoroughly to decide whether to change the frame, and if so how to give all enterprises in the frame an industrial classification.

4. ESTIMATING VARIABLES NOT AVAILABLE IN THE ADMINISTRATIVE DATA

Since administrative data do not normally contain all items important to a survey, it is necessary to develop tools for estimating the items that are not directly available. This might be done in several different ways, e.g. distributing totals from the administrative data using data for large enterprises surveyed via questionnaires as the key for distribution or using questionnaires for a small sub-sample to create such a key. Another method, that has proved to be useful for certain indicators, is estimating data using only different administrative sources. In Sweden, this method has been used for estimating data for investments and changes in equity. Since the methods are similar, this presentation focuses on changes in equity. To simplify matters a little, only joint stock companies are considered in this presentation.

Data on changes in equity are important for the national accounts in their calculations of sector accounts. For these calculations to be as correct as possible, it is necessary to calculate changes in equity for small enterprises surveyed through administrative data as well, data for large enterprises are not enough.

The changes in equity for a certain year can be presented as a very simple formula:

Equity at the beginning of the year + profit/loss for the year + other changes = Equity at the end of the year

What proves more difficult is to divide the item ”other changes” into the items that are of interest to national accounts, namely:

- New share issue, including agio
- Decrease of share capital
- Dividend paid
- Shareholders´ contribution
- Write-ups and write-downs using revaluation reserve

The model uses three kinds of administrative data; SRU for the year, SRU for the previous year and data from The patent and registration office (Patent- och registreringsverket, PRV) about which enterprises are newly started during the year, and which have been deregistered during the year. The calculation is then done as follows:
1. The equity total, both at the beginning of the year and at the end of the year is divided into the following items:

Share capital (SC)
Revaluation reserve (RR)
Other restricted reserves (OR)
Unrestricted reserves (UR)

This can only be done for enterprises that have SRUs that have been considered correct for both years. Enterprises that do not have correct SRUs for both years are treated separately, see point 10 below.

2. The difference between the end of the year and the beginning of the year is calculated for all items, as well as for the equity total (ET), as follows, using balance sheet data for two years of SRU and income statement data about profit/loss for the year (PR) for this year:

$\Delta SC = SC\ (t) - SC\ (t-1)$
$\Delta RR = RR\ (t) - RR\ (t-1)$
$\Delta OR = OR\ (t) - OR\ (t-1)$
$\Delta UR = UR\ (t) - UR\ (t-1) - PR\ (t)$
$\Delta ET = ET\ (t) - ET\ (t-1) - PR\ (t)$

3. The changes in equity are calculated on the basis of the changes in the different items using different model assumptions. The model goes from the simplest possible cases to more and more complicated ones, and finally leave some of the most complicated cases for manual research. The different steps are described below.

4. The simplest case is where the change in equity total is equal to the change in unrestricted reserves i.e.:

$\Delta ET = \Delta UR \Rightarrow$ Shareholders´ contribution (if positive) / dividend paid (if negative)

5. If 4 is not fulfilled, then the model checks if the change in equity total is equal to the change in unrestricted reserves plus the change in other restricted reserves:

$\Delta ET = \Delta UR + \Delta OR \Rightarrow$ If $\Delta ET>0$ then ΔET=shareholders´ contribution, else ΔET=dividend paid

6. If 5 is not fulfilled, then the revaluation reserve is introduced as an explanatory variable, i.e:

$\Delta ET = \Delta UR + \Delta OR + \Delta RR$

Then, ΔRR = write-up (if positive) / write-down (if negative). The rest of ΔET is treated as above, i.e:

If $\Delta ET-\Delta RR>0$, then $\Delta ET-\Delta RR$=shareholders´ contribution, else $\Delta ET-\Delta RR$=dividend paid

7. Up to now, there has been no change in share capital. If points 4-6 are not enough to explain the total change in equity, share capital is introduced as an explanatory variable. The simplest case is where

$\Delta ET = \Delta SC$

Then, if $\Delta ET>0$, ΔET=new share issue, else ΔET=decrease of share capital

8. In case 7 is not fulfilled, the model tests if share capital and one other explanatory variable is enough. In this case, there are several different possibilities that have to be tested:

a. $\Delta ET>0$ and $\Delta ET>\Delta SC$
b. $\Delta ET<0$ and $\Delta ET>\Delta SC$
c. $\Delta ET>0$ and $\Delta ET<SC$
d. $\Delta ET<0$ and $\Delta ET<\Delta SC$
Here, we present only the model for case a above. The other cases are similar, although a little more complicated.

This is a relatively simple case, since both changes have to be positive. Depending on which variable is involved, the model makes different assumptions, as follows:

$\Delta ET = \Delta SC + \Delta OR \Rightarrow$ new share issue including agio. It could also have been a new share issue and a shareholders´ contribution, this estimate is based on the model assumption.

$\Delta ET = \Delta SC + \Delta RR \Rightarrow \Delta SC =$ new share issue, $\Delta RR =$ write-up

$\Delta ET = \Delta SC + \Delta UR \Rightarrow \Delta SC =$ new share issue, $\Delta RR =$ shareholders´ contribution

9. If 8 is not fulfilled, the change in total equity is considered to be too complicated to be modelled adequately. To model it, you would have to take into account three explanatory variables, with possibilities of both positive and negative changes for each variable, which would mean a large number of different cases, each relevant for only a very small number of enterprises. Therefore, enterprises in this category are not calculated, but studied manually. Large changes are entered into the database, whereas small changes are left as unexplained.
10. For enterprises that do not have correct SRUs for both years, a different approach is necessary. These enterprises can be divided into three categories:

a. Enterprises that existed in t-1, but do no longer exist
b. Enterprises that exist in t, but did not exist in t-1
c. Enterprises that existed both years, but where one or both SRUs are unusable

Data on which enterprises belong to categories a and b are collected from PRV. For enterprises that belong to category a, all variables for t are set to zero. Then, the changes in the different variables are treated as follows:

$\Delta SC =$ decrease of share capital, $\Delta RR =$ write-down, $\Delta OR + \Delta UR =$ dividend paid

In a similar way, for enterprises belonging to category b, all variables for t-1 are set to zero, and the changes are calculated as follows:

$\Delta SC + \Delta OR =$ new share issue including agio, $\Delta RR =$ write-up, $\Delta UR =$ shareholders´ contribution

Enterprises belonging to category c are left for estimation.

The model has been used for the SBS material for 1996 and 1997. It was evaluated for the 1996 material. Almost 200 000 enterprises were included in the evaluation, of which 162 000 were joint-stock companies. Of these, almost 133 000 enterprise had no ”other changes” in total equity at all (we accepted difference of 100 SEK). For 25 000 enterprises, changes in equity were calculated and 4 000 enterprises were left for manual studies. Of these less than 100 enterprise had large enough changes in total equity for a manual check to be considered meaningful. For the rest of the enterprises in this group, the total change in equity was considered unexplained.

To evaluate the model, a number of enterprises in each category (points 4-8 above) were chosen for manual checking against annual reports. This checking showed that the model as a whole gave good estimates. Only in a few cases did the model estimate wrong figures, and in most of these cases the reason was errors in the SRU data and not problems with the model assumptions.

Using these kinds of models have proven effective for data on equity changes and investments. By using these kinds of models, there is a risk of creating large model assumption errors in the data. The evaluation of 1996 data showed that these errors were small for this model. On the other hand, other indicators are certainly less suited for these kinds of models and must be estimated using other kinds of models.

5. CONCLUDING REMARKS

As seen from the previous discussion, some notable gains have been achieved from the use of SRUs: It has been possible to reduce the response burden towards Statistics Sweden for smaller businesses very considerably. The survey costs have been reduced as a result of this. The SRUs can be useful as a new data source for quality checks and improvement of the Business Register. Precision of estimates for small and medium sized businesses have been very much improved.

There are, however, also some problems with the new design. Even though the SRU material is a very rich data source it does not cover all items of interest. We have seen an example of this in section 4, where it has been necessary to rely on modelling. To control whatever model assumption errors that may arise from this, it may become necessary to perform some evaluation survey, possibly on a small scale and on a non-recurrent basis , to estimate and control for such error.

The SRU material is certainly not error free when it arrives at Statistics Sweden. One must keep in mind when using administrative material that in general the primary user (in this case the National Tax Board) has no particular interest in the material as a statistical data source. Hence only those parts of the material that are important for the primary use (the collection of taxes) can be expected to have first class quality. As a consequence, the SRU material is subjected to a fairly detailed editing procedure at Statistics Sweden, which adds to the survey costs.

USE OF BUSINESS INCOME TAX DATA TO EXTEND THE INFORMATION AVAILABLE FROM THE ABS ECONOMY WIDE ECONOMIC ACTIVITY SURVEY

Steve Crabb & Paul Sutcliffe, Australian Bureau of Statistics
Steve Crabb, Australian Bureau of Statistics, P.O. Box 10 Belconnen ACT 2616, Australia
steve.crabb@abs.gov.au

The views expressed in this paper are those of the authors and do not necessarily reflect those of the Australian Bureau of Statistics (ABS).

ABSTRACT

The Australian Bureau of Statistics (ABS) is committed to increasing the range and quality of the statistics it provides to users, while at the same time keeping the statistical burden placed on businesses to acceptable levels. One important initiative in this regard is the supplementation of the relatively small scale annual Economic Activity Survey conducted by the ABS with business income tax data provided by the Australian Taxation Office (ATO). The statistics produced from this survey have been improved by the use of business income tax data for a large (super) sample of small and medium employing businesses, and by extending the coverage of the survey to include non-employing businesses. As well as improving the quality of broad national estimates, use of a large sample of ATO data in combination with ABS data enables the release of a much finer dissection of business input costs, essential for an input-output approach to the compilation of national accounts.

The methodology used to achieve these improvements is relatively simple in concept, but operationally complex. It involves the use of ABS collected data for the relatively few large and complex businesses, supplemented with ATO collected data for the many small and medium sized, simply structured businesses. A subsample of these small and medium sized businesses is also approached by the ABS to obtain data not available from the ATO and to be used in preliminary estimates. The paper presents details of the methodology used to compile the statistics, together with some of the hurdles that needed to be resolved and some of the challenges that still need to be overcome.

Key Words: respondent burden, taxation data, two-phase estimation

1. INTRODUCTION

In the last decade the Australian Bureau of Statistics (ABS), like many other official statistical agencies, has been under increasing pressure to increase the range of statistics it makes available; maintain and improve the quality of these statistics; and provide them in timely fashion, while becoming more cost effective. In particular, in the last few years the need to reduce the statistical reporting burden placed on businesses has led to more effective uses of administrative data, especially business income tax data provided by the Australian Taxation Office (ATO). The Economic Activity Survey (EAS) which commenced in the early 1990s has recently been modified to use taxation data to supplement data directly collected data from businesses.

The ABS has been using ATO data to compile economic statistics, in one form or another, for many years. The use it has made of these data has varied as a consequence of such things as changes to international statistical standards, changes to Australian economic statistics strategy, and changes to Australian taxation law and administration. Basic information about the Australian business population is used to maintain and update the ABS Business Register, which is used as a framework for most ABS business collections. In recent years, this register has increasingly relied on data from the ATO's Group Employer system. This will soon be replaced by data available from the new Australian Business Register, to be maintained by the ATO.

Aggregate data from the income tax system has been used by the national accounts area of the ABS for many years as a source for annual benchmarks for items such as turnover and gross operating surplus. However, use of these simple aggregates suffered from many deficiencies including consolidation of data for large businesses, inaccurate industry classifications, timeliness/completeness of the dataset and differences in the content and definition of data items. Since the mid-1980s the ABS has been given access to identifiable income tax unit records and has had some significant influence on the data item content of income tax returns. These two events, together with the application of appropriate methodologies to address other deficiencies in tax data, have led to a significant improvement in the quality of the data available for benchmarking the national accounts. While the ABS has a relatively long history of using tax data to maintain its Business Register, it is only in recent years that it has been using unit record income tax data as an alternative to financial data directly collected in its surveys.

2. METHODOLOGY

Currently EAS is designed to provide two main types of data outputs. Firstly, a single phase survey to collect financial data (base data) such as main income and expenditure items, which are readily available from the annual accounts of each business. Secondly, a smaller second phase subsample collecting detailed income and expenditure data required for the Input-Output (I-O) Survey which is used in the compilation of the national accounts. Traditionally these data have been collected directly from businesses.

Over the past few years investigations have analyzed the quality of the tax data reported by small and medium businesses by comparing it with the data currently collected by the ABS. In broad terms these investigations have suggested that data are comparable. Where there are significant differences it is due to the different timing requirements for completing returns. Annual ABS industry surveys generally obtain information from businesses six months prior to the finalisation of accounts for income tax purposes. While the output requirements from EAS have not diminished, the availability of taxation data has opened up a range of possibilities for reducing the amount of direct collection and improving the quality (sampling and non-sampling error) of the outputs.

Incorporation of taxation data into the design for EAS is achieved through the partitioning of the economy wide population framework into sub-populations from which either direct collection or administrative data, or both can be used. The larger and more complicated businesses are only surveyed using direct collection. However for some classes of businesses, the directly collected EAS data (used in the preliminary publication) and the taxation data available later are closely comparable at the unit level. For these businesses a two-phase approach allows the use of a ratio estimator to 'pro-rate' the benchmark totals collected in a first phase sample from tax by the finer breakdown of data collected from EAS.

2.1 Population Framework

At present the ABS is not using the taxation data as a population framework in its own right, instead preferring to use it as an updating source for our own register of businesses. All sampling frames for business surveys are extracted from this register maintaining scope and consistency across our surveys. At present the ABS register does not include non-employing businesses. However, the availability of data about the non-employing businesses from tax has enabled the ABS to expand the scope of the published estimates for business surveys, even though the businesses are not added to the register population.

Although EAS is an economy wide survey the tax strategy outlined in this paper refers to only to the service industries component. Other surveys such as Manufacturing, Mining and Agriculture are developing their own strategies for the use of taxation data.

Studies have demonstrated that the best use of taxation data is restricted to business with a simple structure defined by a single legal entity, the majority being single location businesses. Therefore, in order to effectively operationalise the use of the two sources of data the population has been divided into a number of streams. They are: (1) large businesses which have complex structure; (2) complex small and medium employing businesses; (3) simple small and medium businesses suitable for income tax data substitution; and (4) non-employing businesses which are only available from the income tax files (i.e. currently out of scope of the ABS Business Register). Table 1 below shows the relationship of the four streams.

Table 1: Summary of Data Sources for the EAS/Tax Strategy

Selected industries

Type of business	Large businesses Stream 1	Complex small & medium employing businesses Stream 2	Simple small & medium employing businesses Stream 3	Non-employing businesses Stream 4
The source used to identify businesses of each type	ABS Business Register	ABS Business Register	ABS Business Register	ATO income tax files
The number of each type of businesses	2 000	102 000	497 000	952 000
The number of businesses that are selected to provide data	2 000	3 000	88 000	952 000
Source of data	ABS survey	ABS survey	ATO business income tax returns	ATO business income tax returns
Contribution to total business income for service industries	32%	14%	48%	6%

The majority of large employing businesses (businesses with 200+ employees) have more than one legal entity, making it difficult to identify all legal entities for that business on the tax file. Additionally, business income tax data do not include all of the detailed information that the ABS requires from large businesses. Finally while large businesses with more than 200 employees are relatively few in number, their contribution to the estimates is relatively large. Hence the business financial data for these units is sourced from the EAS direct collection.

Similarly, it is not appropriate to source business financial data from ATO business income tax returns for all small and medium employing businesses. Some small and medium businesses have more than one legal entity while not-for-profit businesses are generally not required to complete businesses income tax returns. These types of businesses are collectively termed 'Complex small and medium employing businesses' and are both sourced from the EAS direct collection.

Small and medium employing businesses that have simple structures (i.e. Businesses with one legal entity) are better suited to having their data sourced from their business income tax returns. There were approximately 497,000 simple small and medium employing businesses on the ABS Business Register as at 30 June 1997-98. Of these, 88,000 were selected to have their data sourced from their business income tax returns. A sample is selected to give adequate attention to the detection and treatment of non-sampling errors and to keep the processing costs down to an acceptable level. The selected businesses have to be identified on the business income tax files to obtain their business financial data. To successfully achieve this requires the identification of the ATO Reference Number for each selected business. The ABS and the ATO have been collectively working towards having the ATO Reference Number for each employing business on the ABS Business Register.

The scope of the ABS Business Register is all employing businesses. Even though taxation data for non-employing businesses has become available the records have not been added to the sampling framework. This doesn’t restrict the scope of the outputs as an aggregate contribution for non-employing businesses is included in estimation. The inclusion of non-employers as stream 4 is made by producing an aggregate estimate of the contribution using all the taxation data available for these businesses. As the businesses contributing to the estimate of stream 4 are not reconciled with the units contributing to the estimates in stream 3 there is potential for duplication. A recent study has measured the over estimation attributed to duplication at around 3% of the final estimate.

2.2 Output Requirements

When EAS commenced in the early 1990s the main output requirement was the provision of a set of standard economy wide financial data (base data) available from annual accounts of businesses; for example, data collected such as wages and salaries, sales, total expenses, total income, other expenses and other income paper in the annual profit and loss statements required for all businesses. Once the Input-Output methodology became the method by which the national accounts are compiled, the need to collect more detailed income and expenditure data was required from ABS business surveys. The approach taken was to select a smaller sub-sample of the businesses selected in EAS who were asked to also complete the more detailed I-O form. The benchmark data from the larger EAS sample is then be apportioned into the detailed I-O output.

To ensure that data are available in a timely fashion for the compilation of national accounts, a preliminary publication is produced based on data from a sample (n=5,000) collected directly from businesses. While the data items are the same in the preliminary publication the level at which the data is presented (e.g. industry level) is much broader than is published in the final publication. This sample is designed to meet a set of constraints for the first publication only, but is also used in the second phase calculation of the base data.

The advent of the use of taxation data as a large set of business records has changed the methodology used by EAS. In effect this has meant that EAS can decide how and when to utilise the taxation data. For example, since the taxation data is available too late for preliminary publication purposes, directly collected data from a single phase sample is still used at that stage. The final publication, however, is now augmented using taxation data.

For the final EAS publication a large sample (88,000) is selected from the ABS Business Register, and then attempts to match this to taxation records. Because the data items available from tax are limited, a much smaller ‘second phase’ sample is directly approached so that the full set of base financial data can be published. In practice this ‘second phase’ sample is the same as that used for the preliminary publication, meaning that there are two quite distinct sets of objectives for the design. To date, the estimates for the single phase preliminary publication have been given priority. One interesting feature of this process is that the directly collected second phase sample is available earlier than the data for the first phase sample of tax records.

2.3 Estimating Base Financial Data For The Final Publication

Users of the tax estimates examine industry level estimates as a whole, not separately by streams, hence the sample was allocated to minimise the overall sample error for the entire tax estimate. Since the use of taxation data is restricted at present to the compilation of the base data and in particular the estimates from stream 3, the rest of this section will concentrate on how the stream 3 estimates are produced.

As first phase data in stream 3 is obtained as an administrative by-product, there is no provider load cost from the ABS due to the survey. The processing costs to the ABS are also small compared to surveys where the ABS collects the data, since the ATO has already incurred many of these collection costs. Therefore, for the limited set of data available from Tax there is really no restriction on the size of the sample in the first phase. A census of stream 3 was considered but was rejected for a number of reasons. Firstly, the population, as previously mentioned, is selected from the ABS register of businesses and it is known that not all data would be available for all businesses from tax when required for estimation. Secondly, a sample means less units need to be checked for non-sample error when editing reported data, allowing for a greater focus on the methods of editing. Thirdly, the final estimate produced by the combined streams using the two collection methods will always contain a source of sampling error from the sample of stream 2 units. Finally, the sampling errors for stream 3 can be minimised efficiently using a large first phase sample.

The current sample (*n*= 88,000) from the population of 497,000 businesses in the first phase is designed to achieve lower relative standard errors (3%) at the ANZSIC subdivision level than the existing directly collected EAS (7%) which is used for the preliminary publication.

The use of tax data is not without its problems. In our case they surface when estimating the benchmark total for the first phase x'_{Tax}. Once the first phase sample has been selected from the ABS Business Register the strategy then requires the selected sample to be matched against the appropriate taxation file. In an effort to produce timely data, the ABS uses business income tax files that are generated 12 months after the end of the financial year. At this point in time not all businesses have provided their income tax returns by that date (indications are that income tax returns for 10% of businesses are still be outstanding). Additionally, a proportion of businesses that are included on the ABS Business Register would not have traded for the year in question. After taking these factors into account, it is estimated that data for between 15% and 20% of live businesses selected in the sample are not available for processing when estimates were produced. Conversely, 80% to 85% (match rate) will match with businesses on the ABS Business Register. Data for these unmatched businesses must be estimated.

In practice studies have shown that the first phase selections which match to the taxation data have a higher proportion of live businesses than the non-matched units. Thus the non-matched units cannot be considered to be ‘standard non-respondent’ units and should not be estimated using the average values of the respondent units. To account for this ‘response bias’ the estimates are adjusted for the lower proportions of non-matched units that are live using an adjustment factor.

Within each first phase stratum *h* the estimate can be decomposed into two simple expansion estimates, one based on those units where a match $(i \in hm)$ is made between the taxation data and the first phase tax sample, and the other based on those which do not match $(i \in h\bar{m})$, where *m* and $\bar{m}$ denote matched and unmatched units respectively.

$$x'_{Tax} = \sum_{h=1}^{H} \frac{N_h}{n_h} \left[\sum_{i \in hm} x_{hi} + \sum_{i \in h\bar{m}} x_{hi} \right]$$

The first component can be directly calculated from the matched sample. However, the second component requires estimation and can be rearranged as shown below.

$$x'_{Tax} = \sum_{h=1}^{H} \frac{N_h}{n_h} \left[\sum_{i \in hm} x_{hi} + n_{h\bar{m}} \bar{x}_{h\bar{m}} \right]$$

We can use the relationship $N_{h\bar{m}} \bar{X}_{h\bar{m}} = N_{h\bar{m}l} X_{hml} + N_{h\bar{m}d} X_{h\bar{m}d}$ where *l* and *d* distinguish between live and defunct businesses. By noting that $N_{h\bar{m}d} X_{h\bar{m}d} = 0$ because $\bar{X}_{h\bar{m}d} = 0$, and estimating the population means using the sample means provided, the relationship $N_{h\bar{m}} \bar{x}_{h\bar{m}} = N_{h\bar{m}l} \bar{x}_{hml}$ can be used to obtain the following.

$$x'_{Tax} = \sum_{h=1}^{H} \frac{N_h}{n_h} \left[\sum_{i \in hm} x_{hi} + n_{h\bar{m}} \frac{N_{h\bar{m}l}}{N_{h\bar{m}}} \bar{x}_{h\bar{m}l} \right]$$

By assuming that the mean value of the live unmatched units is equal to the mean value of the live matched units, i.e. $\bar{x}_{h\bar{m}l} = \bar{x}_{hml}$, the second component can be expressed in terms of the matched units only.

$$x'_{Tax} = \sum_{h=1}^{H} \frac{N_h}{n_h} \left[\sum_{i \in hm} x_{hi} + n_{h\bar{m}} \frac{N_{h\bar{m}l}}{N_{h\bar{m}}} \frac{N_{hm}}{N_{hml}} \bar{x}_{hm} \right]$$

This equation can be simplified into a term including the ratio of unmatched to matched units in the sample, a factor f_{sm} incorporating the ratio of live populations for matched and unmatched units and the sum of all matched units in sample.

$$x'_{Tax} = \sum_{h=1}^{H} \frac{N_h}{n_h} \left[\sum_{i \in hm} x_{hi} + \frac{n_{h\bar{m}}}{n_{hm}} f_{sm} \sum_{i \in hm} x_{hi} \right]$$

This then simplifies to an adjusted simple expansion estimator.

$$x'_{Tax} = \sum_{h=1}^{H} \frac{N_h}{n_h} \left[1 + f_{sm} \frac{n_{h\bar{m}}}{n_{hm}} \right] \sum_{i \in hm} x_{hi}$$

$f_{sm} \dfrac{n_{h\bar{m}}}{n_{hm}}$ is known as the adjustment factor, where *s* represents the level at which the factor is calculated. Note that when there is full matching between the selected sample from the Business Register and the records from tax this factor is zero and the estimator reduces to the standard simple expansion estimator.

However, the population counts in f_{sm} are not known and need to be estimated. The components N_{hm} and $N_{h\bar{m}}$ are estimated using information from the tax first phase sample where a unit's matched/unmatched status is determined and the components N_{hml} and $N_{h\bar{m}l}$ are estimated using information from both the EAS survey where a unit's live/dead status is determined and the first phase sample of tax records. In practice this is done using:

$$f_{sm} = \left(\sum_{h \in s} \frac{N_h}{n_h} \hat{n}_{h\bar{m}l} \Big/ \sum_{h \in s} \frac{N_h}{n_h} n_{h\bar{m}} \right) \left(\sum_{h \in s} \frac{N_h}{n_h} n_{hm} \Big/ \sum_{h \in s} \frac{N_h}{n_h} \hat{n}_{hml} \right)$$

where $\hat{n}_{h\bar{m}l} = \left(\dfrac{m_{g\bar{m}l}}{m_{g\bar{m}}} \right) n_{h\bar{m}}$ and $\hat{n}_{hml} = \left(\dfrac{m_{gml}}{m_{gm}} \right) n_{hm}$ are based on the proportion of live businesses (*m*) in EAS second phase stratum *g*.

The large first phase sample provides accurate benchmark data while the second phase sample collects detailed breakdowns that can be used to pro-rate the benchmarks for output purposes. The pro-rating factors refer to the different benchmark variables which are used for each type of output.

Pro-rated estimates are currently calculated at the ANZSIC subdivision level. The EAS based pro-rating factor, $\dfrac{y'_{EAS}}{x'_{EAS}}$, and the total from Tax, x'_{Tax}, are all calculated at the ANZSIC subdivision level, and then the final pro-rated estimate, $y'_{pro-rated}$, is calculated at this level.

For example, in stream 3 the first phase estimates of total income is calculated using the data from taxation records, but estimates of all dissections, or components of total income, are estimated using pro-rating factors. Total income is also collected in EAS, and the factors estimated from EAS are the proportion each component contributes to the total. Thus the pro-rated estimates are $y'_{pro-rated} = \dfrac{y'_{EAS}}{x'_{EAS}} x'_{Tax}$ which is the standard form of the two-phase ratio estimator where each term is defined as follows:

$y'_{pro-rated}$ = final pro-rated estimate (e.g. estimate of interest income)

y'_{EAS} = estimate from EAS of finer level item (e.g. interest income)

x'_{EAS} = estimate from EAS of total (e.g. total income)

x'_{Tax} = estimate from Tax of total (e.g. total income).

While the estimator outlined above is in the standard form of a two-phase ratio estimator, it should be noted that the benchmark information collected from tax collected in the first phase and from EAS in the second phase come from different sources. Therefore, at the businesses level, reported values may not be equal for the same base data item and so the population totals are not equal $\left(X_{Tax} \neq X_{EAS}\right)$ as would be the case in a traditional two-phase survey.

The variance estimator for the two-phase phase estimator requires modification to account for the variance associated with the matching of taxation data in the first phase. However, the current variance estimation component for the first phase ignores this random nature of the adjustment factor and treats f_{sm} as a constant:

$$var'\left(x'_{Tax}\right) = \sum_{h=1}^{H}\left(\frac{N_h}{n_h}\right) n_{hm}\left(N_h - n_{mh}\right)\left[1 - f_{sm}\left(\frac{n_{h\bar{m}}}{n_{hm}}\right)^2\right] S_h^2$$

where

$$S_h^2 = \frac{1}{\left(n_{hm} - 1\right)}\left[\sum x_i^2 - \frac{\left(\sum x_i\right)^2}{n_{hm}}\right]$$

Similar formula are used for $var'\left(x'_{EAS}\right)$ and $var'\left(y'_{EAS}\right)$.

The overall variance estimator is linearised using Taylors Series expansion.

$$\begin{aligned}
var'\left(y'_{pro-rated}\right) &= var'\left(\frac{y'_{EAS}}{x'_{EAS}} x'_{Tax}\right) \\
&\cong \left(\frac{x'_{Tax}}{x'_{EAS}}\right)^2 var'\left(y'_{EAS}\right) + \left(\frac{y'_{EAS}}{x'_{EAS}}\right)^2\left(\frac{x'_{Tax}}{x'_{EAS}}\right)^2 var'\left(x'_{EAS}\right) + \left(\frac{y'_{EAS}}{x'_{EAS}}\right)^2 var'\left(x'_{Tax}\right) \\
&\quad - 2\left(\frac{y'_{EAS}}{x'_{EAS}}\right)\left(\frac{x'_{Tax}}{x'_{EAS}}\right)^2 cov'\left(y'_{EAS}, x'_{EAS}\right) + 2\left(\frac{y'_{EAS}}{x'_{EAS}}\right)\left(\frac{x'_{Tax}}{x'_{EAS}}\right) cov'\left(y'_{EAS}, x'_{Tax}\right) \\
&\quad - 2\left(\frac{y'_{EAS}}{x'_{EAS}}\right)^2\left(\frac{x'_{Tax}}{x'_{EAS}}\right) cov'\left(y'_{EAS}, x'_{Tax}\right)
\end{aligned}$$

A recent study concluded that the correct variance estimation which accounts for the adjustment factor shows an increase in the estimated variance for stream 3 which results in the overall relative standard errors (RSE) increasing from 3% to 3.9%. These changes in the RSEs are significant enough to consider implementing the correct variance formula into the estimation system. Therefore, we are currently investigating the use of the jackknife variance estimation technique which is used by the generalised estimation system (GENEST) developed by the ABS.

2.4 Estimating Detailed Input-Output Data

The method used to select the I-O sub-sample is to take the EAS selections, restratify them into I-O strata, and select the required second phase sample. The I-O selections are then a simple random sample from the EAS selections within each I-O stratum. This results in units within the same I-O stratum having different selection probabilities depending on which EAS stratum they were selected. The incorporation of the taxation data has effectively added a third phase to the estimation process. The availability of more accurate benchmark data from the first phase sample of tax records allows for more options when determining what level to apply pro-rating. Storm (1997) modified the standard pro-rating technique in a way that overcame the following problems:

- Providers often reported a value of total other expenses on the I-O supplementary form different to what they reported on the main EAS form, and this was not properly edited, or checked for consistency.
- Unequal probabilities at the third phase of collection within I-O strata makes estimation complicated, meaning pro-rating was easiest at the second phase level.
- The I-O sample size was quite small, and consequently there were many units with large influence (weights) on the estimates at stratum level.

2.5 Future Improvements to Methodology

As we have seen, the EAS/Tax strategy accommodates both a preliminary sample for national accounts purposes and the need to support detailed Input-Output requirements. Therefore, in the short term it is unlikely that the direct sample could be reduced from 5,000 businesses. However, the current design of the second phase is based on standard outputs from the old EAS and does not reflect the purpose of two phase pro-rating using tax records. Thus, we are currently investigating designing the second phase based on constraints relating to the accuracy of the pro-rating factors rather that estimate of level.

To move to a standard two-phase design approach we are investigating the incorporation of a cost function relating the cost of collecting data for each phase into the overall design, in order to obtain a more optimal balance between the size of first and second phase samples.

3. OPERATIONALISING THE USE OF INCOME TAX DATA

Operationalising the methodology has not been straight forward and plenty of effort has been required to ensure that the first set of published ‘experimental’ estimates were of high quality. This section will discuss some of the hurdles faced throughout this difficult process, and the arrangements put in place which work to minimise the difficulties.

A key to the quality of estimates has been the close relationship maintained between the ABS and ATO. Apart from very regular contact with many officers at many levels in both organisations, a number of more formal processes have been established. A Memorandum of Understanding sets out an agreement between the two organisations for mutual co-operation in the use of tax data and in other activities of joint interest. The memorandum sets a number of milestones which the two organisations work towards and which are updated annually. The relationship is further aided by the existence of an ABS officer outposted to the ATO whose role is to coordinate statistical activities and communication. At the highest level a small committee of senior executives from the ATO and ABS meet about three times a year. This group focuses on the strategic aspects of the relationship and monitors overall progress on issues of mutual interest.

While extensive editing and quality control is undertaken by the ATO before passing data to the ABS, this is largely driven by the requirement for tax compliance. The ABS also undertakes a significant amount of editing on the data for statistical purposes, often focusing on the economic relationships between finer level income and expense items on the forms. The results of this work forms the core input into annual quality reports which the ABS provides by way of feedback to the ATO. Information from these reports can lead to amendments to ATO quality control processes and on a number of occasions to the design of the tax forms themselves.

In using administrative tax data the ABS has needed to address issues of timing versus reliability of the information. As a consequence of the time allowed for business to submit taxation returns and ATO processing and quality control, a final business income tax tape is not available to the ABS until 18 months after the end of the reference period. A preliminary tape is available 9 months after the end of the reference year, but would contain only around 60% of records. To derive estimates which are reliable, yet as timely as possible, the ABS methodology currently utilises a tape provided 12 months after the reference period containing approximately 90% of records as the basis for estimates. As noted in Section 2, use of the ABS register as the population frame provides a basis on which to estimate for missing records. Use of tax data 12 months after the end of the reference year nevertheless delays publication of data by six months necessitating timely release of broad preliminary aggregates which do not utilise tax data.

The large numbers of records on the tax files, over two million, has influenced the choice of methodology as described in Section 2. Primarily this is reflected in the decision to match a sample of records on the register to the tax files rather than the entire population. As only a sample of records are matched to the tax files, it is not possible to identify all those businesses that are not on the register. The ABS Business Register is primarily sourced from ATO Group Employer registrations, therefore those not on the register for the most part will be non-employers. The ABS uses a set of defining characteristics (i.e. based on reported values for wages and salaries, employee superannuation expenses and size of reported income and expenses) to identify non-employing businesses on the business income tax files. It is recognised that the current set of defining characteristics leaves open the possibility for some businesses to be included both in the population of employing businesses as well as the population of non-employing businesses. Additionally, other businesses could be excluded from both of these populations. The potential overlap, estimated to be approximately 3% of the final estimate, will be addressed in future improvements to methodology.

Use of administrative data in the manner described in this paper has resulted in cultural shift for processing staff accustomed to direct collection methods and regular contact with data providers. Data quality issues, including resolution of edit failures, are not usually able to be followed through with the tax return provider, resulting in more reliance on alternative investigation methods. The result has been an improved understanding of unit record structures and of the conceptual differences underlying tax and economic statistics.

Estimates derived using tax data continue to carry an 'experimental' label in ABS publications. The ABS is working to build confidence in the quality of the estimates through a range of measures, including documentation of concepts and methodology, regular reference group meetings to discuss the use of tax data and regular quality reviews. An important component of the quality reviews has been visits to accountants and businesses to discuss factors leading to differences between tax data and data reported to the ABS. A common theme which has emerged from the visits is that while the timeliness of taxation data delays publication by six months the quality of the data is improved by being based on accounts which are finalised for tax purposes.

4. CHALLENGES FOR THE FUTURE

Supplementation of the Economic Activity Survey with business income tax data has enabled the ABS to improve the range and quality of statistical output produced while at the same time minimising the reporting load on providers. The main challenges for the future are to utilise tax data in satisfying other areas of unmet demand such as finer industry statistics and geographic breakdowns and to promote the use of business income tax data by other collection areas in the ABS.

While the methodology described in this paper has been successful in improving the quality of industry data available regarding small and medium sized businesses, user demand remains for data at finer levels of industry classification, and in particular 4 digit ANZSIC Class level data. The current methodology employed limits the availability of finer level industry output due to the sample design in the small and medium employing business streams (Streams 2 and 3 in Table 1). Future work proposed by the ABS will target those industries for which reliable Class level results could be obtained using tax data. For instance, for those industries dominated by simply structured small and medium employing business or non-employing businesses, an increase in the tax sample of small and medium business may result in reliable Class level estimates.

There is also strong demand amongst users of economic statistics in Australia for breakdowns by State and Territory and even finer regional breakdowns. Taxation data is a particularly useful source of information for the many small businesses with a single location. For example, a project currently underway in the ABS focuses on using tax data for small business to derive a regional view of the data. The definition of small business used was total income less than $5m. It was intended that such a definition would largely eliminate multiple-location entities or headquarter-type businesses, which may have operations outside of the region in question.

While efforts are made to ensure general comparability of estimates derived using tax data, the application of specialist methodologies to derive reliable fine level industry estimates and geographic breakdowns may result in a degree of inconsistency with EAS estimates. The ABS is currently preparing a Concepts, Source and Methods document which will describe the individual methodologies and provide some indication of the relative quality of estimates.

Work continues in the ABS to promote the use of tax data by other collection areas in the ABS. In addition to using tax data as a supplement to direct collection as was the case with the EAS, a number of other uses are encouraged. These include using taxation data as a substitute for direct collection, for the purposes of frame checking or supplementation, as input into an efficient sample design, as input into editing, imputation and outlier analysis and for the purpose of post enumeration studies, data confrontation and quality assurance. To assist in the process, the ABS has conviened the Business Income Tax Reference Group which provides a forum for current and potential users of business income tax data within the ABS to share information.

A significant influence on the future use of business income tax data in the ABS will be the current process of tax reform in Australia. Tax reform covers many changes to the tax system including the introduction of an Australian Business Register (ABR), the introduction of a Goods and Services Tax, a review of business taxation and introduction of a new Pay as You Go system for the payment of income and other taxes. The reform process provides many statistical opportunities including the availability of regular Business Activity Statements (BAS). The BAS is a new government requirement for businesses to claim credits from the new tax system. The statements contain information about turnover, wages paid, capital expenditure, exports and other expenses and is available on a quarterly basis. The BAS, while not containing the detail of a business income tax form, will be available on a far more timely basis and should provide a rich source of information with which to supplement ABS direct collection surveys in the future.

5. REFERENCES

"Business Operations and Industry Performance 1996-97" ABS Publication 8140.0

"ABS - Use of Business Income Tax Data in ABS Economic Surveys" ABS Publication 5672.0

Storm, A. (1997), "Estimation for the 1996 Input-Output Subsample", unpublished report, Canberra, Australian Bureau of Statistics.

NEW APPLICATIONS OF OLD WEIGHTING TECHNIQUES

Constructing a Consistent Set of Estimates Based on Data from Different Sources

A.H. Kroese and R.H. Renssen, Statistics Netherlands[1]
A.H. Kroese, Statistics Netherlands, P.O. Box 4000, 2270 JM Voorburg, the Netherlands
akse@cbs.nl

ABSTRACT

The increasing demand for consistent data and the need to work more efficiently make it necessary to redesign the statistical process in our institute. The use of administrative registers is central to the new approach. The outline of the new statistical process has already been developed and described before.

In this paper we discuss the estimation strategy to be applied in the new process. To that purpose, the method of mass imputation is evaluated. After that a recently developed estimation approach is presented as a more satisfactory alternative. The method uses (exact) record linking and (repeated) weighting in order to obtain an extended set of approximately design unbiased estimates that are totally consistent with each other as well as with all possible aggregations of the register variables involved. The approach combines the attractive statistical properties of model assisted estimates with total coherency of all estimates, both from the surveys and the registers involved.

In the paper we focus on business statistics. Various examples are given.

Key Words: calibration estimation, mass imputation, micro-database, repeated weighting

1. INTRODUCTION

At Statistics Netherlands the design and organization of the statistical process is changing rapidly. This change is motivated by the need to produce more consistent data and by political pressures to cut down staff costs and the response burden.

In Keller et al. (1999) an outline is given of the statistical process in the near future. This outline is briefly discussed in section 2. A number of methodological problems have to be solved in order to implement the new statistical process. In particular, an appropriate estimation technique has to be found. One possibility to generate estimates is mass imputation. In general, this technique may lead to unreliable estimates, see e.g. Kroese and Renssen (1999). In section 3 the specific situation of business statistics is considered.

In section 4 a weighting approach to estimation is described. The approach can be seen as a new application of old weighting techniques. Several rectangular micro-datasets are constructed, each of which is used to obtain a specific set of estimates by means of weighting. Here we allow the same rectangular micro-dataset to be weighted several times.

2. THE NEW STATISTICAL PROCESS

2.1. Drawbacks of the old approach

Traditionally, the production process of a statistical publication is designed according to the ‘stovepipe’ model. A questionnaire is designed with questions aimed at collecting the information needed to compile a particular publication serving specific user needs. The questionnaire is sent to the units (mostly businesses, households or persons) in the sample and the response is edited and subsequently entered in some kind of statistical package. Publication totals are estimated by calculating one set of weights for the responding units.

In the past, there were many isolated stovepipes in our institute so that harmonization was a very difficult task. It is not guaranteed that the estimates resulting from the various stovepipes can be related without some further action. It is even quite possible that different estimates are published for the same concept in different statistical publications. The need for more consistency has been the main motivation for a change in the statistical process.

The change is also motivated by cuts on the budget of Statistics Netherlands and the political desire to lower the response burden on especially businesses. The new approach has to be more efficient than the stovepipe approach.

[1] The views expressed in this paper are those of the authors and do not necessarily reflect the policies of Statistics Netherlands.

2.2. The new statistical process

In this subsection a short description is given of the new statistical process as envisaged in our institute for the near future. In Keller et al. (1999) a more detailed description can be found.

In figure 1 a diagram of the new statistical process is given.

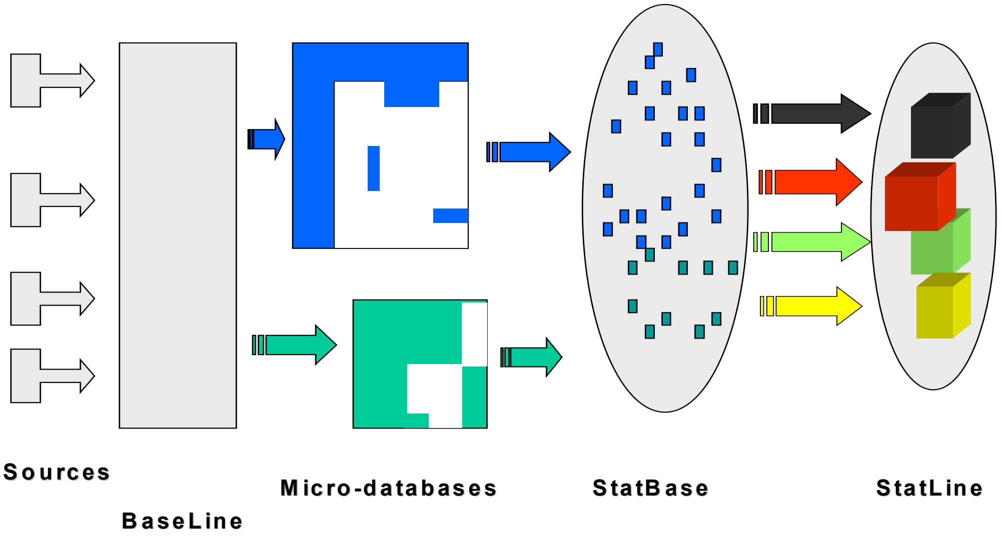

Figure 1: a diagram of the new statistical process

Data from different sources (administrative registers, surveys, EDI) are matched in the input-database '*BaseLine*'. The data in the sources refer to observation units which are not necessarily equal to statistical units. In BaseLine all input-data are adapted to data about statistical units.

As a second step, a *micro-database* is constructed for each object-type (persons, households, businesses etc.). The micro-database for businesses can be seen as a table with a row for each business in the population (the business register). Each column shows a variable. The information in the micro-database is edited and the variables are harmonized. The values of some variables are known for all businesses in the population (SIC-code, size class); the values of other variables are only known for a subset of the population, for example, corresponding to a sample survey. Consequently, the micro-database contains many empty cells.

The next step in the new statistical process is to estimate population totals based on the data in the micro-database. If the value of a variable is known for each business, the population total is estimated by simply adding all values. If the value of a variable is not known for each business a statistical estimation technique has to be applied. All estimates are placed in *StatBase*, which can be regarded as a database containing all data Statistics Netherlands considers worthwhile publishing. The set of estimates in StatBase should satisfy the requirements of

- Reliability: the estimates should either be generated by means of an approximately design-unbiased method (variances reasonably small) or by a model-based method where the model used is plausible in some sense.
- Consistency: by confronting estimates in StatBase no contradictions may be obtained; for example, it should not be possible to derive two different total 'number of employees' in a certain industry by combining estimates.
- Disclosure control: it should not be possible to derive information about individual businesses by combining estimates.

The estimates are disseminated by means of *StatLine*, which can be seen as a 'view' on StatBase. In StatLine the aggregates are presented in a user-oriented way, i.e. the aggregates are arranged in (multi-dimensional) tables (called data-cubes) that reflect standard 'areas of interest' or 'themes' in the society. For example, it is our ultimate goal to gather all data referring to the theme 'health' in one such data-cube, from which users can select the data they are interested in.

2.3. The estimation problem

Before the new statistical process can be implemented in practice, a number of methodological problems have to be solved. One of the main methodological challenges is how to proceed from micro-databases to StatBase. An estimation technique has to be developed that starts from a number of only partly filled micro-databases and that results in a set of reliable estimates that are both consistent, safe and as comprehensive as needed to satisfy user wishes.

Part of this question is the core topic of this paper. Not much attention will be paid here to the confidentiality condition. This means that we will focus on the question how to ensure that the conditions of reliability and consistency are met. To that purpose, we suggest and discuss a new estimation strategy, which can be seen as a new application of old weighting methods. It leads to a larger set of consistent estimates than ordinary weighting.

Before presenting this weighting method, the method of mass imputation is discussed in the specific situation of business statistics. Mass imputation can be considered as an attractive alternative for weighting, because it induces consistent estimates in a straightforward way. However, if the imputations are less accurate, then, generally, cross-tabulations should be considered unreliable as imputations can have harmful effects on relationships.

Both the weighting method and the mass imputation method will be illustrated by means of a very simple fictitious micro-database presented in figure 2. The backbone of this micro-database is the business register with information about SIC-code, number of employees and legal status.

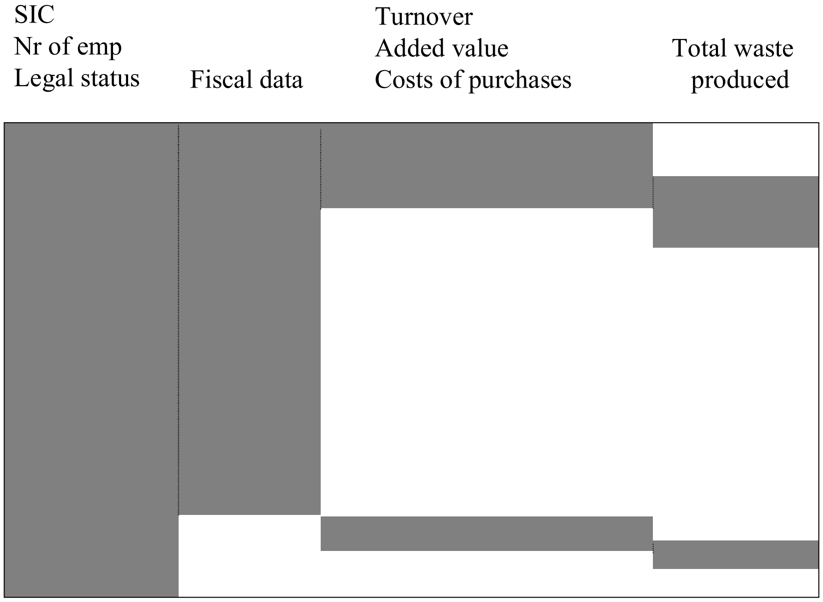

Figure 2: a simple fictitious micro-database for businesses; gray indicates data, white empty cells

Matched to this register is an administrative register originating from the Ministry of Finance. The register contains the tax declarations of many (but not all) businesses in the register. The register contains variables from the profit and loss account, but defined and measured in a different way than is needed for statistical publications.

Also matched in the micro-database are data from two business surveys: the production-survey (here with three variables: ‘turnover’, ‘added value’, and ‘costs of purchases’) and the environment-survey (one variable: ‘total waste produced’). Note that some businesses are observed in both the production-survey and the environment-survey. It is assumed that the design of both surveys is STSI, that is, both samples are drawn by simple random sampling within strata. To be more precisely, the population is partitioned in G size classes and in A activity classes. The size class a business belongs to depends on its number of employees. The activity class a business belongs to depends on its SIC code. For both surveys, each combination of a size class and an activity class defines a stratum, giving $K=G.A$ strata. For later use we define some notation.

u_k : the set of all businesses in stratum k

$s_k^{(ps)}$, $s_k^{(es)}$: the set of all businesses in stratum k that are observed in the production-survey, environment survey

$y_{k,i}^{(tu)}$, $y_{k,i}^{(ad)}$, $y_{k,i}^{(cp)}$: 'turnover', 'added value', 'costs of purchases' of the *i*th business in stratum k

$x_{k,i}$: 'number of employees' of the *i*th business in stratum k

$t_k^{(x)} = \sum_{u_k} x_{k,i}$ and $t_k^{(tu)} = \sum_{u_k} y_{k,i}^{(tu)}$.

In the following we focus on the estimation of population totals of the survey variables. The administrative information is only used as auxiliary information.

3. MASS IMPUTATION

3.1. The method

The method of estimation by means of mass imputation consists of two steps. In the first step, all empty cells of figure 2, i.e. all cells for which the true value is unknown, are filled with some imputation value. The result of this step is a completely filled micro-database: a table with a row for each business, a column for each variable and a value in each cell. Obviously, values are either observed or imputed. In the second step an estimate of a population total is obtained by adding observed and imputed values.

In practice it is not necessary to impute the variables in the register with fiscal data: no population totals of these variables are estimated. The fiscal data may be used, however, to improve the imputations for the survey variables.

The issue here is whether mass imputation generates consistent and reliable estimates. For a specific set of estimates, the answer mainly depends on the method of imputation. Before we discuss this in general we describe a specific (simple) procedure.

In our institute the stratified ratio-estimator ('number of employees' as auxiliary variable) is often applied in business statistics. This motivates the following imputation procedure for the micro-database in figure 2. For example, to impute 'turnover' for businesses not in the production-survey the following steps are carried out. First we calculate

$$\hat{B}_k^{(tu)} = \frac{\sum_{s_k^{(ps)}} y_{k,i}^{(tu)}}{\sum_{s_k^{(ps)}} x_{k,i}} \quad \text{for } k=1 \ldots K.$$

Second, if the *i*th business in stratum k is not observed in the production survey we impute $\hat{y}_{k,i}^{(tu)} := \hat{B}_k^{(tu)}.x_{k,i}$. We note that 'added value', 'costs of purchases', and 'total waste produced' are imputed analogously. Regression coefficients $\hat{B}_k^{(ad)}, \hat{B}_k^{(cp)}$ and $\hat{B}_k^{(tw)}$ are estimated and multiplied with 'number of employees' for businesses that are not observed in the relevant survey. For 'total waste produced', the summation is over $s_k^{(es)}$ instead of over $s_k^{(ps)}$.

Total turnover in the *k*th stratum is estimated by adding observed and imputed values

$$\hat{t}_k^{(tu)} = \sum_{s_k^{(ps)}} y_{k,i}^{(tu)} + \sum_{u_k \backslash s_k^{(ps)}} \hat{y}_{k,i}^{(tu)} = \hat{B}_k^{(tu)}.\sum_{s_k^{(ps)}} x_{k,i} + \sum_{u_k \backslash s_k^{(ps)}} \hat{B}_k^{(tu)}.x_{k,i} = \hat{B}_k^{(tu)}.t_k^{(x)} , \quad (1)$$

which is the standard ratio estimator for a stratum total. Total turnover in an activity class or size class is estimated by adding the estimates for the constituent strata. This results in the standard stratified ratio estimator.

3.2. Consistency

After the micro-database is imputed many estimates can be obtained by adding observed and imputed values. The issue here is whether all these estimates are consistent.

A record in the imputed micro-database contains scores on all variables for one business. Some of these values are observed, others are imputed. A necessary and sufficient condition for consistency of all estimates is that all records are consistent on the micro-level. A record is said to be consistent if there are no (hard) edit rules violated. An edit rule (in this context) is a rule that relates two or more variables.

An example of such an edit rule in the micro-database of figure 2 is 'turnover'='added value'+'costs of purchases'. Suppose this edit rule is violated in a number of records. In that case the estimated population total of 'turnover' (obtained by adding all observed and imputed values for this variable) is not necessarily equal to the sum of the estimated population totals of 'added value' and 'costs of purchases'. For the imputation procedure described in the previous subsection, the edit rule is satisfied for all imputed records, assuming the edit rule is satisfied for all observed records. This follows from

$$\hat{B}_k^{(tu)} = \frac{\sum_{s_k^{(ps)}} y_{k,i}^{(tu)}}{\sum_{s_k^{(ps)}} x_{k,i}} = \frac{\sum_{s_k^{(ps)}} (y_{k,i}^{(ad)} + y_{k,i}^{(cp)})}{\sum_{s_k^{(ps)}} x_{k,i}} = \hat{B}_k^{(ad)} + \hat{B}_k^{(cp)}, \tag{2}$$

Consequently, if there are no other edit rules in our simple micro-database this imputation method leads to consistent records on the micro-level and hence also to consistent estimates at the aggregate level.

The imputation method described in the previous subsection is a special form of regression imputation. In general, suppose we have a survey with variables that are subject to a number of linear edit rules. Suppose the observed records all satisfy these edit rules. Then, if all survey variables are imputed using the same regression model, all imputed records also satisfy the edit rules. Here the condition that the same regression model is used for all variables is essential. Another imputation method that is often used to ensure consistency at the micro level is 'simultaneous hot-deck'. For an overview of the various imputation methods that are used in practice see Kalton and Kasprzyk (1986).

It is sensible to use the information in the fiscal register to construct imputations. In order to do this the population can be split into two parts: businesses that are in the fiscal register and businesses that are not. The latter part is imputed, for example, by the simple imputation method described above. The former part is imputed using, for example, the method of regression imputation with auxiliary variables from the fiscal register. As correlations are usually high between fiscal variables and survey variables the 'quality' of the resulting imputations is expected to be high. As long as the same fiscal variables are used for all survey variables consistency is ensured.

Note that we have only discussed edit rules that concern variables in the same survey. If there is an edit rule that states a relation between variables from different surveys the situation is much more complicated.

3.3. Reliability

After the micro-database is imputed many estimates can be obtained by adding observed and imputed values. The question is whether all these estimates are reliable. Before addressing the issue in general we discuss the specific imputation procedure described in subsection 3.1.

3.3.1. Reliability of estimates generated by the imputation procedure of subsection 3.1

Formula (1) shows that the estimator of total turnover in a stratum obtained by the imputation method is the standard ratio estimator. This estimator is known to be approximately design unbiased, see e.g. Särndal et al. (1992). Therefore, also estimates for total turnover in a size class or an activity class, which are obtained by adding the estimates of total turnover in the constituent strata, are approximately design unbiased. Clearly, if the imputed micro-database is only used to publish total turnover (added value, costs of purchases, total waste produced) for strata, size classes or activity classes all estimates are reliable, provided that there are sufficient observations to ensure reasonably small design variances.

The situation is different if other estimates are generated. Suppose for example, we are interested in total turnover of all limited liability companies. Whether a business is a limited liability company or not can be deduced from its legal status (known for all businesses in the micro-database of figure 2). An estimate of their total turnover is obtained by adding all observed and imputed values of limited liability companies in the imputed micro-database. The question is whether this estimate is reliable.

The corresponding estimator is not approximately design-unbiased. Even if there are no limited liability companies observed in the production survey the estimate is still defined; only imputed values are added up in that case. Only if the limited liability companies have the same 'expected' turnover as other companies with the same number of employees in the same stratum, this estimate is reliable. Whether this assumption is plausible or not has to be decided by subject matter specialists.

In general, total turnover (added value, costs of purchases, total waste produced) can be estimated for an arbitrary subset of businesses by adding the relevant observed and imputed values. Such estimates are only reliable if it assumed that, after correction for the explanatory variables 'number of employees', 'activity class' and 'size class', businesses in the specific subset are 'no different' from other businesses.

We need the same kind of assumption, which we will call the conditional independence assumption (CIA), when estimating the relation between variables in different surveys. For example, suppose we are interested in the relation between 'turnover' and 'total waste produced'. Is it true that businesses with a large turnover produce a lot of waste? A way to measure this relationship is to estimate covariances. However, using both observed and imputed data the imputation procedure only leads to reliable estimates if 'turnover' and 'total waste produced' are 'independent' conditionally on the auxiliary variables 'number of employees', 'activity class' and 'size class'. If this CIA-assumption is not satisfied, the estimate obtained from the imputed dataset underestimates the relation; in the imputed values only the relation is found that is 'explained' by the auxiliary variables.

3.3.2. Reliability of estimates generated by imputation procedures in general

In Kroese and Renssen (1999) it is argued that, in general, there is no imputation procedure that generates only reliable estimates. In constructing the imputations only a limited set of explanatory variables can be dealt with. Only if the interest is in totals of subsets of the population defined by the explanatory variables in the model, the imputation approach leads to approximately design-unbiased and hence to reliable estimates (at least if the variances are reasonably small). If the interest is also in totals of subsets defined by other variables, the imputation approach may lead to unreliable estimates, unless the CIA-assumption is satisfied for each subset. In general, there is no reason to belief such an assumption uncritically.

Usually, only a very limited number of explanatory variables can be included in the model and hence, only a very limited number of estimates are approximately design unbiased. This is obviously a drawback of the (mass) imputation approach.

The disadvantages of mass imputation are larger for statistics about persons than for business-statistics. There are several reasons for this

- Sampling fractions are usually larger for business statistics. As a consequence the fraction imputations is larger in the imputed micro-database for persons than in the imputed database for businesses.
- Missing data usually concerns small businesses. The (small) set of large businesses accounts for most of the total turnover (added value etc.) in the population. If these large businesses are all observed it is relatively less important how we estimate the total for the smaller businesses.
- Registers are less informative for business statistics than for person statistics. In business statistics there is not a lot of demand for totals of all kinds of subsets. Usually only activity code and size class is used to generate subsets.
- In our institute there is more demand to study relations between variables in different person surveys than to study relations between variables in different business surveys.

Still mass imputation is not the optimal way to proceed from micro-databases to StatBase. In the next section it will be shown that weighting is more promising.

4. WEIGHTING

4.1. The method

The method described here can be seen as a new application of old weighting techniques and consists of four steps: constructing rectangular micro-datasets, assigning a set of weights to each micro-dataset, calculating estimates that can be obtained consistently, and reweighting the micro-datasets to obtain other estimates.

All steps will be illustrated by means of the micro-database in figure 2. To obtain estimates this micro-database is split into two parts: the records for which fiscal data are available and the records for which these are not available. Estimates are made for the two parts separately and these estimates are added to obtain overall estimates. Motivation for this approach is that the variables in the fiscal register are highly correlated with the survey variables. Consequently, it is attractive to use the fiscal variables in the estimation process. This can only be done for the records for which this information is available.

In the following subsections we only consider the part of the micro-database for which fiscal data are available.

4.1.1. Constructing rectangular micro-datasets

In figure 3 the generation of rectangular micro-datasets is illustrated by means of the micro-database of figure 2. It is shown that the part of the micro-database for which fiscal information is available generates four rectangular micro-datasets.

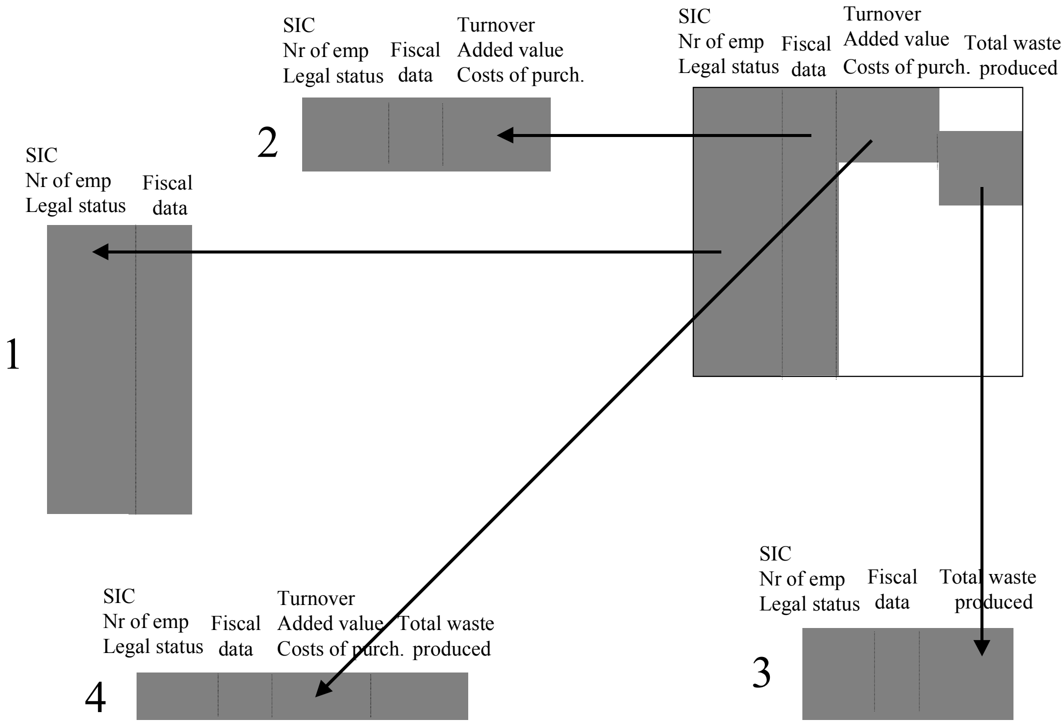

Figure 3: splitting part of the micro-database of figure 2 into four rectangular micro-datasets

Dataset 1 is the business register matched with the fiscal register; it is used to estimate totals like 'total number of employees by activity class' and 'total number of limited liability companies'.

Dataset 2 is the production-survey enriched with register information. This dataset is used to estimate totals like 'total turnover by size class' and 'average added value by activity class'.

Dataset 3 is the environment-survey enriched with register information. This dataset is used to estimate totals like 'total waste produced by legal status'.

Dataset 4 contains the set of businesses that are observed in both the production-survey and in the environment-survey. It is used to estimate the relation between 'total waste produced' and the variables in the production-survey.

In general, each partly filled micro-database can always be split into a finite number of rectangular micro-datasets. In Kroese and Renssen (1999) a more complex example is given.

4.1.2. Assigning a set of weights to each micro-dataset

In order to obtain estimates each rectangular micro-dataset is assigned a set of weights. The order in which the sets of weights are calculated is important. Rectangular micro-datasets with a lot of records are assigned weights first. In assigning a set of weights to a rectangular micro-dataset we account for sets of weights that are assigned before to other rectangular micro-datasets. In the micro-database of figure 2 this is done as follows.

All records in dataset 1 are assigned weight 1, because this dataset corresponds to an integral registration. To obtain the sets of weights for the remaining datasets we follow the standard weighting approach in sampling theory, see e.g. Deville and Särndal (1992).

For example, to obtain a set of weights for the second dataset we first calculate the set of *inclusion weights*. Being the inverse of the net inclusion probability of the production-survey, the inclusion weight $d_{k,i}$ of the *i*th business in stratum *k* (if it is observed in the production-survey) is equal to the 'total number of businesses in stratum *k*' divided by the 'number of businesses observed in the production-survey in stratum *k*'. Secondly, the inclusion weights are *calibrated* with respect to one or more variables with known population totals. Let $G_{k,i}(.,.)$ be a distance measure and $z_{k,i}$ a vector that contains (for the *i*th business of stratum *k*) the scores on one or more of such variables. Let t_z denote the corresponding vector of known population totals. Then, a set of (calibration) weights $w_{k,i}$ is calculated that minimizes

$$\sum_{k}\sum_{s_k^{(ps)}} G_{k,i}(d_{k,i}, w_{k,i}) \tag{3}$$

under the restriction that

$$\sum_{k}\sum_{s_k^{(ps)}} w_{k,i} z_{k,i} = t_z \; . \tag{4}$$

Under some not very restrictive conditions, the estimator obtained by using calibration weights is asymptotically equivalent to the general regression estimator, which is known to be 'approximately design-unbiased', see e.g. Deville and Särndal (1992). For $G_{k,i}(d_{k,i}, w_{k,i}) = (w_{k,i} - d_{k,i})^2 \sigma_{k,i}^2 / 2d_{k,i}$ the resulting calibration estimates precisely correspond to general regression estimates ($\sigma_{k,i}^2$ specifies the variance structure, see Särndal et al. (1992)).

In the situation of dataset 2, suppose we choose $\sigma_{k,i}^2 \propto x_{k,i}$ and $z_{k,i}$ a *K*-dimensional vector, where the *k*th element is equal to $x_{k,i}$ and all other elements are equal to zero. Then the weight $w_{k,i}$ of the *i*th business in stratum *k* equals

$$w_{k,i} = \frac{\sum_{u_k} x_{k,i}}{\sum_{s_k^{(ps)}} x_{k,i}} \tag{5}$$

The corresponding estimator of the total turnover in stratum *k* is

$$\hat{t}_k^{(tu)} = \sum_{s_k^{(ps)}} w_{k,i} y_{k,i}^{(tu)} = \frac{\sum_{s_k^{(ps)}} y_{k,i}^{(tu)}}{\sum_{s_k^{(ps)}} x_{k,i}} \sum_{u_k} x_{k,i} \; , \tag{6}$$

which is the same estimator as in formula (1), i.e. the standard ratio estimator.

The above choice of $z_{k,i}$ corresponds to calibrating the weights with respect to the total number of employees of each stratum. In general, the weights can be calibrated with respect to all kinds of known population totals. There are several reasons to include a population total in the calibration restrictions of formula (4): correction for selective nonresponse, variance reduction and consistency of estimates. It is very attractive to calibrate the weights with respect to the information in the fiscal register. The survey variables are highly correlated with the fiscal variables and, hence, calibrating the weights with respect to these totals may lead to a large reduction in variance and nonresponse bias.

In a similar way we can derive (calibration) weights for datasets 3 and 4. The inclusion weights for the third dataset are easily derived from the first order inclusion probabilities of the environment-survey. The inclusion weights for the fourth dataset are slightly harder to derive. If the production-survey and the environment-survey are drawn independently, a natural choice is the following: the inclusion weight of a record in dataset 4 is the product of the inclusion weights of the corresponding records in datasets 2 and 3. Subsequently, calibration weights are calculated according to the formulas (3) and (4).

We note that, in view of the consistency requirement, the choice of the calibration restrictions is essential. For example, dataset 4 is used to estimate the relation between 'total waste produced' and the variables observed in the production-survey, while datasets 2 and 3 are used to estimate the marginal distributions of these variables. In order to ensure consistency, dataset 4 should be calibrated with respect to all marginal totals that can be estimated by using the weights of datasets 2 and 3.

The need for consistency can best be explained in terms of categorical variables. Suppose we are interested in the relation between 'turnover' and 'total waste produced'. A way to describe this relation is to transform these continuous variables into categorical variables. Classes are defined by a lower bound and an upper bound. A business belongs to a 'turnover' class if its turnover is larger than (or equal to) the lower bound and smaller than the upper bound.

Estimating the relation between two categorical variables comes down to estimating a contingency table. The marginal totals of this table are estimated by means of datasets 2 and 3. The interior is estimated by using dataset 4. In order to have the interior consistent with the estimated marginal totals, the weights of dataset 4 have to be calibrated with respect to these estimated marginal totals. In other words, the calibration restrictions for the weights of dataset 4 should contain the requirement that the estimated marginal distributions of the (categorical) variables 'turnover' and 'total waste produced' are reproduced.

4.1.3. Calculating estimates that can be obtained consistently and reweighting the micro-datasets to obtain other estimates

Estimates of population totals are obtained by weighted counting in the relevant rectangular micro-dataset. Dataset 1 is used for estimates of population totals that only involve register variables like 'total number of limited liability companies'. All such estimates can be obtained by counting and are all consistent and reliable. Dataset 2 is used for estimates of population totals of variables in the production-survey, for example, 'total turnover in the population'. Such estimates are obtained by weighted counting, using the calibrated weights assigned to the dataset. The dataset is also used to estimate the relations between variables in the production-survey and variables in the register, for example 'the relation between turnover and number of employees for each stratum' and 'total turnover by legal status'. There are two possibilities

- The register variables have been included in the calibration restrictions. In this case the estimate can be obtained by using the calibrated weights assigned to the dataset.
- The register variables have not been included in the calibration restrictions. In this case it is less attractive to obtain the estimate by using the calibrated weights assigned to the dataset.

In the example of subsection 4.1.2, the weights of micro-dataset 2 are calibrated with respect to 'total number of employees per stratum'. This means that 'the relation between turnover and number of employees for each stratum' can be estimated by using the calibrated weights. The estimate is consistent with the total number of employees in each stratum which is known from the register.

Suppose we are also interested in an estimate of 'total turnover by legal status'. For a specific category of legal status, the sum of the derived calibration weights is not necessarily equal to the known number of businesses in that category. So, the estimated turnover in a category of legal status does not necessarily apply to the correct number of businesses. This inconsistency would become evident if, together with the total turnover in a category of legal status, an estimate would be given of the number of businesses in the category based on weighted counting in dataset 2.

A solution is to *reweight* the micro-dataset 2 in order to obtain a consistent estimate of 'total turnover by legal status'. A new set of weights $w_{k,i}$ is calculated that minimizes the expression in formula (3) under the calibration restrictions of formula (4). The calibration restrictions (defined by the vector $z_{k,i}$) are chosen such that a consistent estimator of 'total turnover by legal status' can be obtained by the resulting weights. As a consequence, the calibration restrictions should contain *at least* the requirements that

- The known total number of businesses for the different categories of legal status is reproduced and
- The estimated total turnover is reproduced.

Other calibration restrictions may be added in order to reduce variance and to correct for nonresponse. For example, it is natural to add calibration restrictions based on fiscal information. After the calibration weights have been calculated, the estimate of 'total turnover by legal status' can be obtained. We note that the rectangular micro-dataset is calibrated with respect to an estimated population total. In Renssen and Nieuwenbroek (1997) a similar two-step procedure has been described.

Obtaining estimates from datasets 3 and 4 can be done in a similar way. Dataset 4 is used to obtain estimates of the relations between 'total waste produced' and variables in the production-survey (and possibly variables in the register). If the weights of dataset 4 are calibrated with respect to all variables involved, the estimate can be obtained by weighted counting using the calibrated weights assigned to the dataset. If some variables are missing from the calibration restrictions, it is better to reweight the dataset and to calculate the estimate by using the new set of weights.

4.2. Consistency and reliability

In the previous subsection a weighting approach is described to proceed from micro-databases to StatBase. All estimates are obtained by using weights that are calibrated to known or estimated population totals. As a consequence, all estimates are approximately design-unbiased. Margins can be calculated and if these are reasonably small, the estimates are reliable.

The claim in this paper is that the estimates are also consistent. It is not possible to obtain contradictions by combining estimates obtained by the weighting method as described above (provided the micro-data contains no contradictions).

The reasoning behind this claim is the following: suppose we are interested in the relation between the variables Y_1 and Y_2. An estimate of this relation is obtained by weighting the largest rectangular micro-dataset D that contains both variables. If the marginal totals of these variables can be estimated from even larger datasets, the weights of D are calibrated with respect to these estimated marginal totals. In this way the estimate of the relation is consistent with the estimates of the marginal totals.

The term ‘relation’ is left deliberately vague. The relation between two categorical variables is simply the contingency table. Consistency means that the row- and column-totals match the interior. For continuous variables the term ‘relation’ is less well defined. A way to present the relation between ‘turnover’ and ‘legal status’ is a table with total turnover for each category of legal status. Consistency here means that, for each category of legal status, the estimate of total turnover is based on the known total number (of employees) of businesses.

It can be seen immediately whether a contingency table is consistent or not. This is not true for a table with total turnover by legal status. The inconsistencies only become obvious if also a table is published with the average turnover by category of legal status or if turnover is categorized and a contingency table is published.

In reweighting rectangular datasets in order to ensure consistency special attention has to be given to edit rules.

5. CONCLUSIONS

In this paper two estimation methods are described for proceeding from micro-databases to StatBase. The method of imputation leads, under some conditions, to a consistent set of estimates. Only a very limited number of estimates, however, is reliable. A weighting method is described as a more attractive approach. It is shown how information from administrative registers can be incorporated. At the moment, this approach is tested extensively in our institute.

6. REFERENCES

Deville, J.C. and C-E. Särndal (1992), “Calibration estimators in survey sampling”, *Journal of the American Statistical Association,* **87**, pp. 376-382.

Kalton, G., and D. Kasprzyk. (1986), "The treatment of missing survey data," *Survey Methodology,* **12**, pp. 1-16.

Keller, W., Bethlehem, J., Willeboordse, A., and W. Ypma (1999), "Statistical processing in the next millenium," *Proceedings of the XVIth Annual International Methodology Symposium on Combining Data from Different Sources, May 1999 Canada*.

Kroese, A.H., and R.H. Renssen (1999), "Weighting and imputation at Statistics Netherlands," *Proceedings of the IASS conference on Small Area Estimation, Riga August 1999*, pp. 109-120.

Renssen, R.H., and N..J. Nieuwenbroek. (1997), "Aligning estimates for common variables in two or more sample surveys," *Journal of the American Statistical Association ,* **92**, pp. 396-374.

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

COMBINING SURVEY AND ADMINISTRATIVE DATA: DISCUSSION PAPER

Pierre Lavallée, Statistics Canada
Statistics Canada, Ottawa, Ontario, K1A 0T6, Canada, plavall@statcan.ca

1. INTRODUCTION

The present paper is a discussion on the papers "Use of Administrative Data as Substitutes for Survey Data for Small Enterprises in the Swedish Annual Structural Business Statistics" by Erikson and Nordberg (E&N), "Use of Business Income Tax Data to Extend the Information Available from the ABS Economy Wide Economic Activity Survey" by Crabb and Sutcliffe (C&S), and "New Applications of Old Weighting Techniques – Constructing a Consistent Set of Estimates Based on Data from Different Sources" by Kroese and Renssen (K&R).

Administrative sources have been used since years in northern countries of Europe (e.g. Sweden and Finland) in the production of statistics. This practice is also extending to other countries, as these sources become available in computer readable format and offer some great potentials to replace direct surveys. Although the gains in using administrative data can be large, some problems often occur in their use. Most of the time, unfortunately, these problems were under estimated when planning the use of the chosen administrative data. The three papers mentioned above describe some useful uses of administrative data combined to survey data. They also reveal the problems that were needed to be faced before their proper use. This will be the focus of the present discussion.

2. TYPES OF USE OF ADMINISTRATIVE DATA

E&N used administrative data for two purposes. First, administrative data were used for replacing direct data collection for small enterprises less than 50 employees for the Structural Business Survey (SBS). For the enterprises with 50 employees or more, the administrative data was not found of sufficient quality. The second purpose of the use of administrative data by E&N was to quantify frame errors.

K&R discussed the problem of estimation when several databases are merged together. These databases might contain data from administrative registers, surveys or electronic data reporters. The problem comes from the fact that merging these databases lead to "holes" in the resulting dataset because not all units of the population have data in each of the initial database. K&R discussed two methods to produce reliable and consistent estimates from the resulting database. The two methods are mass imputation and weighting.

Although the work of K&R is quite valuable, they seemed however to miss two important points. First, the main problem with mass imputation is the fact that the users have the misconception that the database has been completely filled with real collected data, which is obviously not true because of the use of imputation. Because of this, users have the feel that they can produce estimates for any subpopulation of interest. This can lead to estimates based only on imputed data, and consequently to the reliability problem mentioned by K&R. This cannot happen with the solution based on weighting only because a subpopulation that would be based on only imputed data would, in this case, have no data to produce the estimates. A second point missed by K&R is the relationship existing between the definition of the imputation classes and the calibration variables. Weighting corresponds in fact to impute missing data by using the mean of the available units computed within the imputation classes. These classes are often corresponding to the strata, but they do not need to. In fact, it can be argued that the problem of reliability with mass imputation is quite related to the problem of choosing the calibration variables for the weighting solution.

C&S described an effective use of administrative data that lead to a reduction of the sample size of the Australian Economic Activity Survey (EAS), and consequently also to a reduction of the response burden of firms. The EAS was a yearly survey collecting financial information directly from a sample firms. Administrative data have been used to replace some of the financial variables collected from the EAS. For complex firms, the new EAS still collect the data directly from the firms. For the non-complex firms, a "two-phase" sample design is used. A "first-phase" sample is first selected from the population of administrative data to get some financial variables. To get the financial variables for the breakdowns, a "second-phase" sample of firms is selected and the additional information is then collected directly from the firms. Statistics Canada uses a similar type of sample design to the EAS described by C&S for its Survey on Employment, Payroll and Hours.

3. BENEFITS OBTAINED BY THE USE OF ADMINISTRATIVE DATA

3.1 Imputation for non-response (E&N, K&R)

For several surveys, administrative data is used as auxiliary information to impute missing data collected through surveys. This auxiliary information is used in imputation methods such as modeling where the administrative data is used as covariates, or nearest neighbor imputation where the closest neighbor is found using the administrative data.

3.2 Compensation of bias due to cut-off sampling (K&R)

With cut-off sampling, a part of the target population is not covered by the survey process. In other words, some units have a zero selection probability. Cut-off sampling is generally used for business surveys where we do not sample small firms below a given threshold (or cut-off point). Administrative data is then used to calibrate the estimates to the whole population, or to impute for the small excluded firms.

3.3 Availability in computer form (K&R)

Most of the administrative data is available in computer form. Therefore, if is free of capture costs that are usually non-negligible. Note however that administrative data do normally have some processing costs to transform the data in a usable format for a given application.

3.4 No sampling error (E&N, C&S)

Administrative data is often available on a census basis. In that case, there is no sampling error associated to them.

3.5 Reduction of response burden (E&N, K&R, C&S)

The use of administrative data instead of data obtained by direct collection reduces the response burden of the firms of the target population. Filling up of administrative records by firms is usually mandatory. Collecting data that is already available from administrative sources is simply adding to the burden of the firms.

3.6 Reduction of collection costs (E&N, K&R, C&S)

Collecting data directly from firms is a costly process. Hence, taking advantage of administrative data that is already available enables to reduce the amount of variables that are measured by questionnaire.

3.7 Production of estimates in finer details (E&N, C&S)

As mentioned earlier, administrative data are often available on a census basis. When it is the case, it is possible to produce statistics for any domain of interest, even if this domain is very small and therefore does not contain a lot of firms. Note that the choice of the domain will be restrained to what can be identified from the data itself. For example, one cannot produce statistics for a given city if the geographical coding is not lower than the provincial level.

3.8 Frame update (E&N, C&S)

Administrative data is often used as a source for updating frames. In national statistical institutes, the frame is usually extracted from a business register and one source of update of the register is administrative data coming from tax records, for example.

3.9 Supplement data collected though surveys (C&S)

This benefit is close to the one that is related to imputation. One way to reduce the response burden of firms is to diminish the number of collected variables. In that case, administrative data can be used to assign (or impute) a value to the variables excluded from the questionnaire.

3.10 Benchmarking (K&R, C&S)

This is an important use of administrative data. As administrative data are often available on a census basis, aggregated values can be obtained that are merely free of sampling errors. These values can then be used as control totals to adjust the survey estimates for the following reasons. First, it produces consistent estimates between surveys. Second, if can adjust the survey estimates for under- or over-coverage of the population. Finally, to some extend, it can correct for non-response.

4. LOSSES FROM USING ADMINISTRATIVE DATA

4.1 Limited control on the data (quality checks, processing time, etc.) (E&N, C&S)

The primary user of the administrative data is often with low interest in the data as a statistical source. Therefore, the provider of administrative data might take some decisions that can affect the statistics program. For example, it might be decided to change from monthly to quarterly the frequency at which the administrative data are released. This can then jeopardize monthly surveys that use these data. As another example, by removing one variable from the administrative data file, one might affect the imputation process, if this variable was used in the imputation model. Finally, because the primary user does not consider a certain variable important, some limited quality checks might be done on it. In that case, the statistics produce for that variable might contain a large amount of non-sampling errors.

4.2 Restricted number of available variables (E&N, K&R, C&S)

Administrative data often contain a limited number of variables. This is because data entry is expensive and then the choice is often made to have a limited number of variables on all firms, instead of the opposite. The available variables are usually restricted to the ones required by the primary users, which might be different from the ones needed for statistical purposes. Because, of this, one needs to supplement administrative data by the use of others sources of data such as surveys.

4.3 Missing units and missing items (E&N, K&R, C&S)

Missing units and missing items are two types of non-response that are present with administrative data. Although providing answers to administrative data collection is usually mandatory, some non-response can still exist. In that case, it more often caused by processing problems (computer problems, non-readable fields, corrupted files, etc.) than non-response due to non-compliance. In the case where some units and/or items are missing, imputation systems need to be developed.

4.4 Processing errors (E&N, C&S)

Several processing errors can occur when dealing with administrative data files. This is because these files are often very large and then require large computer resources to be processed. As mentioned earlier, these processing errors can lead to non-response units and items that need to be corrected. Note that the processing errors are not solely related to computer problems and may also be due to keying errors, consistency problems, coding errors, etc.

4.5 Coverage (or frame) errors (K&R)

Although administrative data can be useful to supplement survey data, the population covered by the administrative data can be different from the target population of the survey. For example, administrative data can only be available for firms with more than 10 employees while the survey covers all firms. As another example, administrative data can be restrained to some economic activities.

4.6 Lack of industrial classification (K&R)

An industrial classifications such as the North-American Industrial Classification System (NAICS) or the *Nomenclature d'Activité économique de la Communauté Européenne* (NACE) is vital for the production of economic statistics. The assignment of industrial codes is however a costly process that should not be under-estimated for the following two reasons. First, it usually needs to be assigned to all units for which we have administrative data, which might be very large. Second, because administrative data have a limited description of the firm activity (if not any at all), it is often difficult and costly to assign a code.

4.7 Timeliness (C&S)

Timeliness is often neglected when deciding on choosing or not some sources of administrative data. This might be important however if, for example, the administrative data is available only several years after their reference period. In that case, it might slow down the production of the statistical product up to the point when they become useless.

4.8 Editing and processing costs (C&S)

Although administrative data are often available in computer form for the entire population, they are not free of editing and processing costs. As mentioned earlier, because of processing errors, missing items, missing units, etc., administrative data need to be processed in order to make them useful for the production of statistical figures.