Stefan Berg

Addressing coverage and measurement errors using multiple administrative data sources

1. Summary

In the Swedish SBS administrative data in the form of tax data has been used since the mid 1990s. From reference year 2003 onwards the tax data is combined with data from three separate sample surveys within the SBS framework. The use of tax data has been beneficial for both the producer Statistics Sweden and for the respondents. As in any traditional survey there are problems with non-response, measurement and coverage errors. In this paper we describe how we obtain a rough estimate for the coverage errors using several administrative data sources. We also give two examples on how we work with reducing measurement errors, stemming from the use of administrative data.

2. Re-designing the Swedish SBS

During the last ten years we have used three different survey designs for the Swedish SBS. Each new survey design has been a valuable step towards a better SBS.

Statistics Sweden started to compile statistics about the financial accounts of enterprises in 1950. The first survey design was used until 1996, however with different modifications introduced during the period. The sampling design was stratified sampling. The stratification was made according to kind of activity and size. Since the survey was based on a sample, all estimates were affected by a sampling error. Estimates of change over time concerning two and three digit level by kind of activity tended to be reliable, but estimates of change at a more detailed level could give strange and un-reliable results.

The second step was taken in 1997. Statistics Sweden then started to use administrative data to compile SBS-statistics. This was done at the same time, as the statistics were adapted to new EU-regulations.

With this new survey design we gained access to data for (virtually) the whole population, as opposed to earlier when data only was obtained for the sampled (and responding) objects, which gave us estimates free of sampling errors for variables in the balance sheet and for the principal variables in the income statement, such as turnover. For these variables it was now possible to calculate reliable estimates of change even for the most detailed level of kind of activity. This was a very significant improvement of the quality and usability of the SBS statistics.

The administrative data was used only for enterprises with less than 50 employees. Enterprises with more than 50 employees were surveyed

through a questionnaire. This resulted in two different sets of variables in the SBS. One very detailed set of variables for enterprises with more than 50 employees based on a questionnaire and one less detailed for enterprises with less than 50 employees based on administrative data.

After a few years, this survey design was once again found inappropriate. The survey was very demanding for both the respondents and for Statistics Sweden. Search began for a survey design, which used the administrative data to a larger extent. Also the 50 employee cut-off was questioned. By having a 50 employee cut-off the coverage rate relating to the detailed set of variables varied considerably between different industries. The estimates for the service sector, which is very dominated by small enterprises, became misleading. As the service sector grew in importance for the economy it became necessary to drop the cut-off in order to produce relevant estimates. For the reference year 2003 we therefore took the third step.

An important objective with the re-design of the SBS was to keep the response burden as low as possible. Therefore a lot of effort was spent on finding out whether or not we could use data from already existing surveys. From the beginning of the project it was clear that the tax data could not serve as the sole data source for the Swedish SBS. Instead we would need to combine this data with other data sources. In the end we decided to use data to some extent from no less than seven different surveys conducted at Statistics Sweden, the advantage of this being of course cost reductions both for the producer and for the respondents. The disadvantages being that we had to settle with for example existing definitions of variables and definitions of populations, which means introducing both measurement errors and non-response errors. In each of these cases we thought that the advantages were greater than the disadvantages.

Despite the fact that we use a number of external data sources it was still not enough to meet all the requirements on the SBS statistics. Therefore we also needed to perform our own data collection on a sample basis. We decided to use a compromise for the solution of our questionnaire. For the biggest enterprises we decided to maintain data collection in spring. For those about 500 enterprises we also decided not to use tax data at all. Instead all the information needed was collected directly from the enterprises.

For all other enterprises we decided to use a later data collection period, from October to December. This gave us the opportunity to use tax data not only in the estimation phase, but also in the design stage. And in addition by this approach we could work with pre-printing of tax-data in the questionnaire and make more efficient sample designs bringing the sample sizes down.

The three sample surveys, which we conduct within the SBS framework differs somewhat with respect to survey design.

The SpecA is a survey on shares and assets. This survey is the smallest in terms of number of selected enterprises. This is due to both that these variables have an extremely skew distribution very few enterprises account for a big share of the population totals and also we only need estimates for those population totals. No breakdowns according to NACE or others are needed. The design for SpecA is stratified random sampling. Stratification is made according to the total amount of the shares and assets, variables which are found in the tax data.

The SpecI survey is a survey on investments. The overall sampling size is about 2 500 firms. We use a cut-off strategy rather than probability sampling meaning that all enterprises which according to the tax data have made investments exceeding about 0.5 € million are included in the sample. For the enterprises beneath the threshold model estimation is used.

Finally we have the SpecRR survey on income and costs. The participating enterprises should provide detailed information on for example turnover by product. We use a nomenclature based on the European CPA-classification for production in the service sector. We use stratification according to activity. Within each of the about 150 strata we use Πps-sampling.

3. Coverage errors in the SBS

Like most surveys the Swedish SBS follows a series of steps while planned and later conducted. Crucial steps in the planning phase are:

a. Defining which variables to use
b. Defining which statistical measures to use to summarise the data
c. Defining the population for the survey
d. Deciding how to gather the necessary information for the survey

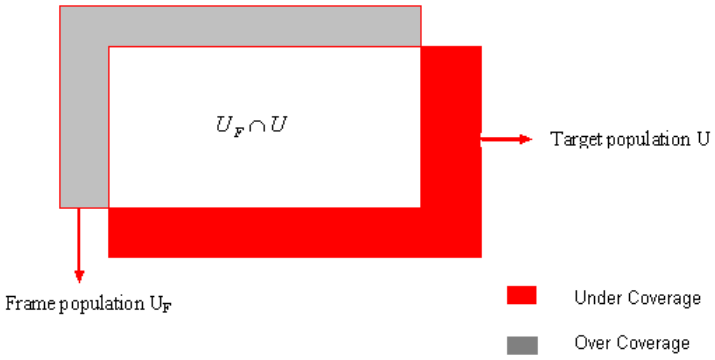
To a very large extent a) and b) are given by demands from the users. In the case of the SBS the principal users of the statistics are Eurostat and the Swedish national accounts.

When it comes to c) the *target* population for the Swedish SBS consists of all enterprises, with economic activity during the reference year. Ideally we would like our statistics to refer exactly to these objects, no extras and none excluded. Unfortunately no register containing exactly these objects is available. We have to settle with a *frame* population constructed from the Swedish business register (BR) and containing all enterprises deemed as active in November of the reference year. Our assumption is that the frame population is a close approximation of the target population.

In order to have coherence between the different surveys within the Swedish economic-statistical system all surveys, with calendar year as reference period, uses the same version of the BR for constructing their frame populations. Therefore it is of utmost importance that the quality of the BR is of as high quality as possible in November. Much work by the unit

responsible for the maintenance of the BR is spent on ensuring this. Despite this we know that we still have problems with both under coverage and over coverage. The coverage problem is graphically illustrated in the following figure.

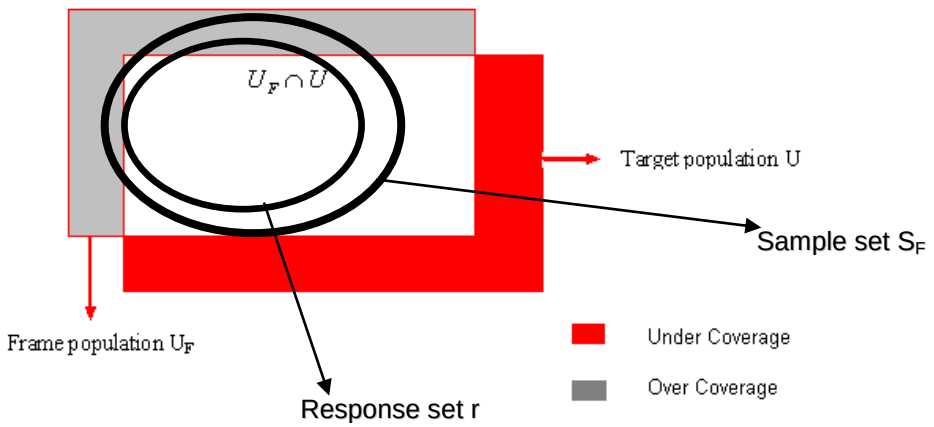
Figure 1. The coverage problem in the SBS



Once we established our population of operation, namely our frame population, at least for this stage of the survey, attention is turned towards point d) how to gather the desired information. This task can be divided into two parts closely linked. First, should we make a census or a sample survey? Second, how do we obtain the data needed, by direct inquiry of the objects of interest or through other sources such as administrative data? In the Swedish SBS we use a combination of methods. The basis of the survey is a census based on tax data. In order to meet all demands on the survey we have added a sample survey, in which specifications of the tax data is asked for through direct inquiries of the selected enterprises.

In a more traditional survey, not using administrative data, one would expect to end up in a situation like the one described in the figure below after the data collection phase.

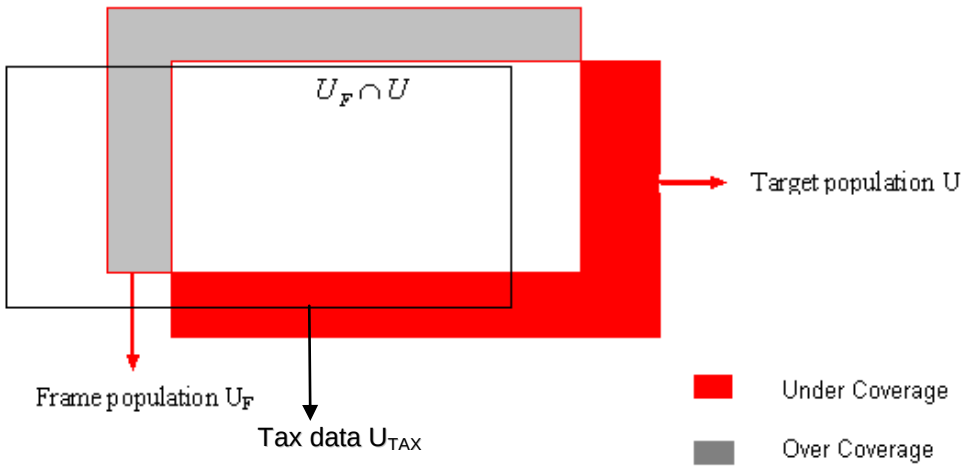
Figure 2. An expected outcome of a survey with non-response and coverage errors



As indicated in the figure we are likely to sample objects from the part of the frame population, which also belongs to the target population $(U_F \cap U)$ as well as from parts falling outside the target population-over coverage (U_F and not U). In addition to coverage errors we can also be

sure of suffering from non-response (indicated by the area between the two ellipses). Therefore in the estimation phase we have to account for both non-response and coverage errors. Two questions in relation to this is often particularly troublesome, how to identify the over coverage in the non-response set and how to compensate for the part of the target population not being subject to the sample, that is the under coverage set. When using the tax data in the SBS the situation becomes somewhat different. First of all no sample is taken, instead we strive to make this part of the SBS as a census for the frame population U_F. Secondly the Swedish Tax Agency, who delivers the tax data to Statistics Sweden, send all the tax records regardless of whether the enterprises in question belongs to the frame population or not. We thus receive data from objects outside the frame population, which is normally not the case in a traditional survey. So instead of the situation in figure 2 the situation for the SBS can be described as follows.

Figure 3. The relation between tax data and target and frame population



As indicated in figure 3 the set of tax data differs from both the frame population and the target population set. The question is how to account for this in the estimation process. There are a number of possible strategies:

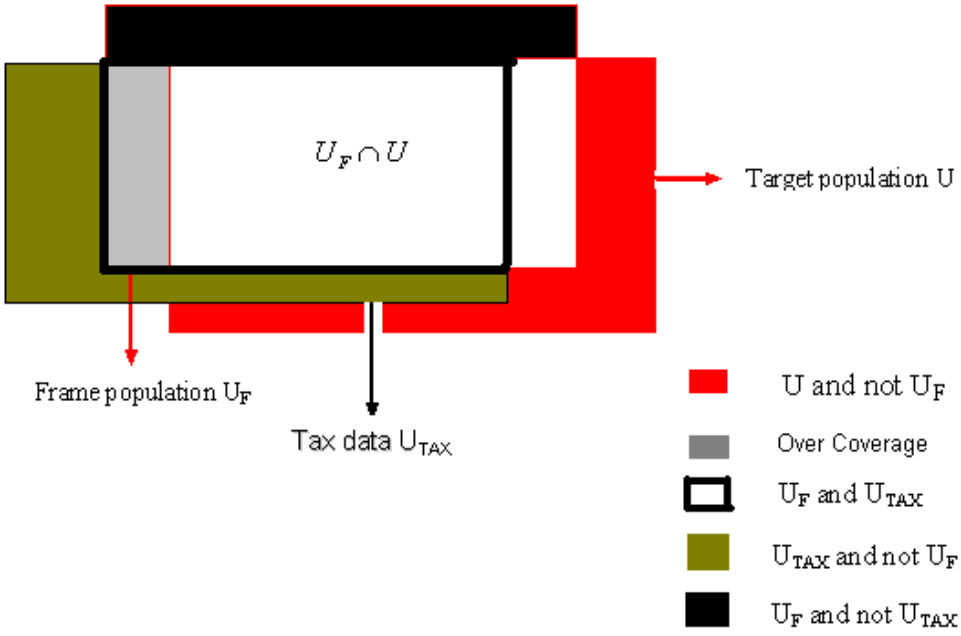
- Strategy 1: Calculate estimates for U_F. Compensate for the non-response in the tax data that is objects belonging to U_F but who are missing in U_{TAX}. No efforts are made to compensate for the coverage errors, which at least resembles estimates for U under the (implicit) assumptions that the under coverage errors and the over coverage errors cancel out.
- Strategy 2: Replace U_F with U_{TAX} and make the estimates relating to U_{TAX} instead of U_F. In this case we do not need any methods to adjust for non-response, since we have full response by definition. No efforts are made to compensate for the coverage errors in the tax data set.
- Strategy 3: Calculate estimates related to the target population U, by using the combined set $U_F \cap U$ including methods for handling non-

response in the tax data and methods for compensating for the under coverage (U and not U_F).

Each of these strategies has its pros and cons. The advantage with strategy 2, as already mentioned, is that no method for non-response adjustment needs to be applied. On the other hand by using this strategy the producer, Statistics Sweden, loses control of the definition of the population. We have to settle with whatever set of objects the Tax Agency supplies data for. This is a big disadvantage when it comes to assessing the quality of the statistics, in particular with respect to the coverage errors. We know from experience that a tax return *alone* is a rather weak indicator of a business being economically active in the SBS sense. Therefore strategy 2 does not seem to be a real alternative.

Strategy 3 seems to be the most appealing from a theoretical point of view, at least when coherence to other surveys within the economic-statistical system is less emphasized. However in order to carry out the calculations associated with strategy 3 flawlessly we need to identify certain subsets exactly, namely: a) $U_F \cap U$, b) U and not U_F. Since we don't have any data file containing members of U this is of course difficult if not impossible in practice.

Figure 4. Different intersections between tax data and target and frame population



What we can do in practice is to determine the subsets $U_F \cap U_{TAX}$, U_F and not U_{TAX} and U_{TAX} and not U_F respectively, which can be of some help. A rather reasonable assumption is then to say that $U_F \cap U_{TAX} \subseteq U$. Thus meaning that if an object in the frame population also occurs in the tax data material this could be taken as strong indication of economic activity in the SBS sense, therefore qualifying for membership in U. The

second subset U_F and not U_{TAX} is more difficult. This subset could in principal be further divided into one part belonging to U, that is U_F and not U_{TAX} and U and another part falling outside U. In the first case we have non-response in the tax data, which needs addressing by appropriate methods, but in the second case we have over coverage and the proper action would then be to exclude these set of objects from the further calculations. The question is how to make the distinction between non-response objects and over coverage objects in practice.

Last but not least in order to carry out strategy 3 we need to determine the under coverage set U and not U_F. One possible way of pursuing this would be to make the following assumption. An object k belongs to the set U and not U_F if and only if it belongs to the set U_{TAX} and not U_F meaning that objects in the tax data not belonging to the frame population coincides exactly with the under coverage set. This is however not true. As mentioned earlier we know for a fact that a tax record alone is not strong indication of economic activity according to SBS definitions and we also know that some objects of interest for the SBS are missing in the tax data material.

The current estimation strategy applied in the Swedish SBS mostly resembles strategy 1. Estimation is done with respect to U_F rather than U. However a small number of objects appearing in the tax data but not in the frame population are added to the population after separate analysis carried out by a subject matter specialist. The objects subjects to analysis are objects, which according to the tax data show signs of significant economic activity. The number of objects dealt with in this way is strongly limited by the amount of time available with the subject matter specialists. Another weakness with this approach is that virtually nothing is done in order to address the over coverage errors.

Since the current method for handling coverage errors in the Swedish SBS, at least in theory, is far from perfect it is of interest to search for a better strategy, more resembling the theoretically more appealing strategy 3.In Sweden we have a number of other administrative data sources, which can be of great assistance in finding the necessary data set when strategy 3 is used. The basic idea is to use all administrative sources available in order to determine whether or not an object is active, i.e. qualified for membership in U.

A part from the tax data we have access to administrative data stemming from the VAT payments, the VAT register, “Gross pay and preliminary tax based on statements of income (LSUM)” and “Gross pay, payroll taxes and preliminary tax from employers monthly tax returns (LAPS)”. The following table shows one way of using the different administrative materials combined in order to determine in which subset an object k is likely to fall.

Table 1. Classification of objects using several administrative materials

Presence in the materials[1]					Classification of the object		
UF	UTAX	VAT	LSUM	LAPS	$U_F \cap U$	U_F and not U	U and not U_F
1	1	0/1	0/1	0/1	1	0	0
1	0	1	0/1	0/1	1	0	0
1	0	0/1	1	0/1	1	0	0
1	0	0/1	0/1	1	1	0	0
1	0	0	0	0	0	1	0
0	1	1	0/1	0/1	0	0	1
0	1	0/1	1	0/1	0	0	1
0	1	0/1	0/1	1	0	0	1
0	0	1	1	1	0	0	1

[1] 1 indicating presence, 0 otherwise

We have applied the technique described in table 1 to each object in the set $U_F \cup U_{TAX} \cup U_{VAT} \cup U_{LSUM} \cup U_{LAPS}$ in order to determine whether each object qualifies for membership in U or not. Aggregated results based on data from 2005 are presented in table 2.

Table 2. An estimation of the coverage errors in SBS 2005

Subset	No. of enterprises	Turnover, SEK billion
Population used for actual SBS estimates	824 389	5 646
Over coverage acc. to table 1 (U_F and not U)	61 191	28
Under coverage acc. to table 1 (U and not U_F)	105 813	42

If we subtract the over coverage and then add the under coverage we end up with the following adjusted estimate for the total turnover: 5 646-28+42=5 660 SEK billion, an increase by 14 SEK billion or 0.25% compared to the original estimate. For the time being this procedure only provides an estimate of the coverage error. Still it can be valuable when it comes to describing the quality of the statistics. In order to perform this adjustment in the real production of SBS statistics one would need to solve some remaining issues. The most important being how to establish values for the important domain indicators such as for example the NACE code for each unit in the under coverage set. But despite remaining obstacles combining different administrative sources could be one way of assessing the issue of coverage errors.

In the long run what one should strive for is to use the administrative data more efficient in the maintenance of the BR, so that the quality of the sampling frame improves, which would be beneficial for all surveys within the economical-statistical system. Such work has just been started at the unit responsible for the BR at Statistics Sweden.

4. Dealing with measurement errors in tax data - two examples

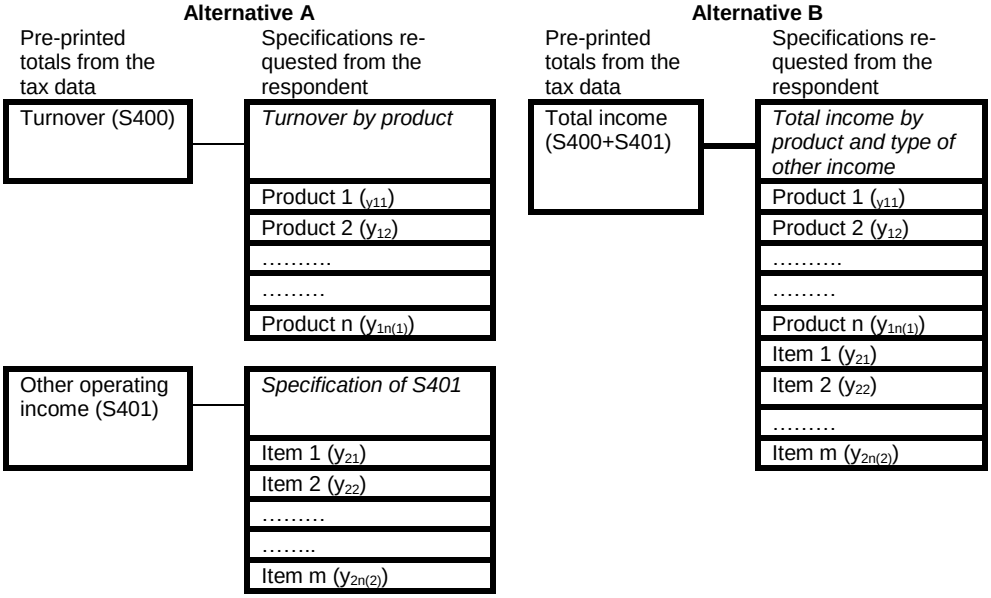
Using administrative data in the production of statistics has many advantages. It is highly cost-efficient, in the sense that it allows us to obtain data

from a large number of enterprises at a fraction of the cost, which would have been associated with collecting the same information directly and individually from each enterprise. It is also beneficial for the respondents, who only need to provide the information to one authority, thus reducing the respondent burden. However there are some also drawbacks that come along with using administrative data. As a producer of statistics we have to settle with the set of information the authority responsible for the administrative data, provides us with, both in terms of which objects information is delivered as discussed in the previous section and the definition and quality of the variables provided.

In this section we will give two examples how we work with measurement errors associated with the use of tax data in our SBS. Our first example illustrates that it sometimes can be wise to refrain from using all the details in the tax data. In the second example we show how we use information about the NACE code, available in the BR, when editing the administrative data.

One important part of the SBS is statistics concerning the incomes and costs of enterprises. The tax data contains the variable "Total income", but it also contains the two sub-components "Turnover" and "Other operating income", each with its own interest in the SBS. In addition to this division we also need to estimate for example "Turnover" by product. Since this information is not available in the tax data we collect this information through a sample survey. In order to keep sample sizes at a minimum we await the arrival of the tax data, so we can use it in the sample selection. This also gives us the opportunity to pre-print tax data-values in the questionnaires before sending them out to the respondents. Two possibilities when it comes to the income part of the questionnaire are illustrated in figure 5 below.

Figure 5. Two alternative ways of pre-printing data in the questionnaires



In alternative A the values from the tax data concerning turnover (S400) and other operating income (S401) are pre-printed separately and the respondent are requested to make further specifications of each of these items. In alternative B however the respondent is requested to give a specification of the pre-printed value of total income (S400+S401). Using the A alternative means, to a higher extent than with alternative B, trusting the division of the total income into its components turnover and other operating income that the respondents make when reporting to the Tax Agency. If, indeed, the division of the total income as given to the Tax Agency coincides with the definitions of the corresponding SBS variables when we would of course be better of using alternative A since it wouldn't be necessary to base our estimations of turnover and other operating income on the sample. Therefore those estimates wouldn't be affected by any sampling error. If however the division of the total income in the tax data doesn't coincide with the SBS definitions we risk ending up with biased estimates of turnover and other operating income when using alternative A. In most industries our experience is that the definitions in the tax data do coincide with the SBS defintions. There are exceptions however especially in the service sector. Therefore we apply alternative B in certain industries. Applying alternative B in the questionnaire still give the opportunity to make estimation as if trusting the division of total income in the tax data. This way we can compare the two resulting sets of estimates.

The estimators for the two alternatives are given by (variable Turnover) and domain (industry) g :

$$\text{Estimator A}: \hat{\mathrm{T}}_{400\mathrm{gA}} = \sum_{k \in Ug} S400_k$$

$$\text{Estimator B}: \hat{\mathrm{T}}_{400\mathrm{gB}} = \frac{\sum_{k \in rg} \sum_{p=1}^{\mathrm{n(1)}} \frac{y_{1pk}}{\pi_k}}{\sum_{k \in rg} \sum_{i=1}^{2} \sum_{p=1}^{\mathrm{n(i)}} \frac{y_{ipk}}{\pi_k}}$$

In table 3, a comparison between the results for the two estimators applied to data from the SBS 2005 is given.

Table 3. The effects of using estimator B, SEK million

Industry	Turnover		Other operating income	
(NACE Rev. 1.1)	Estimator A	Estimator B	Estimator A	Estimator B
80.10 Primary education	8 945	9 502	825	267
80.42 Adult education	10 110	10 103	268	275
85.32 Social work activities without accomodation	7 104	7 794	748	58

Source: SBS 2005

Some differences in the estimates of the distribution of total income can be found in NACE 8010 and 8532. In NACE 8042 virtually no differ-

ence in the two estimates can be found. This being an indicator that for this industry we could use the distribution obtained from the tax data, that is we could switch from alternative B to alternative A, whereas for the two other industries we should hold on to the more restricted use of the tax data. Otherwise we risk ending up with biased estimates for the distribution of the total income.

Another way of evaluating the quality of the tax data, with respect to measurement errors, is to compare the tax data with data from other sources. Since we have a number of other administrative sources, as described earlier in the paper, there are many possibilities. In this example we show how we use information from our BR as auxiliary information in the editing process.

The variable "Total costs" (TC) is divided into a number of components in the tax data. The three most important are "Costs for purchasing goods for resale, raw materials and consumables" (S500), "Other external costs" (S530) and "Personnel costs" (S0128). The distribution of the total costs for an enterprise is of course to some extent individual, but it is reasonable to assume that enterprises operating in the same kind of industry (NACE) are more alike than enterprises operating in different industries. The following table shows how the ratio (called P500) between S500 and TC varies between three industries.

Table 4. Ratio (P500) between S500 and TC

NACE Rev 1.1	No. of enterprises	P500		
		1st quartile (Q1)	Median (Q2)	3rd quartile (Q3)
28 Manufacture of fabricated metal products	10 535	.16	.36	.50
52 Retail trade	55 613	.47	.64	.76
80 Education	11 957	.13	.26	.27

Source: SBS 2005

The three industries in table 4 are quite representative for their kinds of industries (manufacturing, trade and service) in terms of the ratio p500. Trade business generally and not very surprising spend a large portion of the total costs on purchasing goods, whereas business in the service sector spend significantly smaller parts of their total costs on goods. Finally within the manufacturing industries the ratio p500 tends to vary more than it does in the other industries.

The ratio p500 is a very useful tool when editing the tax data. We use the individual value of p500 for a business to see whether it falls within a certain interval or not. The interval is based on the distribution of p500 for the industry in which the business operates according to the BR. Business whose value fall outside the interval is checked further manually. In this process it is not un-common that the values in the tax data is deemed to be correct and that the cause of error is a change of activity or even a misclassification error in the BR.

Finally some end remarks. In the Swedish SBS we have used administrative data (mainly tax data) since the mid 1990s. This has meant a vast improvement to the quality of the SBS statistics, but it has also been a necessity in order to keep the response burden at a reasonable level. Despite the improvement in the quality, there are remaining problems, some of which we have brought up here. We still lack a fully satisfactory way of estimating the consequences (non-response errors and measurement errors), of using administrative data to the total quality of the statistics.