

**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

HANDLING INCONSISTENCIES IN INTEGRATED BUSINESS DATA

Invited Paper

Submitted by Statistics Netherlands¹

Abstract: At Statistics Netherlands we want to increase the use of businesses registers and decrease the use of business surveys. To investigate whether one data source can replace another, the consistency of business registers and surveys is assessed. Target populations, classifications, units, and definitions of variables are harmonized whenever possible. Data sources are then matched to a central business register. A few key variables are considered, such as turnover, purchase value, wages and number of employees. To assess consistency of sources for these variables a top-down approach is used, looking first at inconsistencies in aggregates before looking at individual records. We developed a method to obtain aggregates for incomplete registers. We aim to detect records that have a substantial contribution to inconsistencies at an aggregated level. Score functions are used to detect records with influential inconsistencies.

I. INTRODUCTION

1. In the past ten years a lot of effort has been made at Statistics Netherlands to harmonize business statistics. As a result both SBS and STS are produced by a uniform statistical process. However, there is still a lack of harmony between SBS, STS, and other statistics such as International Trade, Survey of Employment and Earnings, and statistics for groups of enterprises. As a result, figures are published that can be inconsistent with figures already published. Furthermore, the division National Accounts has to make a lot of checks and adjustments when statistics are integrated, both on the macro level and micro level.

2. At Statistics Netherlands the current strategy is to increase the use of businesses registers and decrease the use of business surveys. The past few years several business registers have come available. Due to a statistical law enforced in 2004 Statistics Netherlands is not allowed to collect information via surveys that is also available in registers. There are therefore several projects with the aim to use registers to produce statistics. Because of the pressure on these projects to use registers as soon as possible the quality of these registers is not always assessed in detail. As a result inconsistencies between registers and surveys are either ignored or solved quick and dirty. Projects that integrate surveys and/or registers are also increasing in numbers. Unfortunately, there is a lack of co-ordination between these projects.

3. The division Business Statistics aims to integrate business registers and surveys using a general and sound approach. The ultimate goal is to use businesses registers instead of surveys, whenever possible.

¹ Prepared by Jeffrey Hoogland (jhgd@cbs.nl) & Ilona Verburg.

When business data sources are checked for inconsistencies early in the production process the department of National Accounts will need less time to publish. When causes for inconsistencies are found and solved at the micro level the quality of data will also improve.

4. To integrate data we need an environment for data storage, matching, and manipulations. Since 2003 an Economical Statistical Database (ESD) is under development (Heerschap and Willenborg, 2006). There are three main projects for ESD. First, there is a software development project which aims to deliver a basic version of ESD (ESD Base) in spring 2007. This system is used to store business microdata for statistical units (enterprises and groups of enterprises) and aggregates. Selections and aggregates can be made in order to publish directly from ESD Base. Classifications and metadata are also stored in ESD Base. Second, there is a project to transform all business microdata, metadata and classifications in the current database (Microlab) to ESD Base. Third, two research projects started in 2004. ESD Integration aims to integrate small and medium-sized enterprises, and CONGO aims to integrate large enterprises. This paper covers the goals, design, methodology, and some preliminary results of ESD Integration.

5. The goals of ESD Integration are

- Improvement of transparency and quality of business statistics and registers due to assessment of statistical processes and confrontation of sources;
- Improvement of consistency of data sources by adjusting micro data and weights;
- Improvement of usability of business registers to determine reliable aggregates;
- Decrease of workload for the department of National Accounts by implementing a step-by-step plan for integration that is in accordance with the priorities of National Accounts;
- Determination of necessary functionality for a database system that is an extension of ESD Base. This database system will be used to obtain consistent data and publication figures.

6. The paper is organized as follows. In section II we describe a general strategy for integration of data sources. In section III we describe specific data sources, units, and variables that are integrated in research project ESD Integration. We discuss some causes, consequences and possible solutions for data inconsistencies in section IV. Methodological challenges that we are confronting to obtain consistent data are the topic of section V. Finally, we draw conclusions in section VI.

II. STRATEGY FOR INTEGRATING DATA SOURCES

7. We propose the following step-by-step plan for integration. First, it is important to determine the target population, output variables, branches and periods of interest, and the statistical unit. We can then make a choice from the available data sources. These sources should contain units that can be linked to the statistical unit of interest. For each data source we have to decide for

- the version to be used. There may be different versions, for instance raw and edited tax data. The choice depends on the required quality and timeliness.
- the variables to be used; it is important to scrutinize variable definitions and analyze values of variables; some variables may contain a lot of empty or improper values. Furthermore, the variables should be related to output variables or provide information about the quality of output related variables;
- the records to be used; some records may be outside the target population.

8. The records in the chosen data sources should then be converted to the statistical unit. This may not be possible for all records due to linkage problems. In case of $1:m$ (one record relates to m statistical units) or $n:m$ relations one should decide if and how this transformation is performed. It implies that values of variables have to be divided over units. It is advisable to analyze variables of data sources after unit transformation for validation purposes. For instance, outliers may indicate linkage errors or improper handling of $1:m$ or $n:m$ relations. In section V.A we will discuss some methods for detection of outliers.

9. When data sources have been matched it may be necessary to compute derived variables that are comparable with output variables, according to variable definitions. In the case of subtle differences between variable definitions it may be decided to compare variables anyway. Data analysis, such as regression analysis, will show whether variables have comparable values.
10. For efficiency reasons, obvious mistakes should be removed before computing initial aggregates since this type of editing can be easily automated, but also because they can have a large influence on aggregates. Matched data can be helpful to detect these errors. For instance, thousand errors in structural business statistics (SBS) can be detected by comparing turnover in SBS with VAT-turnover.
11. For an incomplete register or a survey that is not designed for our target population we have to assign (different) weights to records and/or impute records. This methodological challenge is discussed in section V.B. We could use matched registers to obtain necessary information for the weighting and imputation process.
12. We can now compute initial aggregates. When aggregates for comparable variables across sources differ too much they need to be examined. For instance, we have two sources (SBS and VAT) and for small enterprises in publication cell ‘painting and placing glass’ the two aggregates for turnover differ by more than a fixed percentage or significantly.
13. There are several causes for inconsistencies at a macro level, which we will discuss in section IV. Most causes can be related to the micro level. That is, we want to detect inconsistent records that have a substantial influence on aggregates. This methodological challenge is discussed in section V.C. The use of score functions or robust regression analyses are possible approaches.
14. A difficult step is to solve influential inconsistencies in records. There are many possible causes and it may be unclear which source has to be adjusted. We need to have access to the General Business Register (GBR). The GBR is used as the sampling frame for business surveys and it contains monthly information about the building blocks of statistical units. For instance, enterprises consist of one or more legal units. Furthermore, the relation between enterprises and legal units can change every month. A possible cause for an inconsistency between data sources is that they are based on different versions of the GBR. For example, an enterprise in one data source may consist of more legal units than the ‘same’ enterprise in another data source due to a take-over purchase. We also need process metadata for each source, such as remarks of editors, which can be crucial for our confidence in a data source for specific records. Finally, we should be able to get in touch with experts for each source when we cannot explain an inconsistency. The causes, consequences, and solutions for inconsistencies should be stored in a database for efficient handling of records.
15. Solving an inconsistency will result in either
- adapting values of variables for a data source; it is advisable to keep an original version for each data source, such that original aggregates remain available and changes can be undone;
 - remove or add records for a data source;
 - adapting weights for a data source;
 - adapting the classification (NACE or size class) for a specific record;
 - changing the way in which records from different data source are matched.

Aggregates can then be recalculated. When aggregates for comparable variables across sources still differ too much they need to be considered again. When it is unclear which changes have to be made a possible approach to obtain consistent aggregates is to use repeated weighting techniques (Houbiers, 2004). This means that for each aggregate the weights are adjusted in a minimal way according to a certain criterium, such that all restrictions (consistency with comparable aggregates and edit rules) are satisfied.

III. INTEGRATING DATA SOURCES IN PRACTICE

16. For the first phase of ESD Integration we are interested in annual business data sources available on enterprise level for 2001-2004. Our target population is all active enterprises for the branch of industrial manufacture and construction industry. We want to integrate six key variables, which are given in table 1. These variables are available in several sources and they are very important for National Accounts and other publications. Furthermore, we want to obtain consistent data for 2004 at an aggregated level. The aggregates are based on all enterprises within a combination of publication cell (group of NACE) and size group. We consider two size groups, namely small enterprises (less than 10 employees) and medium-sized enterprises (10-99 employees). Statistics Netherlands aims to make large enterprises consistent during observation. That is, in the near future a large enterprise will be asked to fill in all questionnaires at once.

Table 1. Available annual sources on enterprise level for six key variables.

	GBR	VAT	CT	TS	SSD	SBS	SEE	Pc
Number of employed persons	X				X	X	X	
Gross wages and salaries			X	X	X	X	X	
Total labour costs			X			X		
Net turnover		X	X		X	X		
Purchase value			X			X		X
Profit			X			X		

17. At Statistics Netherlands there are several annual business data sources available on enterprise level, see Figure 1. The General Business Register (GBR) serves as a unit base and is used as a population, sampling, and weighting frame for business surveys. This register is quite complete for most business branches (except for farmers and lawyers), but not always up to date. Furthermore, the GBR also contains non-active businesses. The statistical units in the GBR are enterprise and groups of enterprises. The enterprise is the smallest combination of legal units that is an organisational unit producing goods or services, which benefits from a certain degree of autonomy in decision-making, especially for the allocation of its current resources. An enterprise group is an association of enterprises bound together by legal and/or financial links.

18. There are several tax sources available for businesses, namely Value Added Tax (VAT), Corporation Tax (CT), and Tax on Salaries (TS). It is not easy to use these sources for statistics, because the tax department uses a fiscal unit instead of a statistical unit. However, fiscal units can be linked to legal units, which can be linked to enterprises. There are two reasons why tax information can be unavailable on the enterprise level. First, each type of tax form does not to have be filled in by every fiscal unit. Second, unit transformations can be so complex that values on the fiscal level can not be easily transformed to values on the enterprise level. In these cases it is decided so far not to compute tax values on the enterprise level.

19. The Social Statistical Database (SSD) is much further in its development. It has already been used to store and publish coherent and consistent data for several years (Arts and Hoogteijling, 2002; Houbiers, 2004). One of the files in the SSD is the jobs of employees file (JSSD). It contains information on wages and number of working days. A job is a unique combination of a person and an enterprise. It is therefore easy to derive information on wages and number of employees on the enterprise level, which is used for ESD Integration. Figure 1 shows that there can be enterprises in the JSSD for a specific year that do not exist in the GBR for that year. The main reason is that some changes in enterprises resulting in a different enterprise id number are ignored for JSSD. For instance, when an enterprise changes its name the enterprise id number is also changed in the GBR. However, when there is no change in the composition of the employees the expired enterprise id number remains in use for JSSD to obtain comparability over time.

20. The surveys included in the first phase of ESD Integration are the Survey of Employment and Earnings (SEE), Structural Business Statistics (SBS), and Prodcom (Pc). Other annual business surveys that are available on the enterprise level are Gross Fixed Capital Formation statistics (GFCF), Research & Development statistics, and ICT-statistics. The latter describe investments in and use of ICT-means. These surveys are probably included in the second phase in 2007. There is hardly any overlap between variables of GFCF and SBS. However, a great part of the investments in buildings, for instance, imply turnover for the construction industry. Growth for GFCF-variables therefore implies growth for SBS-variables.

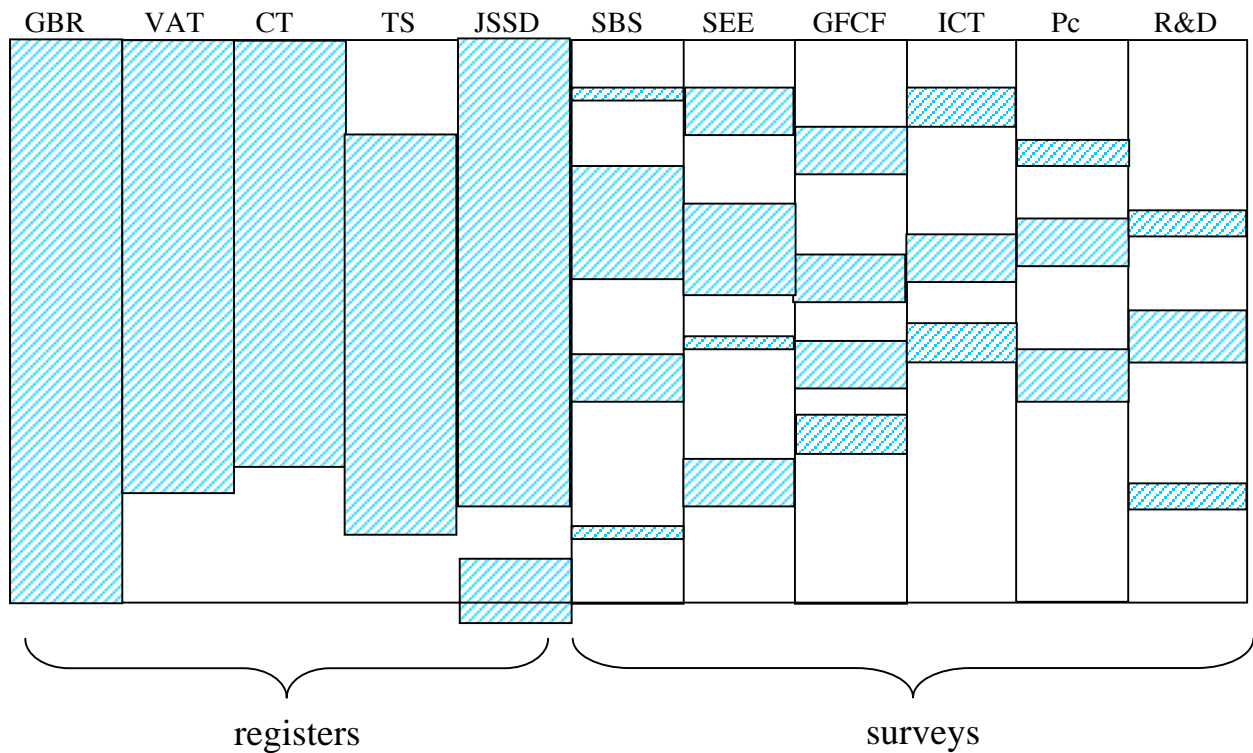


Figure 1. Annual business data sources available on enterprise level.

IV. INCONSISTENCIES BETWEEN DATA SOURCES

A. Causes

21. There are several causes for inconsistencies between data sources at the microlevel, see table 2. They can cause substantial differences between different publication figures regarding the same variable. Problems that we have encountered are, for example, mismatching of tax records with records in surveys, a difference in the date for which a survey questions the number of employees, differences in variable definitions for wage-related variables, and errors in the editing phase.

Table 2. Causes for differences between sources at publication and/or micro level.

Causes for differences at publication level (only)	Causes for differences at publication and micro level
Difference in target population	Matching error
Difference in weights	Difference in variable definition
Classification error	Difference in measurement time (period)
	Measurement errors in variables
	Processing errors in variables, e.g. due to wrong unit transformations
	Difference in editing strategy
	Observed versus imputed value
	Difference in imputation method

B. Consequences and possible solutions

22. Inconsistencies between data sources at the microlevel can cause substantial inconsistencies at an aggregated (publication) level. For instance, it is found that for several enterprises the number of employees according to SBS is very small (after editing), while the number of employees in SSD and SEE is large. It turns out that the difference is caused by the fact that according to SBS the personnel is borrowed from another enterprise. It is difficult to decide which source is right. For SBS more editors and relations between variables are available and there is more time to contact enterprises. However, the editors of SEE in general pay more attention to the variable number of employed persons and shifts of personnel between enterprises. To solve this issue editors from different departments have to be confronted with the differences between sources and the available proces metadata, such as remarks made during the editing phase. This could result in a decision tree which gives the preferred source for each variable in different situations.

23. As long as the reliability of sources, the usability of proces metadata, and differences in variable definitions and editing strategy are unclear, it is a labour-intensive task to produce consistent integrated micro data. We hope that enough insight will be obtained to automatize this process. We could implement the above-mentioned decision tree by programming if-then rules. A disadvantage of this approach is that the decision rules are inflexible and may need a lot of maintainance.

24. We could also follow a more fundamental and mathematical approach to obtain consistent integrated micro data. At Statistics Netherlands there is a lot of experience with automatic editing of business data sources (De Waal, 2000), based on the Fellegi-Holt principle (Fellegi and Holt, 1976). The main difference with integrated data sources is that we also want consistent data between data sources. If inconsistencies occur on the micro level then consistent data can be realized with a set of edits that relates variables from different sources. In figure 2 we give an example for three data sources (A, B, and C) and two variables (x and z) that are comparable across these data sources. We distinguish k cells, where a cell could consist of a combination of NACE and size class of an enterprise. These cells are formed such that they contain enterprises for which variables in the data sources have a certain degree of reliability. A reliability weight represents the degree of confidence we have in a certain variable in a data source. We determine reliability weights r_{cj} for each cell c and variable j . We have five explicit edit rules, which imply several implicit edit rules, such as $x^A = x^C$. For enterprises for which at least one edit is violated values of variables are changed such that all edits are satisfied. We follow the generalized Fellegi-Holt principle, that is, we want to minimize the sum of reliability weights for variables changed. In our example we want to minimize for cell c

$$r_{1x^a} \delta_{x^a} + r_{1y^a} \delta_{y^a} + r_{1x^b} \delta_{x^b} + r_{1z^b} \delta_{z^b} + r_{1x^c} \delta_{x^c} + r_{1z^c} \delta_{z^c},$$

where

$\delta_j = 1$, if variable j is changed

$\delta_j = 0$, otherwise

with the restriction that alle edits must be satisfied.

C. Future goals

25. Within five years we hope to have a general editing system for business registers and surveys, where comparable variables within different data sources are linked. The system should give information about definition of variables, (subtle) differences in definitions, and possible consequences of these differences. Variables regarding the same concept that have differences in definitions that can cause large differences in observed values should be considered as different variables. The system should also be

linked to a database which contains a complete list of possible causes, consequences and solutions for inconsistencies at the micro level.

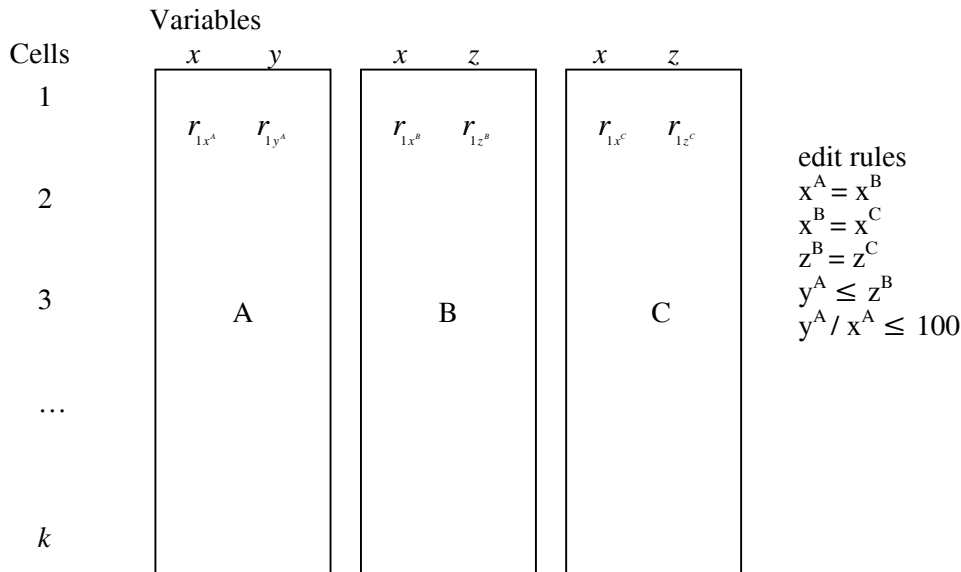


Figure 2. Cells, variables, edit rules and reliability weights for three data sources.

26. To minimize the burden for respondents we should restrict the amount of comparable variables within different data sources. However, this reduces our ability to compare data sources. A trade-off between efficiency and comparability must be found. A general system for storage of microdata is crucial to realize a minimum number of observed variables. The same system could be used for computation and storage of aggregates. Different aggregates for comparable variables may occur. The steps described in paragraphs 13-15 should then be applied. The system should therefore contain information about target populations, classifications, weights, and process metadata for each data source in order to explain and solve differences in aggregates.

V. METHODOLOGICAL CHALLENGES

27. When business data are matched, inconsistencies might occur. These inconsistencies are differences in different data sources, for example for a specific variable that consist in both sources. Therefore, comparison is important to find inconsistencies in one or more sources and to investigate which of the sources can be used as a standard for a specific variable.

28. The target population is all active enterprises in the Netherlands, which are enterprises with turnover or employments for a specific year. The number of active enterprises can be estimated from a unit base register, the GBR. The GBR contains all the enterprises that exist in the Netherlands and is used as unit base register for businesses at Statistics Netherlands.

29. Since the register does not necessarily consist of active businesses only, for each size class and legal form the probability that an enterprise is active is estimated at Statistics Netherlands. Since not all enterprises pass on to Statistics Netherlands when they have become inactive, there is a delay in the GBR. Weights can be used to determine the number of active enterprises in the GBR. Outlier detection is needed to find errors that have a large contribution at the publication level. These errors have to be taken into account while determining raising factors. To estimate population totals weighting is needed for surveys and for incomplete registers

30. To match different registers and surveys, outlier detection for data sources, weighting of incomplete data sources, and detection of influential inconsistencies are of great importance. In this chapter the methodological challenges for these aspects will be discussed.

A. Outlier detection for data sources

31. It is not always the case that outliers are errors; in several cases outliers are correct values. The aim of this paragraph is to discuss methods for outlier detection, for a specific variable, to avoid errors. To compare aggregates for different sources, these aggregates have to be correct. Errors might occur in the raw data, but can also occur after editing. A disadvantage of extreme values is that weighting can lead to disproportionate estimates. This is one of the reasons that outlier detection and editing is needed, before weighting a data source. This raises the question when an error is unacceptable. One has to think about setting a specific threshold value, such that a value is unacceptable if it exceeds the threshold.

32. Outliers may also occur when the data is heterogeneous. This occurs for example when enterprises from different sectors or different sizes are compared. A data source can be divided into different groups, called strata, which are homogeneous groups of data. The idea is to detect errors for these strata and calculate the aggregates, to compare with the aggregates of strata from another data source.

33. Outlier detection can be done within a record and between different times, for one data source, but can also be done between records, for a specific variable. Outliers can also be detected by considering differences in combinations of variables, such as ratios. To find outliers for a variable several methods can be considered, the same holds for finding outliers within records. The dataset and the goal determine which method has to be chosen. It is possible to use methods that find all outliers at the same moment, for example by means of a plausibility indicator (Hoogland, 2002). An advantage of these methods is that they are easy to understand and to apply. A disadvantage is that the raw data has to be considered to conclude whether an outlier is an error or not. Below outlier detection within a record, for several years will be discussed first; afterwards outlier detection for a variable will be discussed.

34. By comparing a record for different time periods extreme values and editing mistakes can be detected. A method is to make use of the variation coefficient, which can be calculated as the quotient of the standard deviation, of the specific record, over the periods and the average value of this record over the periods,

$$F_i = \frac{\frac{1}{n} \sum_{j=1}^n (x_{ij} - \bar{X}_i)^2}{\bar{X}_i}, \text{ with } \bar{X}_i = \frac{1}{n} \sum_{j=1}^n x_{ij}.$$

Here $X_i = (x_{i1}, x_{i2}, \dots, x_{in})$ represent a specific record i , with $i = 1, \dots, m$, m the number of records, and n the number of periods. Enterprises for which the variation coefficient exceeds the specified threshold values are suspected to be outliers.

35. Calculating the so called Mahalanobis distance is another way to find outliers in a record. The Mahalanobis distance is a version of the Euclidian distance that can be computed for different variables simultaneously and takes the covariance between these variables into account. The Mahalanobis distance between $x = (x_1, \dots, x_p)^t$ and $y = (y_1, \dots, y_p)^t$, with $x_i = (x_{i1}, \dots, x_{im})$ and $y_i = (y_{i1}, \dots, y_{im})$ and n, m the number of records is

$$d(x_i, y_i) = \sqrt{(x_i - y_i)^t S^{-1} (x_i - y_i)},$$

where S^{-1} is the inverse of the covariance matrix for $(x - y)$. This method can also be used for comparing variables of different sources. A problem is that a multivariate normal distribution is assumed, which is not the case in the data sources for enterprises. This means that an outlier can have a small Mahalanobis distance and not all outliers will be detected. To solve this problem a robust covariance matrix can be used.

36. By considering differences between records, for the same variable, outliers can be recognized. Enterprises are compared with other enterprises to find outliers. For example when VAT-turnover is considered one can find turnover values in SBS that are disproportionally high. A method to find outliers is based on the average and the standard deviation of enterprises in a stratum. A threshold value will be set to specify when a value in a record will be suspected as an outlier. An idea is to exclude the 10 % records with the highest and lowest value, before applying this method, since these values have a large influence on the average and the standard deviation.

37. Another method that can be used to find outliers is to determine the quartiles (Q_1, Q_3) and median (Q_2) of the records for a variable. When the distance from the first and third quartile is more than three times the inter quartile distance ($Q_3 - Q_1$) a record is suspected as being an outlier. A function F is used to detect outliers. A record can be located as an outlier when $F > 3$.

$$F = \begin{cases} \frac{Q_1 - x_i}{Q_3 - Q_1}, & \text{if } x_i < Q_1 \\ \frac{x_i - Q_3}{Q_3 - Q_1}, & \text{if } x_i > Q_3 \end{cases}$$

38. A method to find outliers, for several variables, is robust regression analysis. This method uses additional information, such as the same variable from another source or a dependent variable from the same source. To find outliers large residuals from a regression analysis can be considered. The least squares method is often used to determine a relationship between two or more variables. However, this method is not robust for outliers existing in the data. An outlier can cause the regression line to be much different than it would have been if there were no outliers. In practice outliers will almost always occur in a business dataset, so that mistakes are unavoidable. To overcome this problem, two things can be done. First, one can remove the outlying points and second, one can use robust regression. In most cases the goal is to assign less weight to an observation which seems to be an outlier. By looking at the residuals, and also at the weights assigned to a record, outliers can much more easily be found than with the least squares method. So robust regression will not be influenced by an outlier.

40. The method of robust regression can be used in the case that information of the same variable from another source is used for outlier detection. For example VAT-turnover is used for outlier detection and weighting of production statistics. In the case of dependency between two variables, such as for example the number of employees and average pay, regression can be used to detect outliers. Another possibility is the use of the ratio of these depended variables when one expects that they differ with a fixed factor. Another idea is to combine more variables to find outliers and inconsistencies. In ESD Integration inconsistency detection can take place by comparing the ratio of two variables, such as sum of wages per employee, for the SEE-survey.

41. Below an example of outlier detection in VAT-data will be discussed. The sector considered in this example is the building industry, in particular demolishing of buildings. We consider size classes 0 to 3, which contain enterprises with 0 to 9 employees as is shown in table 3. The variable of interest is net turnover. The method that is used is the so called boxplot method to find outliers within a variable between different records. The function F is used to select turnover values that have a distance from the first or third quartile that is more than 3 times the inter quartile distance. Below one can find the results for enterprises, which have. The distance of enterprise 14450119 is just small enough for the observation not to be considered an outlier. In this example the median is 118181, the first quartile 55384, and the third quartile 373607. More outliers occur for enterprises with turnover higher than the third quartile. The reason can be that the distribution of enterprises is asymmetric. Most of the enterprises in this group consist of 0 employees, which means that the median is relatively low. Possibly, the considered group in this example is less homogenous as expected and detecting outliers in each size class separately would probably result in less identified outliers.

Table 3. Number of employees of each size class.

Size class group	Size class	Number of employees
Small	0	0
	1	1
	2	2 – 4
Medium	3	5 – 9
	4	10 – 19
	5	20 – 49
Large	6	50 – 99
	7	100 – 199
	8	200 – 499
	9	≥ 500

Table 4. Distance in inter quartile distances for demolishing of buildings.

Enterprise:	Size Class:	Net Turnover:	Distance:	F
14450119	0	1428885	1055278,3	2,82
30156297	2	1611026	1237419,3	3,31
18161553	3	1710296	1336689,3	3,58
10993037	2	1788505	1414898,3	3,79
32594194	2	1800420	1426813,3	3,82
11033797	2	1810914	1437307,3	3,85
31515665	3	1856239	1482632,3	3,97
27378128	3	1899973	1526366,3	4,09
10314539	1	1929560	1555953,3	4,16
25894382	3	1975163	1601556,3	4,29
34482962	0	2119013	1745406,3	4,67
10741569	3	2176769	1803162,3	4,83
10337725	3	2192395	1818788,3	4,87
22063412	2	2294031	1920424,3	5,14
21703701	1	2407107	2033500,3	5,44
32416938	0	2637860	2264253,3	6,06
10963995	3	2650672	2277065,3	6,09
11912278	3	2785782	2412175,3	6,46
11333553	3	2968658	2595051,3	6,95
10325824	3	3577336	3203729,3	8,58

42. The boxplots in figure 3 are a result of taking the four size classes separately. A conclusion of this figure can be that more outliers (stars) occur in the smallest size class with 0 employees and not in size class 3, which one can conclude from table 4. A reason can be that enterprises that have started recently are first placed in size class 0, while several of these are possible larger.

B. Weighting of incomplete sources

43. In order to compare aggregates some surveys and registers have to be weighted. It is of importance that all data sources relate to the same target population. For example corporation tax only concern legal persons, while other sources, such as VAT, also concern natural persons. Corporation tax records therefore have to be weighted. Surveys exist of a (random) sample of a target population that may be smaller than the target population of interest. In these cases the weights for records in a survey have to be adjusted. Furthermore, for both surveys and registers non-response and missing values may occur. Weighting and possibly also imputation have to take place in order to estimate aggregates for the target population of interest. During the determination of weights, some outliers will obtain a weight equal to 1. They are expected to be unique for the target population. The aim of this subsection is to discuss methods for weighting and to discuss the difference between weighting surveys and registers.

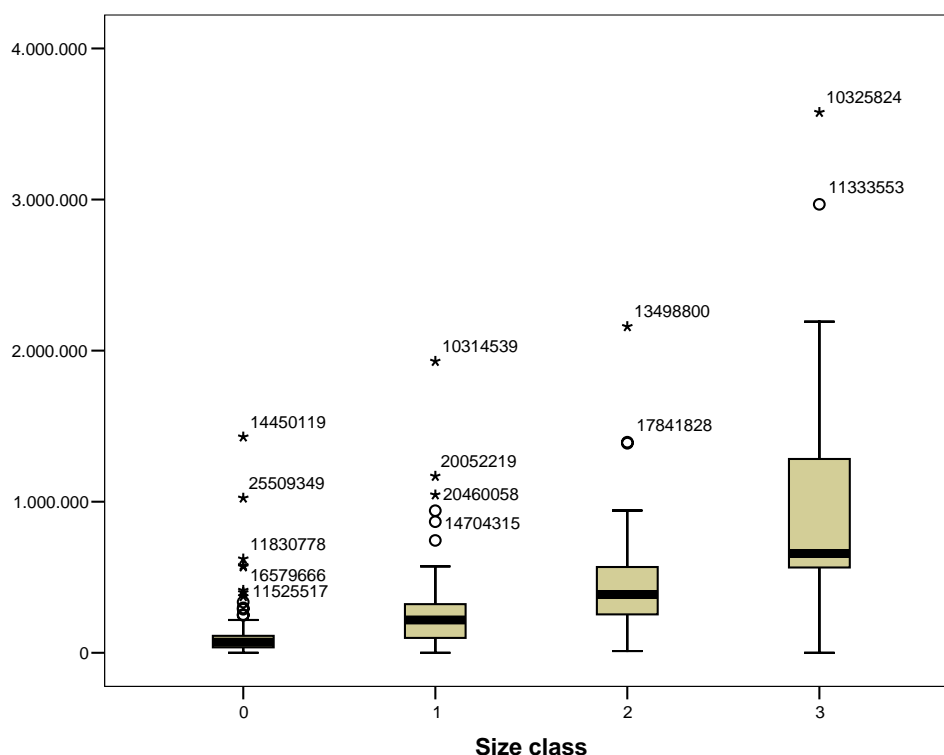


Figure 3. Boxplots for net turnover, for publication cell 'demolishing of buildings' and size classes 0 to 3.

44. The target population can be equal to all active enterprises in the Netherlands, or for example all natural persons in the Netherlands. For ESD Integration the target population is all active enterprises in the Netherlands. We start with the construction industry. The GBR contains all enterprises in the Netherlands. A disadvantage in the use of GBR as a weighting frame is that it does not consist of active enterprises only. Although it is updated every month, most changes cannot be processed at the moment they occur. It is assumed that a change in activity is recorded in the GBR after an enterprise has informed Statistics Netherlands or the Chamber of Commerce. To correct for inactive enterprises, there are correction factors available, called activity probabilities. The number of active enterprises can be estimated using activity probabilities. This can be used, for example, to estimate the total number of employees in the construction industry by means of the GBR.

45. For another example of weighting a register we consider VAT-data obtained from the tax authorities. Information from the tax authorities relate to fiscal units which have to be converted to enterprises. Because of transformation failures a number of enterprises, which are active in the GBR, do not occur in the VAT-data. For other enterprises turnover is unavailable for at least one of the related fiscal units. We weight VAT-records according to all active enterprises in the GBR. This is of great importance for comparing VAT and SBS, since the SBS is also weighted according to all active enterprises in the GBR.

46. We make use of the JSSD-register to weight VAT-records according to the number of active enterprises in the GBR. The reason is that for enterprises in the set JSSDVAT we know that they are active, so net turnover can be estimated. There is probably no reason for further weighting for missing enterprises. The turnover for the enterprises, for which not all data is available, is estimated and

imputation takes place also for these enterprises. Before estimation and imputation of missing VAT-turnover values outlier detection is necessary, because we receive raw data from the tax authorities.

47. To weight VAT-records the enterprises in $GBR \cap (VAT \cup JSSD)$ and the enterprises in $JSSD \setminus VAT$ are determined. Then VAT-turnover per JSSD-employee is calculated for enterprises in $VAT \cap JSSD$, and denoted as

$$F_i = \frac{T_{i,VAT}}{E_{i,SSD}},$$

The average \bar{F} of F_i is calculated and the net turnover is estimated for each enterprise j in $JSSD \setminus VAT$ by multiplying \bar{F} by the number of employees E_j . This is also done for enterprises for which not all VAT-turnover is available. Enterprises for which the turnover per employee is an outlier are not used to calculate \bar{F} . Method that are used to find outliers were described in subsection V.A.

C. Detecting influential inconsistencies at the micro level

48. When combining aggregates influential inconsistencies have to be detected and possibly corrected. Inconsistency detection will be done at micro level; in our case the enterprise level. Only records that influence the aggregates at publication level will be considered in more detail, since small inconsistencies may neutralize each other. A goal of this subsection is to discuss methodology which can be used to detect inconsistencies at the micro level for the same variables in different sources. A second goal is to determine when a difference between aggregates of different sources is unacceptable and which inconsistencies at the company level contribute most to these difference. The focus will be on variables from registers and surveys. The last years more registers became available and a less surveys will be used. A register and a survey that will be compared will have different weights, since a register contains more records than a survey for a specific population.

49. To compare (two) different sources having a measure for the difference between for a specific record i in the different sources is of importance. This measure can be used to draw conclusions about the influence and importance of a specific distance between these records on the total aggregates. A possible measure for comparability of sources is the use of score functions (Groen, 2006). A score function is a monotone function that assigns to every value a certain score. Records with a low score or value will not be of much influence on the publication total. A score function can also tell when a value in a record is probably an error. Score functions do exist in many different forms. The Mahalanobis distance, which is discussed in subsection V.A, and the robust regression method can also be applied to compare variables between data sources. When comparing registers and surveys three separate groups will be distinguished: $A \cap B$, $A \setminus B$, and $B \setminus A$, where A and B are two different sources. In some cases are the aggregates almost the same, while, because of weighting, there is a difference between the individual records. In the case that aggregates are the same there is no further comparison of records.

50. During micro-integration different sources will be compared. At the moment VAT information will be compared with SBS and CT information. We also compare SEE information with SBS and SSD information, table 1 in section III. For this purpose we make use of different score functions. Three score functions, used for enterprises in $A \cap B$, are described below.

51. A score function of Farwell and Raine (2000) can be used for two surveys,

$$SF_1 = \frac{|w_i^A x_i - w_i^B y_i|}{\left| \sum_{i=1}^n w_i^A x_i - \sum_{i=1}^n w_i^B y_i \right|}, \quad (1)$$

This score function determines the difference between x_i and y_i with respect to the difference in aggregates. Here w_i^A and w_i^B are the weights of x_i and y_i respectively, with x_i in A and y_i in B . This score function will be used to detect inconsistencies between comparable variables in different sources.

52. A score function that can be used to compare a survey with a register is

$$SF_2 = \frac{|x_i - y_i| \cdot \max\{w_i^A, w_i^B\}}{\left| \sum_{i=1}^n w_i^A x_i - \sum_{i=1}^n w_i^B y_i \right|}. \quad (2)$$

When a register is complete the weight will be 1 for this register. It is plausible that records of a survey have a higher weight than records of a register. This score function uses the maximum of the two weights to weight the difference between the two records. An advantage is that the real difference of the records x_i and y_i is used to calculate the score.

53. A score function by Hidiroglou en Berthelot (1986) can be used to detect inconsistencies between surveys and registers. Consider the following ratio of x_i and y_i for $x \in A$ and $y \in B$ variables:

$$R_i = \frac{x_i}{y_i}.$$

To detect differences a transformation is needed, such as the transformation S_i ,

$$S_i = \begin{cases} 1 - \frac{R_{med}}{R_i} & 0 < R_i < R_{med} \\ \frac{R_i}{R_{med}} - 1 & R_i \geq R_{med} \end{cases} \quad (3)$$

In S_i is R_{med} the median of R_i . The statistic E_i shows the influence of a record on the aggregate,

$$E_i = S_i * \left\{ \max(w_i^A x_i, w_i^B y_i) \right\}^u.$$

Here u is the size parameter, which describes the influence of a record on the aggregate value. The idea is to use a standard value for u . Often $u = 1/2$ is used, just as Hunt, Johnson and King (1999) did in their research for the Monthly Retail Trade Survey.

54. A problem with the above score functions is that only one variable within a record is used. When there are several variables of interest then several score functions should be computed. However, often a relationship between several variables exists, which cannot be taken in account with these functions. An example will be given below. This is also an example for outlier detection according to subsection V.A.

55. For the situations $i \in A \setminus B$ or $(i \in B \setminus A)$ the score function

$$SF_3 = \frac{|w_i x_i|}{\left| \sum_{i=1}^m w_i x_i \right|}, \quad (4)$$

can be used. Here m is the number of records. For this method a missing value is expected to be zero.

56. Now comparing VAT, SBS, and CT will be discussed as an example. Aggregates of the variable net turnover are compared for the building industry. This is done for enterprises in the intersections of these sources (VAT \square CT \square SBS). The aggregates of the strata (Publication cell x Size Class) do not

appear to be equal. After finding differences in the aggregates, records are compared with the same record in another source. For comparison score function (3) of Hidioglou en Berthelot is applied. Not for all records one source is completely consistent with another source. Several records of SBS are consistent with VAT and other records are consistent with CT. For other records VAT and CT are consistent with each other, but not consistent with SBS.

57. For production of SBS VAT-turnover is used to edit SBS-data, identify outliers, and weight SBS-data. For this reason also raw SBS data is considered and compared with CT. Note that for several enterprises raw SBS is more consistent with CT, while SBS after editing is more consistent with VAT. In table 5 three records are described for which the raw SBS-net turnover is consistent with CT-turnover and SBS- turnover after editing is more consistent with VAT. VAT and CT are not consistent in this example.

Table 5. Score function (3) applied on the sources SBS, VAT, and CT.

Enterprise:	Cell:	Size class:	Net turnover		Net turnover		VAT-CT:	SBS-VAT:	SBS-CT:	SBS raw-CT:
			VAT:	SBS:	raw SBS:	CT:				
10824847	45211	4	1203,341	1154	900	902,076	11,59	-1,45	9,49	-0,07
10345949	45211	5	6757,692	6276	4577	4577,435	39,16	-6,23	29,4	0
11065230	45332	5	7605,597	8080	2941	2940,993	139,34	5,31	157,07	0

58. The cells 45211 and 45332 stand for respectively general civil and commercial and industrial building, and installation of central heating. The closer a score is to zero the better consistency between sources for that specific record. One can conclude that the score for these records for SBS raw versus CT is close to zero, while SBS versus CT has a higher score than SBS versus VAT. This can also be seen, by considering the net turnover.

59. Often a data source contains two or more variables which are (linearly) dependent. Combining these variables means that not only, for example, turnover is considered, but also variables like earnings, and size of the enterprise. It can happen that an enterprise as a whole is different in comparison with other enterprises in one source, but also between different sources. This is the case when for example the size of an enterprise differs from other enterprises in the stratum, but the relation between different variables is the same.

60. By calculating the difference matrix Z and the score function matrix SF_2 , the records which contain high differences with respect to the other records for the same variables can be indicated. It is also possible to check whether variables satisfy editing rules. When a record or combination of records does not satisfy one or more edit rules values of variables have to be adapted. The example below treats three dependent variables and the use of score function SF_2 .

61. In this example x_1 and y_1 are employees in heads, x_2 and y_2 are employees in FTE, and x_3 and y_3 total earnings for respectively source A and B. The editing rules that have to be satisfied are $x_1 \leq x_2$, $y_1 \leq y_2$, $5 \leq x_3 / x_2 \leq 50$, and $5 \leq y_3 / y_2 \leq 50$. The variable z_i is the difference between x_i and y_i for $i=1,2,3$.

$$\begin{array}{ccc}
 \begin{matrix} x_1 & x_2 & x_3 \\ \left(\begin{array}{ccc} 10 & 5 & 100 \\ 4 & 8 & 200 \\ 200 & 150 & 3000 \\ 16 & 10 & 250 \\ 20 & 20 & 400 \\ - & - & - \end{array} \right) & \begin{matrix} y_1 & y_2 & y_3 \\ \left(\begin{array}{ccc} 8 & 4 & 1000 \\ 10 & 8 & 200 \\ 220 & 170 & 3500 \\ 15 & 10 & 250 \\ 20 & 16 & 400 \\ 10 & 10 & 100 \end{array} \right) & = & \begin{matrix} z_1 & z_2 & z_3 \\ \left(\begin{array}{ccc} 2 & 1 & -900 \\ -6 & 0 & 0 \\ -20 & -20 & -500 \\ 1 & 0 & 0 \\ 0 & 4 & 0 \\ 10 & 10 & 100 \end{array} \right) \\ \text{A} & \text{B} & \text{Z}
 \end{matrix}
 \end{array}$$

The matrix below will show the results of score function SF_2 for x_{ij} and y_{ij} , where $i=1,\dots,6$ and $j=1,\dots,3$:

$$SF_2 = \begin{pmatrix} 0.087 & 0.043 & 39.13 \\ 0.26 & 0 & 0 \\ 0.87 & 0.87 & 21.74 \\ 0.043 & 0 & 0 \\ 0 & 0.17 & 0 \\ 0.43 & 0.43 & 4.35 \end{pmatrix}$$

62. When one considers the data sources A and B, one can see that for the first row $y_3 / y_2 = 500$, while $x_3 / x_2 = 25$. Probably the variable earnings (y_3) is inconsistent for the first enterprise of Y. This can also be concluded from the score for $SF_{2_{13}}$. For the second enterprise probably the value $x_4 = 4$ is inconsistent. The number of employees in heads is lower than the number of employees in FTE. This is not easy to see from the score function. The third enterprise has high values for all variables. A reason for this could be that this enterprise has a too low size class. The last enterprise is missing for the first source. One can see this also by considering the difference between the two data sources. By using the score function one also can conclude that the fourth and fifth record are consistent.

V. CONCLUSIONS

63. The division Business Statistics of Statistics Netherlands aims to integrate business registers and surveys using a general and sound approach. The ultimate goal is to use businesses registers in stead of surveys, whenever possible. To integrate business data we need an environment for data storage, matching, and manipulations. A software development project aims to deliver a base version of an Economical Statistical Database in spring 2007.

64. Since 2004 the research project ESD Integration aims to improve transparency and quality of business statistics and registers due to assessment of statistical processes and confrontation of sources. Furthermore, it aims to improve consistency of data sources and a decrease of workload for the department of National Accounts. A general strategy is followed to obtain consistent aggregates. For the first phase of ESD Integration we integrate six key variables from eight annual business data sources available on enterprise level for 2001-2004.

65. Several methodological issues must be solved to obtain consistent aggregates. The main issues are handling of outliers, weighting of incomplete registers, and detection of influential inconsistencies. For ESD Integration an attempt is made to solve these issues. For handling of outliers we use common techniques. To weight incomplete registers we use covariates from a matched register that is almost complete. Score functions are used for detection of influential inconsistencies.

66. Within five years we hope to have a general editing system for business registers and surveys, where comparable variables within different data sources are linked. This system should also be linked to a database which contains a complete list of possible causes, consequences and solutions for inconsistencies at the micro level. This should lead to increased quality of business statistics.

References

- Arts, C., and E. Hoogteijling, 2002, *The Social Statistical Database 1998 and 1999 (In Dutch)*. Sociaal-economische maandstatistiek, Vol. 12, pp. 13-21.
- Farwell, K. and Raine, M. (2000), Some Current Approaches to Editing in the ABS, *Proceedings of the Second International Conference on Establishment Surveys, American Statistical Association*, pp. 529-538.
- Fellegi, I., and D. Holt, 1976, *A systematic approach to automatic edit and imputation*. Journal of the American Statistical Association, Vol. 71-353, pp. 17-35.
- Groen, R., 2006, *Detecting inconsistencies in integrated business data using robust techniques*. Internal paper, Statistics Netherlands, Voorburg
- Heerschap, N., and L. Willenborg, 2006, *Towards an integrated statistical system at Statistics Netherlands*. Forthcoming paper for ISI-Review Vol. 74-3.
- Hidiroglou, M., and J. Berthelot, 1986, *Statistical Editing and Imputation for Periodic Business Surveys*, Survey Methodology, Vol. 12, pp. 73-83.
- Hoekstra, M., 2002, *Extreme values in VAT, research for wholesale trade of 1997 (In Dutch)*. Internal paper, Statistics Netherlands, Voorburg.
- Hoogland, J., 2002, *Selective editing by means of Plausibility Indicators*. Research paper no. 227, Statistics Netherlands, Voorburg.
- Hoogland, J., 2005, *Selective editing using Plausibility Indicators and SLICE*, Statistical Data Editing, Volume 3: Impact on Data Quality.
- Houbiers, M., 2004, *Towards a Social Statistical Database and unified estimates at Statistics Netherlands*, Journal of Official Statistics, Vol. 20, 2004, pp. 55-75.
- Hunt, J., J. Johnson, and C. King, 1999, Detecting Outliers in the Monthly Retail Trade Survey using the Hidiroglou-Berthelot Method, *Proceedings of the section on survey research methods, American Statistical Association*, pp. 539-547.
- Vlag, P., G. Heunen, N. Nieuwenbroek, M. Das, and H. Pustjens, 2001, *Functionalities of weighting process: Technical specifications for outlier detection (In Dutch)*. Internal paper, Statistics Netherlands, Heerlen.
- Waal, T. de, 2000, SLICE: generalised software for statistical data editing and imputation. In: *Proceedings in computational statistics 2000* (ed. J.G. Bethlehem and P.G.M. van der Heijden), Physica-Verlag, Heidelberg, pp. 277-282.
