WP.12
English only

UNITED NATIONS STATISTICAL COMMISSION and
ECONOMIC COMMISSION FOR EUROPE

CONFERENCE OF EUROPEAN STATISTICIANS

Work Session on Statistical Data Editing
(Madrid, Spain, 20-22 October 2003)

Topic (iv): Data editing by respondents and data suppliers

TREATMENT AND EDITING OF TAX DATA FOR SWEDISH STRUCTURAL BUSINESS STATISTICS

Invited Paper

Submitted by Statistics Sweden[1]

ABSTRACT

From 1996, the Swedish Structural Business Statistics (SBS) is a total enumeration of all enterprises. It was possible to make this change since administrative data from tax authorities was available. Since the tax data is processed for the purposes of tax collection and revisions, and not for statistical purposes, the material has to go through an editing process at the statistical office to fit the statistical needs. It is also necessary to calculate or estimate a number of variables that are needed for the statistics but not included in the tax data. The paper deals with the methods used at Statistics Sweden to do these calculations and estimations, and the editing procedures used to make sure that the data is usable for statistics. The method used is based on finding possible extreme values and checking these in a systematic fashion, and also explaining all of the largest of the calculated and estimated missing indicators. The method is described in detail in the paper.

INTRODUCTION

The paper is divided into six chapters. In chapter one, a brief overview of the SBS survey is presented. After that, the paper turns to the treatment of the administrative data. Since the administrative data does not contain all data necessary for the purposes of SBS, calculations and estimations of missing items is necessary. In chapter two, the Swedish models for calculating some of the most important missing items are presented. The paper then turns to the editing and imputation process regarding the administrative data. The editing process can roughly be described as having two key ingredients; checking for extreme values and large errors, and validating the calculations made at the statistical office. This editing and imputation process is presented in chapter three and four. In chapter five, some outlines for further using the administrative data is presented, such as using administrative data to increase the quality of the Business Register. The main findings and conclusions are given in chapter six.

1. AN OVERVIEW OF THE SWEDISH SBS SURVEY

The Swedish Structural business statistics (SBS) is an annual survey designed to fulfil the needs of National accounts as well as providing the data required by the Council Regulation (XX/XX) of

[1] Prepared by Johan Erikson, Dept. of Economic Statistics, Statistics Sweden, S-701 89 Örebro, Sweden (johan.erikson@scb.se).

Structural business statistics. This means that not only does the survey need to provide information on a large number of indicators, it also needs to provide data on three different levels; institutional level, activity code level and regional level. To manage this, the statistical units enterprise, kind-of-activity unit (KAU) and local kind-of-activity unit (LKAU) are used. Data on the institutional level is based on the enterprise, activity code data on KAUs and regional data on LKAUs.

The SBS survey of today is based on two previous surveys – the manufacturing statistics that can be said to have a history that dates back to the nineteenth century and is available in a published format from 1913 onwards, which was based on establishments (local units) as statistical units, and the financial accounts statistics, which was introduced in 1951 and over time grew to incorporate the whole business sector, and was based on enterprise data.

These two surveys ran parallel lives up to 1997, when the new demands coming from entering the EU meant that production data by commodity had to be delivered at short notice. This meant that this data, which was previously collected within the manufacturing statistics, had to be collected in a separate survey. At the same time, it was clear that some double-collecting of information was taking place between the manufacturing statistics and the financial accounts statistics with extra efforts needed to have coherent data being necessary. All of this led to a restructuring of the statistics with the financial accounts statistics and the "rest" of the manufacturing statistics merging into the new SBS survey.

In the beginning of the 1990s, the tax authorities in Sweden began collecting accounting data for tax purposes. The declaration of income by legal entities was expanded with a form where enterprises had to fill in extracts of the profit and loss accounts, balance sheet and some other information needed. This tax form is called Standardised Account Extracts or, in Swedish, Standardiserade räkenskapsutdrag (SRU). The term SRU will be used throughout this paper. Rather soon, it was decided that this material could be used for statistical purposes. The financial accounts statistics was a sample survey, with a total enumeration of enterprises with more than 50 employees, and a sample below that threshold. The manufacturing statistics was a total enumeration of enterprises with at least 10 employees, with an estimate of small enterprises based on accounting data bought from a private firm to complete the picture somewhat. The SRU could be used to increase the quality of both these surveys. After spending a couple of years studying the quality aspects of using this material, it was introduced in the financial accounts statistics from 1995, which was then changed into a total enumeration (for 1995, data based on both methods is available). This satisfactory move was a key element in designing the new SBS survey. It was decided to make this a total enumeration, and to raise the threshold of questionnaires to 50 employees. With the financial accounts statistics and the manufacturing statistics merging, this lowered the burden put on small enterprises considerably. The SRU was to be used for all enterprises below the threshold of 50 employees, and no other data collection would be made for these enterprises. It was also decided, to simplify matters, to treat all small enterprises as having only one KAU, meaning that the SRU material could be used as it was for both enterprise data and KAU data. And that is the way it has been since then. Now, the SBS survey is going through another re-design, with the purpose of using the SRU material even more, but instead collect data that is not available from the administrative source by a sample survey. More about this in chapter four. The main part of this paper will concentrate on the use of SRU as it has been in the SBS survey from 1997 up to now.

2. CALCULATION OF MISSING ITEMS IN THE SRU MATERIAL

The SRU contains profit and loss account data and balance sheet data, but not much more. There is some information on investments, but it is still of lower quality than the rest of the information and difficult to use for statistics. On changes in equity, data which is needed to calculate the sector accounts, there is no information in the SRU. Furthermore, there is no data on the number of employees. This means that there are three areas where there is a need to complement the administrative data with other indicators The methods used vary slightly, but most of these indicators are calculated from the SRU material itself. Data on investment and changes in equity is calculated

from the SRU material itself, but in order to make such calculations, it is necessary to have data for two years, both this year and last year. If you have that information, it is possible to make good calculations. The number of employees is taken from a different source of information.

2.1 Number of employees

The indicator of number of employees is the only indicator where the main source is not a calculation based on the SRU material. While it is possible to come up with a simple model that calculates the number of employees based on the wages and salaries, this indicator is deemed so important it was better to find another source of information. As it happened, this was available from a private company that registers data from annual reports into a database. (The main reasons to use SRU instead of using this source altogether is that it has more detailed information and looks the same for all enterprises whereas profit and loss account data in annual reports vary due to a possibility to choose from two different arrangements of this part of the annual report that gives different information.) Getting this information on number of employees from a different source means that this key indicator will be of good quality with a certain definition, and it will also give an additional opportunity to make quality checks of the SRU material in the editing phase, see chapter three.

2.2 Investments

To make calculations on investments, the statistical office has created a model that uses both profit and loss account data, balance sheet data and the investment data given in the SRU form. Together, this information gives better statistics than using the data given in the SRU as it is. Data must be calculated for both gross investment and net investment (gross investment minus sales), and also by type of asset: machinery and equipment, buildings and land. In order to come up with this, the model uses a three-step approach: first calculating total net investment, then total gross investment and finally distributing investment by type of asset.

The first step is the easiest, and has been proven to give very good calculations of total net investment. Using balance sheet data and profit and loss account data, a simple formula can be created:

BS (t-1) + inv (t) - profit (t) - rev (t) - dep (t) = BS (t) which might be rewritten as
Inv (t) = BS (t) + BS (t-1) + profit (t) + rev (t) – dep (t)

BS=balance sheet total of fixed assets
Inv=net investment
Profit=profit from the sale of fixed assets
Rev=Revaluation=Write-up minus write-down
Dep=Depreciation

BS is available from balance sheet data, while profit, rev and dep are available from profit and loss account data, meaning inv is the only unknown variable and easy to calculate. For example, if the balance sheet total is 1000 in t-1 and 2000 in t, the depreciation was 100 and there was also a write-down of 200 during the year, investment is simply calculated as
Inv=2000-1000+100+100=1200, meaning that an investment of 1200 covers a writedown of 100 and depreciation of 100 while still increasing the balance sheet total by 1000.

To get to gross investment, the simple formula is that

Net inv = Gross inv – Sales, or Gross inv = Net inv + Sales

Sales of machinery and equipment can be taken from the investment data given in the SRU, and a rough estimate of sales of buildings can be calculated, meaning it is quite easy to calculate the gross investment once the net investment has been calculated. The main problem with this step in the model is that the investment data given in the SRU is of a lower quality than other indicators, mainly because

this information is missing in many cases. This together with the fact that the sales of buildings can only be roughly estimated means that gross investments will tend to be underestimated, since data on sales is missing or too small in some cases.

When it comes to distributing the investment by type of asset, the model becomes a bit shaky. The BS indicator is available by asset, as well as the dep indicator. The profit and rev indicators on the other hand are not available by type of asset, meaning they have to be distributed by type of asset within the model. This in turn lowers the accuracy of the model.

All of these calculations are based on the assumption that SRU data is available for both this year and last year. If data for last year is missing (for example if the enterprise was non-responding last year) the model cannot be used. Then there is only the possibility to use the data that is available in the SRU, even if it is of lower quality. So for these enterprises, the model is even simpler, just using existing data and making no adjustments to it.

2.3 Changes in equity

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, ΔRR = write-up

$\Delta ET = \Delta SC + \Delta UR \Rightarrow \Delta SC =$ new share issue, ΔRR = 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:

ΔSC = decrease of share capital, ΔRR = write-down, ΔOR + ΔUR = 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:

ΔSC + ΔOR = new share issue including agio, ΔRR = write-up, ΔUR = shareholders´ contribution

Enterprises belonging to category c are left for imputation (automatic or manual).

3. THE EDITING PROCESS

The editing process for the SRU material is in two steps and has three main objectives. In the first step, enterprises for which the material shows internal incoherence are targeted and selected. These enterprises will be either corrected or left as non-response. The second step has two objectives. The first is to find large errors in the material and the second is to validate single indicators, most importantly those that have been calculated.

In the first editing phase, simple rules have been developed that say that the profit and loss account and the balance sheet must “sum up”, which might be simplified in four rules:

1. The sum of all variables in the profit and loss account must be the same as the profit or loss for the year.
2. The sum of all variables in the balance sheet regarding assets must be the same as the balance sheet total for assets.
3. The sum of all variables in the balance sheet regarding equity and liabilities must be the same as the balance sheet total for equity and liabilities.
4. The balance sheet total for assets must be the same as the balance sheet total for equity and liabilities.

These rules must not be exactly fulfilled for an enterprise to pass this control, but the acceptance limits are quite strict. For an enterprise with a turnover of less than one million SEK, the rules must be fulfilled within an acceptance limit of plus or minus 20 SEK, for larger enterprises the acceptance limit is 1000 SEK. This might seem as very strict rules, but there are two reasons to keep them strict. The first is that experience has shown that for enterprises of at least some size, the normal case is that either are these rules fulfilled exactly, or either are there quite large differences. This means that raising the acceptance limits would not mean many more enterprises that do not pass the control. The

second reason is that many enterprises are so small that their actual totals are close to the acceptance limits. The number of such small enterprises is quite large, and this means that raising the acceptance limits might result in many small errors being left in the material.

Those enterprises that do not pass the control in this phase are put in one of two categories. If the enterprise has large figures (turnover of 100 million SEK or more), it must be corrected manually, it cannot be left as non-response. Such enterprises are checked by comparing SRU data to annual report data, and are corrected manually so that the differences are no longer outside the acceptance limits.

Smaller enterprises are either corrected automatically or left as non-response. The methods of automatic correction are not very well developed, and there are not that many enterprises that are actually corrected automatically. There are only a few rules for automatic corrections, for example when the whole liability side of the balance sheet is with minus signs (this might happen in some accounting softwares when automatic reports are generated) and a few cases where a single value is put on the wrong line in the tax form (for example registered as a loss instead of a profit). In my opinion, the system of automatic corrections might and should be expanded, meaning that more enterprises will be corrected automatically and fewer enterprises left as non-response. Today, those enterprises that can not be corrected automatically based on these limited models are simply left as non-response, meaning that their SRU values will not be used and all variables will instead be imputed.

In the second editing phase, the purpose is to find large errors and validate single indicators. To do this, a set of editing rules have been constructed, in total about 20 rules. Each and every enterprise in the SRU material that passed the first editing phase (including those that were automatically corrected) are tested for these rules, and all enterprises that have at least one suspicious value are put on a list for manual checks.
About half of these editing rules have been constructed to find extreme values, possible large errors that might have an impact on the statistics as a whole. To do this, the rules try to find large changes in important indicators. Three main indicators have been chosen for this purpose: turnover, balance sheet total and value added. This gives an overview of both profit and loss account data and balance sheet data, and in the case of value added also makes it possible to roughly check both income and cost structures as well.

To find the extreme values, the system has been designed to select enterprises for which the changes in these indicators are both large enough in themselves to have some impact, and also have an impact on the statistics. After testing, it was decided to have a combinated threshold of a change of at least 50 million SEK and an impact on the total value for the relevant activity code of at least two per cent, using turnover as an example it would look like this:

(1) If abs (T(t)-T(t-1))<50000000 or 0.98 < TT(t-1)-T(t-1)+T(t) / TT(t-1) < 1.02, then the enterprise is OK, otherwise it is a suspicious value. T denotes turnover for the enterprise, and TT total turnover for the activity code (Nace 4-digit level).

The problem then arises when the enterprise is in another activity code this year than last year. If the enterprise has the same NACE code as last year, there will be only one check made whether the change in turnover has an impact on the total turnover for the NACE code. If, on the other hand, the NACE code changed between the years, two checks are necessary. One to check whether the addition to the new NACE code has an impact on the total for that NACE code, and one check whether the subtraction from the old NACE code has a large impact on that code. One additional advantage of this approach is that enterprises that changed their activity code and have large values for the main indicators are checked, they will also be checked to see whether this change in activity code is correct. This makes it necessary both to add a constraint to the first rule and to create two more editing rules for turnover:

If NACE(t)=NACE(t-1):

(1) If abs (T(t)-T(t-1))<50 SEK or 0.98 < TT(t-1)-T(t-1)+T(t) / TT(t-1) < 1.02, the value is OK

If NACE(t)≠NACE(t-1)

(2) If T(t)<50 million SEK or TT(t, NACE(t-1))+T(t)<1.02, the value is OK

(3) If T(t)<50 million SEK or TT(t-1, NACE(t-1))-T(t)>0.98, the value is OK

The same set of three rules are applied for balance sheet total and value added, except that for value added the threshold value is set at 25 million SEK instead of 50 million SEK. This is because the value added is calculated as income minus costs, and therefore will be somewhat smaller in size than turnover and balance sheet totals. The checks of turnover and balance sheet totals are quite straightforward, while the checks of value added also gives a possibility to find enterprises where there are changes in the cost structure between the years, since not all costs should be deducted from income to get to the value added (personnel costs are not to be deducted).

The second half of the editing rules are for validation of single indicators. The indicators that need to be validated are mostly those that were calculated by the statistical office. There are two main reasons for the checks: to check large values that were calculated and to select those enterprises where the model did not manage to make calculations. Between 1000 and 1500 enterprises fail one or more editing rules every year.

The indicators that are validated this way are:

- Total net investment (all totals of more than 100 million SEK are selected as suspicious values)
- Total gross investment (all totals of more than 100 million SEK are selected as suspicious values)
- Net investment by type of asset (all totals of more than 50 million SEK for a single type of asset are selected as suspicious values) – this check also incorporates a possibility to check large sales of assets.
- Net investment in shares and participations (all totals of more than 50 million SEK for a single type of asset are selected as suspicious values)
- Number of employees (a total of more than 50 employees is selected as a suspicious value).
- New emission of shares (all totals of more than 25 million SEK are selected as suspicious values)
- Dividends paid (all totals of more than 25 million SEK are selected as suspicious values)
- Shareholder´s contributions received (all totals of more than 25 million SEK are selected as suspicious values)
- Group contributions received or paid (all totals of more than 25 million SEK are selected as suspicious values)
- Other changes in equity (uncalculated changes in equity (more than 25 million SEK is considered to be a suspicious value).

This final editing rule is not just an editing rule, it is a signal to make manual adjustments to the data where the model on changes in equity is not good enough to be able to calculate estimates.

The editing is done manually, going into annual reports for each of the enterprises that failed at least one of the checks, and making necessary changes. It is not a process where these enterprises are checked in their entirety, many only have one or two suspicious values and in such cases it is only these values and no other that are checked manually, this is deemed a good way to use resources efficiently. It is better to check a single value for a larger number of enterprises than to check a smaller number of enterprises rigorously.

4. THE IMPUTATION PROCESS

Once the editing is finished, the missing enterprises need to be imputed and estimated. This non-response consists of two groups of enterprises: the ones that failed the first step of the editing process (simple summaries) and those enterprises for which no administrative data at all was received.

One might wonder what the second group of enterprises really is – if the enterprise is existent, it should make a declaration of income and be included in the administrative data. There are several reasons for this non-existence of administrative data. The first might be that the enterprise simply did not fill in the income declaration, the second that they sent it in too late and was not in the database at the tax authorities at the time of delivery to the statistical office. Some enterprises might for some reason not have been entered into the database at the tax authority. And finally, maybe the enterprise is no longer existing but we still need to have data for it.

This last category might seem a bit strange. The reason is that there is a common sample frame for all (or most) surveys in enterprise statistics, and this frame is the population for which data should be estimated. If there are some errors in this frame, this is nothing that can be corrected when the surveys are already running, if it is not changed in all surveys. Such a system might be desirable but is not running today. So it is better to have the same errors in all surveys than to have differences. Also, since there were other reasons why data could be missing, it might be impossible to distinguish whether it was due to one or the other of the factors that a single enterprise is missing. But this is an area where the administrative data could be used even more than today, as will be described in chapter five.

The imputation process in itself is rather simple. The method used is an "all or nothing" approach, meaning that either the SRU is used, and then all indicators are used, or it is not used, and then all indicators are imputed. It is possible to think of more sophisticated models of partial imputation when the profit and loss account data is OK but not the balance sheet data or vice versa, but that is not the situation today. The imputation is made by using average values for the relevant activity code and size class. The model is based on the assumption that a minimum number of observations is needed to be able to use averages as the basis of imputation. Therefore, the model uses the most detailed level of NACE code possible to make an imputation. If there are enough observations at 5-digit level, that basis is used, otherwise you turn to 4-digit level, 3-digit level and 2-digit level. If there are still enterprises that cannot be imputed by average values, a manual work with finding "twin enterprises" is used as the final solution, but this is very rarely needed.

5. FURTHER WORK WITH THE SRU

There is definitely room for improvement when it comes to using and treating the SRU material. Some areas where refined models would be good are:

- Better and expanded methods for automatic corrections, meaning a smaller amount of enterprises to impute.
- Improved editing rules, making the editing process more efficient and less resource-consuming.
- Better models of imputation, looking more at partial non-response and keeping parts of the administrative data that are correct.

Even though the SRU material has been used since 1995/96, there have been only slight adjustments to the models and methods used. This might be both due to lack of resources and lack of knowledge. This area is definitely not as well covered in literature as methods for sample surveys, and there have also been few knowledge-sharing experiences multilaterally. Hopefully this will change over time as administrative data become a larger part of the statistical data.

One specific area that I would like to mention is the possibility to use administrative data to improve the quality of the business register and the sample frames, and also to quantify frame errors.

The frame for the SBS survey is the Business Register (BR). The survey covers all enterprises in BR 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 undercoverage and overcoverage.

For the corporate sector, the undercoverage consists of enterprises that have sent in an SRU but that are not included in the BR frame. These enterprises can be divided into two categories; enterprises that are not included in the BR 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 overcoverage, consisting of enterprises that are classified as active in the BR 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 this presentation, I use numerical examples regarding 1997. Even though these are quite old figures, similar tests have not been carried out since then. There is no reason to think that the picture would look completely different today.

In 1997 there were 88 460 enterprises that sent in an SRU but that were not included in the BR frame. Of these 1903 were not included in the BR 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 BR 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. 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 overcoverage 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.

6. SUMMARY

- The Swedish Structural business statistics (SBS) survey has been using administrative data in the form of tax forms, SRU, since 1995. At the moment it is used for all enterprises with less than 50 employees.
- The SRU material does not contain all indicators necessary for the SBS needs. Therefore a number of models have been created to calculate these missing items. The most complicated models concern investments and changes in equity.
- The models are based on an assumption that data for two consecutive years is available. If this is the case, good calculations can be made. Otherwise, more rudimentary estimations or manual treatment is necessary.
- The SRU data also needs to go through an editing process. This editing process has three steps. In the first step, enterprises for which the data is incoherent are selected and left for imputation, automatic correction or manual treatment. In the second step, extreme values and large changes that have an impact on the published statistics are checked. The third step concerns validation of calculated items and other single indicators.
- Missing data in the SRU is imputed by a rather simple, “all-or-nothing” method, based on average values for the relevant activity code and size class.
- Even if this data has been used for a number of years, the methods for treating and editing the administrative data can be refined and improved. It is also possible to use this data to improve the business register regarding both undercoverage and overcoverage.