

**UNITED NATIONS STATISTICAL COMMISSION and
ECONOMIC COMMISSION FOR EUROPE**

CONFERENCE OF EUROPEAN STATISTICIANS

Work Session on Statistical Data Editing
(Bonn, Germany, 25-27 September 2006)

Topic (ii): Editing data from multiple sources

REGISTER-BASED ECONOMIC STATISTICS ON ENTERPRISES – EDITING ISSUES

Supporting Paper

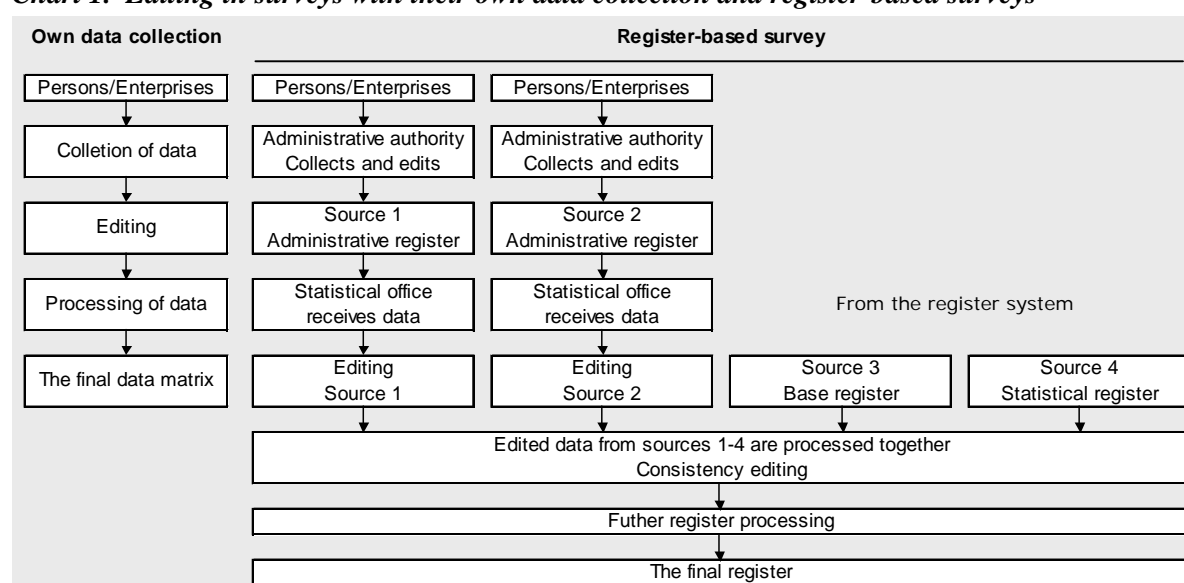
Prepared by Anders Wallgren and Britt Wallgren, Statistics Sweden

I. INTRODUCTION

1. Laan (2002) describes the discussions within Statistics Netherlands concerning a Social Statistics Database (SSD) and an Economic Statistics Database (ESD). In this paper we will shortly describe the methods and editing issues behind the corresponding databases at Statistics Sweden. We distinguish between three kinds of surveys: Sample surveys, censuses and register-based surveys. Sample surveys and censuses are based on statistical data collection. Register-based surveys are based on administrative sources. The methods for such surveys are discussed in Wallgren (2007). The SSD and ESD consist of data from all kinds of surveys, but a large part consists of data from register-based surveys.

2. The editing phase for surveys with their own data collection involves editing of the collected data. It is possible to contact the data providers to correct unreasonable variable values. Errors or suspected errors are interpreted as errors concerning variable values. The aim is to replace wrong or unreasonable values with corrected or reasonable imputed values. With register-based surveys, the data have first been edited by the administrative authority. Every administrative source is also edited when the data has been delivered to the statistical office. After that, data from many sources are edited together. By this *consistency editing* it is possible to find further errors and inconsistencies.

Chart 1. Editing in surveys with their own data collection and register-based surveys

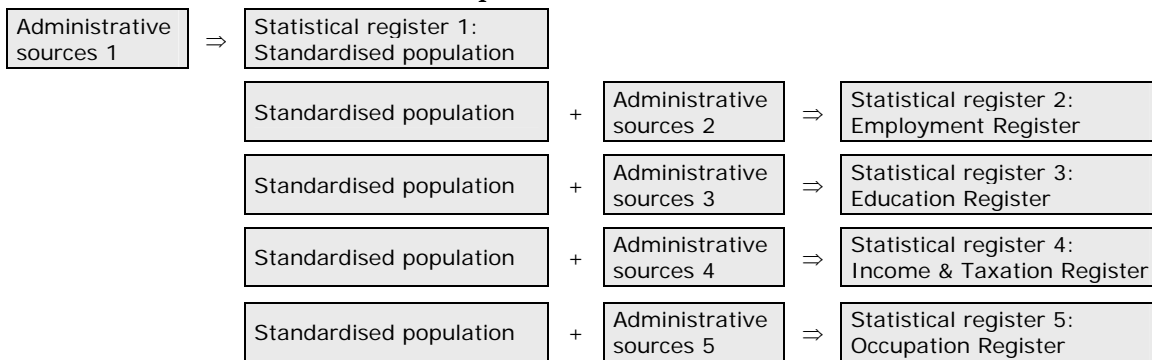


3. In traditional editing of data where the statistical office has collected the data, all errors and suspected errors are interpreted as *errors in variables*. In consistency editing of data in a register-based survey we are editing data from different sources, and suspected errors can both be caused by errors in variables and *errors in objects*. Errors in objects means that we believe that we compare data concerning the *same* object from different sources, however, the data we compare concern *different* objects, which erroneously have the same identity. This will be the case when we get false hits after matching or when we have not created derived objects in a correct way. When we may have errors in objects we should not correct or impute variables values until we have checked that the objects are the same.

II. THE SSD AT STATISTICS SWEDEN

4. The methods used, when registers are created, constitute the most important part of the estimation methods used in register-based surveys. This is discussed in Wallgren (2007). The SSD at Statistics Sweden is created by an estimation process, which results in completely consistent and coherent micro data. The first step in the process implies that a standardized population is created by the team responsible for the Population Register. This population is then the basis for those who work with the other registers. The SSD consists of different parts or registers which are created by other teams at Statistics Sweden. The chart below illustrates the work behind the SSD.

Chart 2. Decentralized but coordinated processes to create the SSD



5. The standardised population is defined as the population of persons at December 31. The population for December 31, year t is created in early February year $t+1$. This standardised population is used as register population in the other statistical registers in chart 2. As the administrative sources 2 – 5 do not overlap regarding statistical variables, the work to create statistical register 2 – 5 can be done independently of each other. The five registers in chart 2 can be integrated into one register with all variables as shown in chart 3. With this SSD it is possible to produce statistics on persons that is completely consistent regarding population and variables.

Chart 3. Integrated register with parts from the Social Statistics Database, SSD

| PIN | Classification variables from the Population Register | | | | | Statistical variables from the ... | | | | | | | | | | | |
|-----|---|------|------|-------|-------|------------------------------------|-------|-------|---------------|-------|-------|-------------------|-------|-------|----------------|-------|-------|
| | var1 | var2 | var3 | var.. | var.. | Employment Reg | | | Education Reg | | | Income & Taxation | | | Occupation Reg | | |
| | var1 | var2 | var3 | var.. | var.. | var.. | var.. | var.. | var.. | var.. | var.. | var.. | var.. | var.. | var.. | var.. | var.. |
| 1 | | | | | | | | | | | | | | | | | |
| 2 | | | | | | | | | | | | | | | | | |
| ... | | | | | | | | | | | | | | | | | |
| N | | | | | | | | | | | | | | | | | |

6. Each source is edited separately, as the variables in different registers do not overlap, there is no need for consistency editing of all sources together. As the object type *person* is stable over time the risk of object errors is comparably small. However, there is risk for object errors due to mismatch as the same person can have different identity numbers (PIN) over time, and also that two persons can have the same PIN. This will give rise to quality problems in longitudinal versions of the SSD. Within the Swedish

Population Register one important task is to keep track of these changes and doublets regarding PIN and edit the PIN variable to reduce this mismatch.

III. CREATING AN YEARLY ESD

7. There are three important differences between the SSD and the ESD:

- First, the population must be defined as the *calendar year population*, i.e. the population of enterprises active somewhere during year t ; during the whole year or during a part of the year.
- Secondly, the administrative sources are often overlapping regarding variables, thus contradicting variable values will give rise to inconsistencies in the unedited microdata.
- Thirdly, the statistical units in the register are also complicated – the administrative units in the administrative sources must be replaced by derived statistical units in the calendar year register.

8. To produce the yearly National Accounts, consistency between different surveys is very important. The calendar year population above is also the target population of the enterprise part of the National Accounts. These needs of the National Accounts should determine the design of the ESD. We will illustrate the methodological problems with data from six sources, one survey where data is collected by Statistics Sweden and five administrative sources, all regarding the calendar year 2002. The table below shows the number of legal units (LU) in these six sources.

Chart 4. Six sources of an Economic Statistics Database, ESD

| Register regarding 2002 | | Number of legal units, LU |
|-------------------------|--|---------------------------|
| Administrative source 1 | Present Business Register, calendar year version | 1 399 403 |
| Administrative source 2 | Enterprises, yearly income declarations | 1 065 516 |
| Administrative source 3 | VAT declarations | 823 144 |
| Administrative source 4 | Yearly wage sums from statements of earnings | 304 181 |
| Administrative source 5 | Monthly wage sums | 300 313 |
| Census, data collected | Business Structure Survey | 770 116 |

9. In this paper we will use only two variables from these sources. For each legal unit we will use *yearly turnover* (administrative sources 2, 3 and census with collected data in chart 4) and *yearly wage sum* (administrative sources 2, 4, 5 and census with collected data in chart 4).

IV. CREATING THE CALENDAR YEAR POPULATION – COVERAGE PROBLEMS

10. Today, the collected data for enterprises is based on a frame population created during November. This frame consists of enterprises active during November according to the administrative source 1 from the National Tax Board. For the purpose of our analysis a special calendar year version of the Business Register was created with data from administrative source 1. The chart below shows the undercoverage in the calendar year version of the Business Register, i.e. the number of legal units missing in spite of the fact that the enterprises have reported to the National Tax Board.

Chart 5. Undercoverage in the calendar year version of the Business Register (BR) 2002

| Number of legal units | Adm source 2 | Adm source 3 | Adm source 4 | Adm source 5 |
|------------------------------|------------------|----------------|----------------|----------------|
| Not in the Business Register | 214 882 | 99 304 | 15 103 | 13 197 |
| In the Business Register | 850 634 | 723 840 | 289 078 | 287 116 |
| Total | 1 065 516 | 823 144 | 304 181 | 300 313 |

11. The main reason of this undercoverage is that the Business Register is based on administrative source 1, which is early available. This is motivated by the fact that the register is used as a frame for data collection. The November version of the Business Register is used when questionnaires are sent out during January the next year. The administrative source 2 (income declarations) regarding 2002 was available at the end of 2003; source 3 was available during spring 2003 for some enterprises but for small enterprises at the end of 2003. Sources 4 and 5 were available during spring 2003. This means that the sources 2 – 5 were not available when the frame was created during November 2002.

12. The undercoverage consists of small enterprises and enterprises that existed during a short period of time, e.g. bankrupt's estates and deceased's estates. These enterprises are in many cases small, but as they are many they will contribute substantially to economic statistics.

13. Our conclusion is that the calendar year version of the Business Register cannot be based on administrative source 1. This source is the National Tax Board's register with all enterprises (legal units) registered for VAT and/or as employers. Sources 2 – 5 give information about all enterprises regarding *actual* turnover and wage sums. The sources 2 – 5 are thus much more reliable than source 1.

14. The calendar version of the Business Register should instead be based on all administrative sources 1 – 5 plus some other administrative sources regarding small enterprises, which are overlooked by Statistics Sweden today. This new kind of calendar year register for year t can be created at the end of year $t+1$. Even if this register cannot be used as sampling frame, it can be used as register population for all yearly enterprise surveys. Work with the yearly National Accounts regarding year t starts at the end of year $t+1$, this means that this kind of register population will also be suitable for the National Accounts.

15. The first part of the ESD has now been created and it is based on all administrative sources. The Business Register's calendar year version with important classification variables will then be used as register population for all other parts of the ESD.

Chart 6. Calendar year population

| BIN | Classification variables from the Business Register | | |
|-----|---|--------|--------|
| | NACE | Sector | Region |
| 1 | | | |
| 2 | | | |
| ... | | | |
| N | | | |

V. CREATING THE VARIABLES IN THE ESD

16. The final ESD should contain all important enterprise variables used for the yearly National Accounts. Turnover, production, value added, export and import, investments, wage sums and number of employed should be included. However, the administrative sources are overlapping regarding these variables, which imply that the work with these sources cannot be decentralized in the same way as the work with creating the SSD.

17. After that the calendar year population has been created, we are planning to create the ESD in the following four steps:

First. *Creating preliminary statistical registers for each administrative source.*

E.g. the administrative VAT-register should be transformed into a statistical VAT-register. This transformation will consist of editing of identifying variables, creating the enterprise units which are used in the calendar population, editing of VAT-variables and checking for doublets. This phase can be decentralized if it is desirable to involve subject matter expertise of different kinds.

Second. *Creating an integrated register for consistency editing.*

All overlapping variables from the primary statistical registers created during the first step above are edited together. The editing process consists of two parts, corrections due to wrong enterprise units and corrections due to wrong variables values. Enterprise units consist of one or many legal units. The grouping of legal units into complex enterprise units is a difficult but important step in the creation of the ESD.

Third. *Creating final statistical registers for each administrative source.*

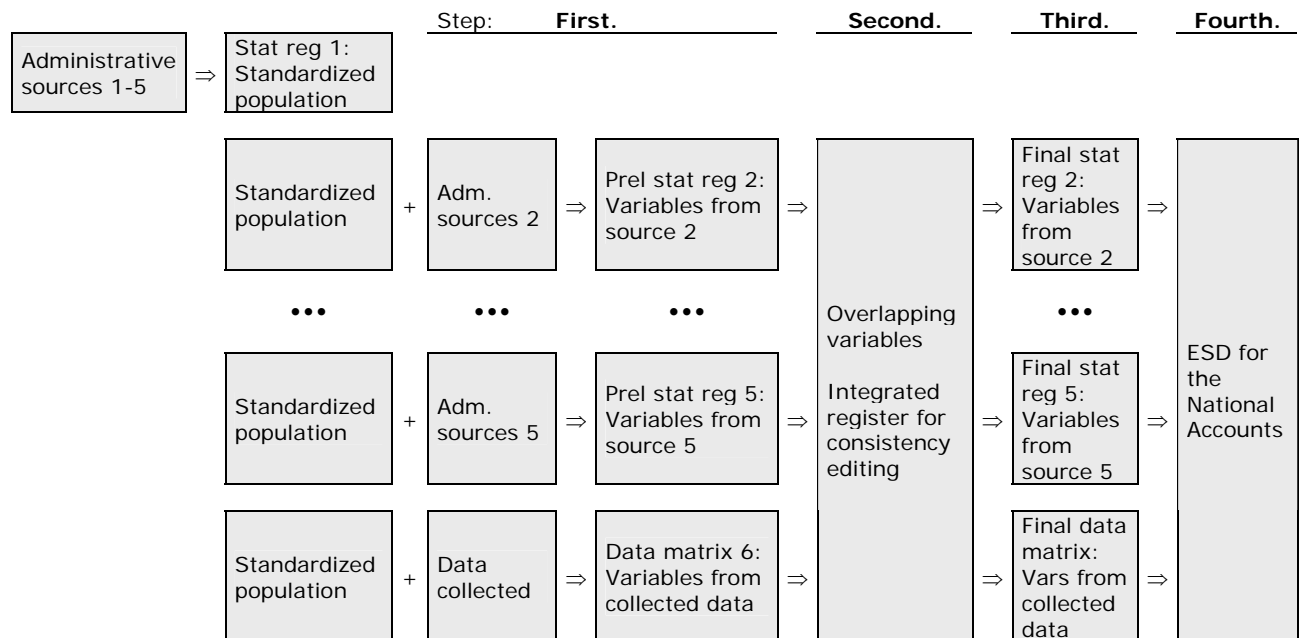
When this integrated register has been edited, the preliminary statistical registers during the first step above are corrected. These final statistical registers are then the consistent input to the ESD.

Fourth. *Creating an integrated register for the yearly National Accounts.*

Those parts of the primary registers above, which are of interest to the National Accounts, are imported to the calendar population register and derived variables are created. The final registers can also be used for other applications, and the statistics produced for these applications will be consistent with the statistics produced for the National Accounts.

18. The chart below illustrates the planned work behind the Economic Statistics Database, ESD:

Chart 7. Coordinated processes to create the Economic Statistics Database, ESD



VI. QUALITY AND EDITING ISSUES IN THE ESD

19. In section IV above it is shown that serious quality issues are related with the present Business Register. Undercoverage gives rise to inconsistencies and underestimation in the National Accounts, but this issue is not discussed in this paper. When overlapping economic variables from different sources are compared, more quality problems become evident. The charts below give an indication of these problems. The tables in the chart are based on data after the first step in chart 7 above. The overlapping variables from different sources are compared before the next step of consistency editing.

20. In the BSS all large enterprises get a questionnaire, and in table 1 the answers to these are compared with the information in the administrative sources. Table 3 shows enterprises with imputed values in the BSS. For all other enterprises in the frame created during November 2001, administrative data from source 2 are used in the BSS, the information for these enterprises are shown in table 2.

Chart 8. Turnover and wage sums in different sources for year 2002, billions of SEK

| Table 1. BSS questionnaire | | | | | | | | | |
|-----------------------------------|-----------------------------|----------------|----------------|-----------------------------|--------------|--------------|--------------|-----------------------|--|
| | Turnover 2002, billions SEK | | | Wage sum 2002, billions SEK | | | | Number of Legal units | |
| | BSS | Source 2 | Source 3 | BSS | Source 2 | Source 4 | Source 5 | | |
| Total | 2 758.7 | 2 849.4 | 3 088.2 | 360.6 | 347.5 | 359.1 | 358.7 | 5 048 | |

| Table 2. BSS uses source 2 | | | | | | | | | |
|-----------------------------------|--------------|----------------|----------------|----------------|--------------|--------------|--------------|-----------------------|--------------|
| | BSS | Source 2 | Source 3 | BSS | Source 2 | Source 4 | Source 5 | Number of Legal units | |
| | Total | 2 003.8 | 2 001.0 | 1 901.3 | 258.4 | 252.7 | 253.3 | | 250.8 |

| Table 3. BSS imputed values | | | | | | | | | |
|------------------------------------|--------------|--------------|-------------|-------------|-------------|------------|-------------|-----------------------|-------------|
| | BSS | Source 2 | Source 3 | BSS | Source 2 | Source 4 | Source 5 | Number of Legal units | |
| | Total | 123.8 | 26.7 | 82.6 | 12.4 | 3.5 | 10.8 | | 10.5 |

21. Comments to chart 8: The BSS in table 1 seems to underestimate turnover. Wage sums in table 2 differ – the BSS seems to overestimate. In most cases the legal units are used as enterprise units, however, 668 legal units in table 1 are grouped into 53 complex enterprise units. The differences between BSS and source 2 in table 2 above (2003.8 – 2001.0 and 258.4 – 252.7) are caused by the editing methods used when editing source 2. Table 3 shows that the present method used for imputation is not good. When values in source 2 are unreliable, values are imputed, but instead it would be better to use the more reliable sources 3, 4 and 5.

VII. ENTERPRISE UNITS AND ERRORS IN OBJECTS – CONSISTENCY EDITING

22. The legal units belonging to the 53 complex enterprise units in the BSS were matched with the administrative sources 2 – 5. One of these 53 enterprise units is shown in chart 9. The BSS values are in this case consistent with all administrative sources. There are no indications of either errors in variables or errors in objects.

Chart 9. Consistency editing of an enterprise unit with consistent data, values in fictive currency

| | | Turnover | | | Wage sum | | | |
|------------|--------------|--------------|------------|--------------|------------|------------|------------|------------|
| | | BSS | Source 2 | Source 3 | BSS | Source 2 | Source 4 | Source 5 |
| EU1 | LeU1 | 1 000 | 997 | 36 | 124 | 124 | 16 | 16 |
| EU1 | LeU2 | | | 702 | | | 69 | 69 |
| EU1 | LeU3 | | | 307 | | | 35 | 35 |
| EU1 | LeU4 | | | 5 | | | 4 | 4 |
| EU1 | Total | 1 000 | 997 | 1 050 | 124 | 124 | 124 | 124 |

23. Another complex enterprise unit in the BSS is shown in chart 10. There are inconsistencies between BSS and source 2 and 3 regarding turnover and inconsistencies between BSS and source 2, 4 and 5 regarding wage sum and a small difference between source 2 and sources 4 and 5. The inconsistencies between the administrative sources can be explained as an error in the definition of enterprise unit EU2. We suspect that at least one more legal unit should belong to the group with administrative data from source 2. This is what we call a probable *error in object*. The BSS values also seem to be too small, the questionnaire has probably failed to give correct measurements.

Chart 10. Consistency editing of an enterprise unit with inconsistent data, values in fictive currency

| | | Turnover | | | Wage sum | | | |
|------------|--------------|--------------|--------------|--------------|-----------|-----------|-----------|-----------|
| | | BSS | Source 2 | Source 3 | BSS | Source 2 | Source 4 | Source 5 |
| EU2 | LeU5 | 1 000 | 1 000 | 1 180 | 65 | 14 | 16 | 16 |
| EU2 | LeU6 | | 12 | 12 | | 5 | 5 | 5 |
| EU2 | LeU7 | | 72 | 1 300 | | 28 | 29 | 29 |
| EU2 | LeU8 | | 317 | 328 | | 0 | 0 | 0 |
| EU2 | LeU9 | | 9 | 9 | | 3 | 3 | 3 |
| EU2 | LeU10 | | 96 | 95 | | 36 | 37 | 37 |
| EU2 | Total | 1 000 | 1 506 | 2 924 | 65 | 86 | 90 | 90 |

Missing legal unit?:

| | | | | | | | | |
|-----|-------|--|-------|---|--|---|---|---|
| EU2 | LeU11 | | 1 400 | 0 | | 4 | 0 | 0 |
|-----|-------|--|-------|---|--|---|---|---|

24. We have tested methods of consistency editing for three subpopulations:

- *Population 1A*: The 53 complex enterprise units used in the BSS
- *Population 1B*: The 4 380 legal units belonging to table 1 in chart 8 but with the 668 legal units belonging to the 53 complex enterprise units excluded. This group should consist of larger enterprises as compared with the subpopulation below.
- *Population 2*: 201 423 of the legal units belonging to table 2 in chart 8. Only limited companies have been included.

25. The next step in the editing work consists of the transformation of all values for each enterprise unit (population 1A) or each legal unit (population 1B and 2) into index values with the BSS value for each variable = 100. In chart 11 the index values for three units belonging to population 1A are shown.

Chart 11. Consistency editing of population 1A

| Reference to chart 9 and 10 | Unit Identity | Turnover | | | Wage sum | | | | Indicator |
|-----------------------------|---------------|----------|----------|----------|----------|----------|----------|----------|-----------|
| | | BSS | Source 2 | Source 3 | BSS | Source 2 | Source 4 | Source 5 | |
| EU1 | ... | ... | ... | ... | ... | ... | ... | ... | ... |
| | BIN04 | 100,0 | 99,7 | 105,0 | 100,0 | 100,0 | 100,4 | 100,3 | 1,2 |
| EU2 | ... | ... | ... | ... | ... | ... | ... | ... | ... |
| | BIN32 | 100,0 | 100,0 | 101,4 | 100,0 | 97,8 | 126,7 | 126,4 | 11,4 |
| EU2 | ... | ... | ... | ... | ... | ... | ... | ... | ... |
| | BIN49 | 100,0 | 150,5 | 292,4 | 100,0 | 99,4 | 103,7 | 103,3 | 50,1 |
| | ... | ... | ... | ... | ... | ... | ... | ... | ... |

26. When all index values for one unit are close to 100, all sources contain consistent values. This means that the line in chart 11 with data for unit EU1 (=BIN04) shows data for an object where all sources are consistent with each other. This enterprise unit is also shown in chart 9. The next line in the chart above shows an enterprise unit where sources 4 and 5 differ from the wage sums in the BSS and source 2. The last line shows the enterprise unit also shown in chart 10. All three measures of turnover are inconsistent and also the wage sources are to some extent inconsistent. Both in the case with BIN32 and BIN49 the most plausible interpretation of the inconsistencies are errors in objects as the administrative sources are based on reports from large enterprises with advanced business systems and professionals who are responsible of the reporting to the tax authorities. It should also be noted that each source have been edited by itself with traditional editing methods before the consistency editing takes place which means that errors in variables have already been corrected. Our conclusion is therefore that these two enterprise units have been defined so that the units are too small – further legal units should be included to achieve consistency between sources.

27. To simplify editing, we have calculated an indicator, here illustrated with data from BIN49:

$$\text{Indicator} = (|150,5-100| + |292,4-100| + |99,4-100| + |103,7-100| + |103,3-100|) / 5 = 50,1$$

Small values are interpreted as no sign of errors in variables or errors in objects. We have used values greater than 11 as sign of plausible errors in objects or variables. In population 1A 40 % of the enterprise units have an indicator value greater than 11, in population 1B almost 25 % of the legal units have an indicator value greater than 11 and in population 2 with small enterprises it seems that errors in objects are not usual.

28. Our conclusions from this test with methods for consistency editing are the following:

- The complex enterprise units used in the BSS must be checked through comparisons with data from different administrative sources.
- More legal units must be grouped into complex enterprise units. During 2002 only 668 legal units were combined into complex enterprise units. Our estimate is that about 2 000 legal units should be grouped if you want to compare data from different sources in a Swedish ESD.
- Objects should be improved first, before inconsistencies are interpreted as errors in variables. Administrative sources with information on relations between legal units should be used in this work.

References

Laan, P. van der (2002): *Creating a Social Statistics Database in the Netherlands: Progress and Priorities*. Paper presented for the second Seminar on Strategies for Social and Spatial Statistics, Copenhagen, Denmark, September 2002.

Wallgren, A., Wallgren, B. (2007): *Register-based Statistics – Administrative Data for Statistical Purposes*. Wiley, forthcoming January 2007.