

The editing of statistical data: methods and techniques for the efficient detection and correction of errors and missing values

Ton de Waal, Jeroen Pannekoek and Sander Scholtus

The views expressed in this paper are those of the author(s)
and do not necessarily reflect the policies of Statistics Netherlands

Discussion paper (201132)



Explanation of symbols

.	= data not available
*	= provisional figure
**	= revised provisional figure (but not definite)
x	= publication prohibited (confidential figure)
–	= nil or less than half of unit concerned
–	= (between two figures) inclusive
o (o.o)	= less than half of unit concerned
blank	= not applicable
2010–2011	= 2010 to 2011 inclusive
2010/2011	= average of 2010 up to and including 2011
2010/'11	= crop year, financial year, school year etc. beginning in 2010 and ending in 2011
2008/'09–	
2010/'11	= crop year, financial year, etc. 2008/'09 to 2010/'11 inclusive

Due to rounding, some totals may not correspond with the sum of the separate figures.

Publisher

Statistics Netherlands
Henri Faasdreef 312
2492 JP The Hague

Prepress

Statistics Netherlands
Grafimedia

Cover

Tel design, Rotterdam

Information

Telephone +31 88 570 70 70
Telefax +31 70 337 59 94
Via contact form:
www.cbs.nl/information

Where to order

E-mail: verkoop@cbs.nl
Telefax +31 45 570 62 68

Internet

www.cbs.nl

ISSN: 1572-0314

© Statistics Netherlands,
The Hague/Heerlen, 2011.
Reproduction is permitted.
'Statistics Netherlands' must be quoted as source.

The editing of statistical data: methods and techniques for the efficient detection and correction of errors and missing values

Ton De Waal, Jeroen Pannekoek and Sander Scholtus

Summary: In order to produce official statistics of sufficient quality, statistical institutes carry out an extensive process of checking and correcting the data that they collect. This process is called statistical data editing. In this article, we give a brief overview of current data editing methodology. In particular, we discuss the application of selective and automatic editing procedures to improve the efficiency and timeliness of the data editing process.

Keywords: data editing, error localization, interactive editing, automatic editing, selective editing, score functions, macro-editing, systematic errors, random errors, deductive correction, Fellegi-Holt paradigm, imputation.

1 Introduction

It is the task of National Statistical Institutes (NSIs) and other official statistical institutes to provide high quality statistical information on many aspects of society, as up-to-date and as accurately as possible. One of the difficulties in performing this task arises from the fact that the data sources that are used for the production of statistical output, both traditional surveys as well as administrative data, inevitably contain errors that may affect the estimates of publication figures. In order to prevent substantial bias and inconsistencies in publication figures, NSIs therefore carry out an extensive process of checking the collected data and correcting them if necessary. This process of improving the data quality by detecting and correcting errors encompasses a variety of procedures, both manual and automatic, that are referred to as *statistical data editing*.

Errors can arise during the measurement process. Reported values are then different from the true values. One reason for a measurement error is that the true value is unknown to the respondent or difficult to obtain. Another reason can be that questions are misinterpreted or misread by the respondent. Also typing errors or other mistakes can cause differences between reported and true values. Besides measurement errors, errors can also arise in the further processing of the data which may include keying or coding. Missing data can arise when a respondent does not know the answer to a question, accidentally skips a question or refuses to give the answer to a certain question. Missing data can be seen as a special kind of error.

In traditional survey processing, statistical data editing was mainly an interactive activity intended to correct all data in every detail. Detected errors or inconsistencies were reported and explained on a computer screen and corrected after consulting the questionnaire, or re-contacting respondents: time and labor-intensive procedures. Alternatives to having all data edited interactively by subject-matter specialists appeared around the 1970s and have been developed ever since. There are basically two approaches to reducing the amount of interactive editing. One is selective editing, which

is based on the notion that not all errors need to be corrected in order to produce reliable publication figures (e.g. tables of estimates of totals, means and percentages). Several studies have shown (see, e.g., Granquist, 1995 and 1997; Granquist and Kovar, 1997) that it is usually sufficient to edit only the most influential errors. Selective editing refers to the process of identifying the possibly influential errors that will be edited interactively. The second approach to a more efficient editing process is to find methods that can detect and correct errors automatically, without any intervention of human editors. In modern statistical processes, NSIs usually apply a combination of selective editing and automatic procedures.

In automatic error detection and correction systems, the following steps can be distinguished. First, the occurrence of an error is detected. This is often done by comparing the values in that record with our knowledge of admissible (or plausible) values and combinations of values of the variables in each record. This knowledge is formulated in a set of rules called edit rules or edits for short. Inconsistency of the data values with the edit rules means that there is an error, but if a violated edit rule involves several variables (e.g. $Profit = Turnover - Total\ costs$), then it is not immediately clear which of the variables are in error. Therefore, in a second step, incorrect values in an inconsistent record have to be localized. This is often called *error localization*. Finally, in a third step, the localized erroneous fields are corrected, that is the values are replaced with better, preferably the correct, values. Replacement with new values is often called *imputation*.

A comprehensive description of statistical data editing is given in De Waal, Pannekoek and Scholtus (2011). In this article we only give a brief overview. The remainder of this article is organized as follows. We first describe the above-mentioned procedures in more detail: Section 2 discusses the use of edit rules for detecting inconsistencies; Section 3 discusses automatic editing; Section 4 discusses imputation methods; and Section 5 discusses selective editing. Next, Section 6 discusses a generic data editing strategy that combines selective editing, automatic editing, and interactive editing. Some concluding remarks follow in Section 7.

2 Checking for inconsistencies: the edit rules

Inconsistencies are most often detected by checking edit rules. It is important to have an extensive set of edits representing as well as possible the prior knowledge about the data values that are valid and those that are invalid. This knowledge is the input for automatic error detection procedures and is also used to guide the editors in the interactive editing process. To illustrate the kind of edits that are often applied in practice, examples of a number of typical classes of edits are given below.

Edits can be divided into *hard* (or *fatal*) edits and *soft* (or *query*) edits. Hard edits are edits that must be satisfied in order for a record to qualify as a valid record. As an example, a hard edit for a business survey specifies that the variable *Total costs* needs to be equal to sum of the variables *Personnel costs*, *Capital costs*, *Transport costs* and *Other costs*. Records that violate one or more hard edits are considered to be

inconsistent and it is deduced that some variable(s) in such a record must be in error. Soft edits are used to identify unlikely or deviating values that are suspected to be in error although this is not a logical necessity. An example is an edit specifying that the turnover per employee of a firm may not be larger than ten times the value of the previous year. The violation of soft edits can trigger further investigation of these edit failures, to either confirm or reject the suspected values.

The simplest edits are edits describing the admissible values of a single variable, sometimes called range restrictions. For categorical variables, a range restriction simply verifies whether the observed category codes for the variable belong to the specified set of codes.

Many edits that are important in the editing process involve more than one variable. These edits describe admissible (or inadmissible) combinations of values of the variables involved. For example an edit on two variables could be that marital status is unmarried for persons with age less than 15. Another example is the so-called *ratio edit* which sets bounds on the allowable range of a ratio between two variables, such as the turnover per employee mentioned above. Ratio edits are often soft edits.

Balance edits are edits that state that the admissible values of a number of variables are related by a linear equality. They occur mainly in business statistics where they are linear equations that should be satisfied according to accounting rules. Typically, balance edits are hard edits. Two examples are:

$$Profit = Turnover - Total\ costs$$

and

$$Total\ costs = Employee\ costs + Other\ costs.$$

As is often the case with balance edits, these two edits are related because they have the variable *Total costs* in common. Balance edits are of great importance for editing economic surveys where there are often a large number of such edits. For instance, in the annual structural business statistics there are typically about a hundred variables with thirty or more balance edits. Furthermore, edits in the form of linear *inequalities* are specified as well. These inter-related systems of linear relations that the values must satisfy provide much information about possible errors and missing values.

3 Automatic editing

3.1 Localization and correction of systematic errors

In automatic editing, a distinction is often made between so-called *systematic errors* and *random errors*, and different methods are used to detect and correct these errors.

A systematic error is an error that occurs frequently between responding units. A well-known type of systematic error is the so-called *unity measure error* which is the error of, for example, reporting financial amounts in Euros instead of the requested

thousands of Euros. See Al-Hamad, Lewis and Silva (2008) for a discussion of the detection of unity measure errors. Systematic errors can lead to substantial bias in aggregates, but once detected, systematic errors can easily and reliably be corrected because the underlying error mechanism is known. It is precisely this knowledge of the underlying cause that makes systematic errors different from random errors. In fact, in automatic editing, all errors without an (as of yet) detectable cause are treated as random errors.

De Waal and Scholtus (2011) make a further distinction between *generic* and *subject-related* systematic errors. Errors of the former type occur for a wide variety of variables in a wide variety of surveys and registers, where the underlying cause is always essentially the same. Apart from the unity measure error, other examples include:

- *simple typing errors*, such as interchanged or mistyped digits (see Scholtus, 2009);
- *sign errors*, such as forgotten minus signs or interchanged pairs of revenues and costs (see Scholtus, 2008 and 2011);
- *rounding errors*, where a balance edit is violated, but the size of violation is very small (see Scholtus, 2008 and 2011).

These errors can often be detected and corrected automatically by using mathematical techniques, as discussed in the above-mentioned references. Methods for correcting simple typing errors, sign errors, and rounding errors have recently been implemented in an R package called *deducorrect*; see Van der Loo, De Jonge and Scholtus (2011).

Subject-related systematic errors are specific to a particular questionnaire or survey. They may be caused by a frequent misunderstanding or misinterpretation of some question such as reporting individual rather than family income. Subject-related systematic errors are usually detected and corrected by applying correction rules specified by subject-matter experts.

Localization and correction of systematic errors is an important first step in the editing process. It can be done automatically and reliably at virtually no costs and hence will improve both the efficiency and the quality of the editing process. It is in fact a very efficient and probably often underused correction approach.

3.2 Automatic detection of random errors

When the systematic errors have been removed, the remaining violations of hard edits now indicate the presence of random errors. It is straightforward to check for violations of edit rules, but it is not so obvious how to decide which variable(s) in an inconsistent record are in error. The most common and fruitful approach to this error localization problem is based on the paradigm of Fellegi and Holt (1976): *The data in each record should be made to satisfy all edits by changing the fewest possible items of data (values of variables).*

This paradigm was later generalized by assigning reliability weights to the variables and minimizing the sum of the weights of the variables that are to be changed to make the record satisfy all edits. In this form the error localization problem according to the Fellegi-Holt paradigm can be formulated as a mathematical optimization problem. We refer to De Waal, Pannekoek and Scholtus (2011) or De Waal (2003) for such a formulation and several solution methods. Alternative references are De Waal and Coutinho (2005) where an overview of algorithms for solving the Fellegi-Holt based error localization problem for numerical data is presented, and De Waal and Quere (2003) where an algorithm that solves the error localization problem for a combination of categorical and numerical data is described.

A solution to the error localization problem is basically just a list of all variables that need to be changed. These variables are set to missing and are subsequently imputed in a separate step.

Provided that the set of edits used is sufficiently powerful, application of the Fellegi-Holt paradigm generally results in data of higher statistical quality.

4 Imputation: correction of missing data and random errors

Missing data is a well-known problem that has to be faced by basically all institutes that collect data on persons or enterprises. In the statistical literature ample attention is hence paid to missing data. The most common solution to handle missing data is imputation, where missing values are estimated and filled in. Missing data arise not only because no response was received for some variables but also during the editing process. Values that were detected as random errors are also treated as “missing”.

An imputation model predicts a missing value using a function of auxiliary variables, the predictors. The auxiliary variables may be obtained from the current survey or from other sources such as historical information (the value of the missing variable in a previous period) or, increasingly important, administrative data. The most common types of imputation models are variants of regression models with parameters estimated from the observed correct data. However, especially for categorical variables, donor methods are also frequently used. Donor methods replace missing values in a record with the corresponding values from a complete and valid record. Often a donor record is chosen such that it resembles as much as possible the record with missing values. It depends on the characteristics of the data set and the research goals, which imputation method is best suited for a particular situation.

There is a vast literature on imputation since it plays an important role, not only in official statistics, but in many other fields in statistics as well. See, e.g., Rubin (1987), Schafer (1997), and Little and Rubin (2002).

At NSIs the imputation problem is further complicated owing to the existence of edit rules that have to be satisfied by the imputed data. Some imputation methods have been developed that can take edit rules into account (Tempelman, 2007), but for many problems such models become too complex. The problem of consistency with the edit

rules can then be solved by the introduction of an adjustment step in which adjustments are made to the imputed values such that the record satisfies all edits and the adjustments are as small as possible. Various formulations of the adjustment problem are discussed in Pannekoek and Zhang (2011).

5 Selective editing

Selective editing is an approach that aims to identify the records with potentially influential errors and restrict the costly interactive editing to those records only. This approach is particularly useful for business surveys, since small and large enterprises often have different contributions to total values, but less so for social surveys, since individuals often have a similar influence on estimates. Methods for the selection of records that will be followed up by editors can be divided into *micro-selection* and *macro-selection methods*.

Micro-selection The extent to which a record potentially contains influential errors can be measured by a score function (cf. Latouche and Berthelot, 1992; Lawrence and McKenzie, 2000; Farwell and Rain, 2000). This function is constructed such that records with high scores likely contain errors that have substantial effects on estimates of target parameters. A score for a record (record or global score) is usually a combination of scores for each of a number of important variables (the local scores). The local scores are generally constructed so that they reflect the following two elements that together constitute an influential error: the likelihood of a potential error (the “risk” component) and the contribution or influence of that record on the estimated target parameter (the “influence” component). Local scores are then defined as the product of these two components, i.e.:

$$s_{ij} = F_{ij} \times R_{ij}, \quad (1)$$

with F_{ij} the influence component and R_{ij} the risk component for unit i and variable j . The risk component can be measured by comparing the observed value with an “expected” value that is often based on information from previous cycles of the same survey. Large deviations from the expected value are taken as an indication that the value may be in error and, if indeed so, that the error is substantial. The influence component can often be measured as the (relative) contribution of the expected value to the estimated total. A global or unit score S_i , say, combines the local scores into a single measure for the whole unit. For the selection, a threshold value for the score S_i is set and all records with scores above this threshold (the *critical stream*) are directed to manual reviewers whereas records with scores below the threshold are treated automatically.

Macro-selection Micro-selection methods use the data of a single record and related auxiliary information to determine possible influential errors. These methods can be applied from the start of the data collection phase, as soon as records become available. In contrast, macro-selection techniques use information from other records and

can only be applied if a substantial part of the data has been collected. Two forms of macro-editing can be distinguished. The first form is sometimes called the aggregation method (see, e.g., Granquist, 1990). It formalizes and systematizes what every statistical agency does before publication: verifying whether figures to be published seem plausible. This is accomplished by comparing quantities in publication tables with the same quantities in previous publications. Only if an unusual value is observed, a micro-editing procedure is applied to the individual records and fields contributing to the suspicious quantity. A second form of macro-editing is the distribution method. The available micro-data are used to characterize the distribution of the variables. Then, all individual values are compared with the distribution. Typically, measures of location and spread are computed. Records containing values that could be considered uncommon (given the distribution) are candidates for further inspection and possibly for editing. Methods for outlier detection described in the general statistical literature can be applied at this stage; see, e.g., Barnett and Lewis (1994), Chambers, Hentges and Zhao (2004), Rocke and Woodruff (1996), Rousseeuw and Leroy (1987), and Todorov, Templ and Filzmoser (2009).

6 A data editing strategy

Data editing is usually performed as a sequence of different detection and/or correction process steps. In this section we give a global description of an editing strategy. This editing strategy, depicted in Figure 1, consists of the following five steps.

1. Treatment of systematic errors: identify and eliminate errors that are evident and easy to treat with sufficient reliability.
2. Micro-selection: select records for interactive treatment that contain influential errors that cannot be treated automatically with sufficient reliability.
3. Automatic editing: apply automatic error localization and imputation procedures to the (many) records that are not selected for interactive editing in step 2. This step treats the remaining (random) errors since the systematic errors are already resolved in step 1.
4. Interactive editing: apply interactive editing to the minority of the records with influential errors.
5. Macro-selection: select records with influential errors by using methods based on outlier detection techniques and other procedures that make use of all or a large fraction of the response. Influential errors not detected in step 2 or 3 (because no edit rule or score function could detect them) can be detected here.

Note that there are two kinds of process steps: those that localize or treat errors and those that direct the records through the different stages of the process. The processes in step 2 and 5 are of the latter kind; they are “selectors” that do not actually treat errors, but select records for specific kinds of further processing.

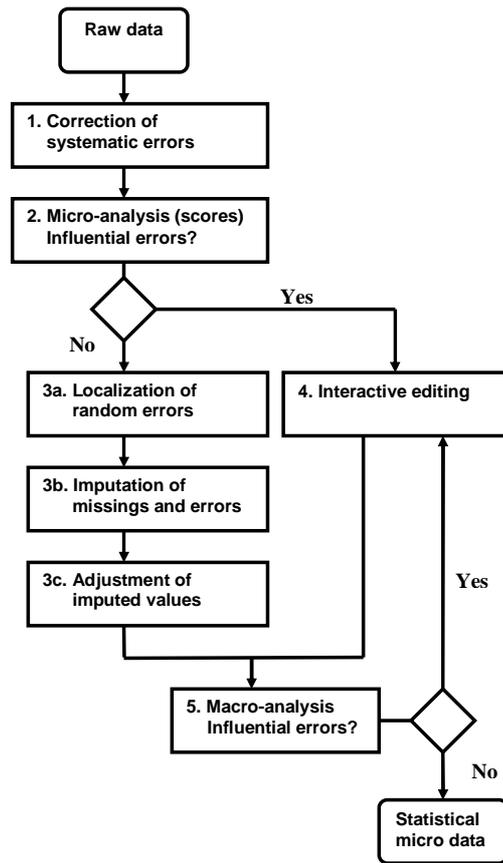


Figure 1. Example of a data editing process flow

The process flow suggested in Figure 1 is just one possibility. Depending on the type of survey and the available resources and auxiliary information, the process flow can be different. Not all steps are always carried out, the order of steps may be different, and the particular methods used in each step can differ between types of surveys or, more generally, data sources.

Pannekoek and De Waal (2005) illustrate how suitable data editing strategies can be constructed for practical cases.

7 Concluding remarks

During the past decades statistical data editing has evolved from an exhaustive labor-intensive process involving human intervention for most corrections to a much more refined ensemble of techniques, involving advanced fully automatic detection and correction procedures for large amounts of records and human intervention for the, cleverly selected, minority of cases where that really pays off. The methodology draws heavily on disciplines such as mathematical optimization, statistical modeling and outlier detection. Application of techniques from these fields in combination with an effective use of subject-matter knowledge can result in efficiency gains without compromising, sometimes even enhancing, the quality of results. This is a most wanted result, especially since many NSIs currently face budget reductions while the need for high quality detailed statistical information only seems to increase.

The way statistical offices are collecting their data is changing and more research is necessary to establish the optimal data editing procedures for different types of data sources. For instance, NSIs seem to be moving towards the use of mixed mode data collection, where data are collected by a mix of different modes, such as paper questionnaires, computer assisted personal interviewing, computer assisted telephone interviewing and web surveys. This obviously has consequences for statistical data editing. Some of these consequences have been examined by Børke (2008), Hoogland and Smit (2008), and Van der Loo (2008).

In order to reduce costs and response burden, most statistical institutes aim to increase the use of administrative sources for producing their statistical outputs (see, e.g., Wallgren and Wallgren, 2007). However, in many cases the required statistical output cannot be obtained from a single administrative source and it is necessary to link several administrative sources and/or surveys to obtain the necessary information. In the editing of these combined sources several challenges arise. The amount of data can be huge, stressing the need for automatic methods. There may be inconsistencies between responses from different sources, not only because of measurement errors but also due to (slight) differences in definitions or times of measurement. Methods to reconcile these differences must be found. The linkage process is not flawless and there is a need for automatic methods to detect and correct linkage errors.

Concluding we may say that while the field of statistical data editing has become a mature and well-developed discipline (see De Waal, Pannekoek and Scholtus, 2011,

and the many references therein) over the last few decades, research in this field is still alive and kicking. In order to accommodate the constantly changing situation at NSIs, for instance with respect to data collection modes and the use of administrative data, the advancement of methods for automatic and selective editing remains an active and rewarding area of research.

References

Al-Hamad, A., D. Lewis and P.L.N. Silva (2008), *Assessing the Performance of the Thousand Pounds Automatic Editing Procedure at the ONS and the Need for an Alternative Approach*. Working Paper No. 21, UN/ECE Work Session on Statistical Data Editing, Vienna.

Barnett, V. and T. Lewis (1994), *Outliers in Statistical Data*. John Wiley & Sons, New York.

Børke, S. (2008), *Using “Traditional” Control (Editing) Systems to Reveal Changes when Introducing New Data Collection Instruments*. Working Paper No. 6, UN/ECE Work Session on Statistical Data Editing, Vienna.

Chambers, R., A. Hentges and X. Zhao (2004), Robust Automatic Methods for Outlier and Error Detection. *Journal of the Royal Statistical Society A* 167, pp. 323-339.

De Waal, T. (2003), *Processing of Erroneous and Unsafe Data*, Ph.D. Thesis, Erasmus University, Rotterdam (see also www.cbs.nl).

De Waal, T. and W. Coutinho (2005), Automatic Editing for Business Surveys: an Assessment for Selected Algorithms. *International Statistical Review* 73, pp. 73-102.

De Waal, T., J. Pannekoek and S. Scholtus (2011), *Handbook of Statistical Data Editing and Imputation*. John Wiley & Sons, Hoboken, New Jersey.

De Waal, T. and R. Quere (2003), A Fast and Simple Algorithm for Automatic Editing of Mixed Data. *Journal of Official Statistics* 19, pp. 383-402.

De Waal, T. and S. Scholtus (2011), *Methods for Automatic Statistical Data Editing*. Paper presented at the 2011 KSS International Conference on Statistics and Probability, Busan.

Farwell, K. and M. Rain (2000), Some Current Approaches to Editing in the ABS. *Proceedings of the Second International Conference on Establishment Surveys*, Buffalo, pp. 529-538.

Fellegi, I.P. and D. Holt (1976), A Systematic Approach to Automatic Edit and Imputation. *Journal of the American Statistical Association* 71, pp. 17-35.

Granquist, L. (1990), A Review of Some Macro-Editing Methods for Rationalizing the Editing Process. *Proceedings of the Statistics Canada Symposium*, pp. 225-234.

- Granquist, L. (1995), Improving the Traditional Editing Process. In: *Business Survey Methods* (ed. Cox, Binder, Chinnappa, Christianson, Colledge, and Kott), John Wiley & Sons, New York, pp. 385-401.
- Granquist, L. (1997), The New View on Editing. *International Statistical Review* 65, pp. 381-387.
- Granquist, L. and J. Kovar (1997), Editing of Survey Data: How Much is Enough? In: *Survey Measurement and Process Quality* (ed. Lyberg, Biemer, Collins, De Leeuw, Dippo, Schwartz, and Trewin), John Wiley & Sons, New York, pp. 415-435.
- Hoogland, J. and R. Smit (2008), *Selective Automatic Editing of Mixed Mode Questionnaires for Structural Business Statistics*. Working Paper No. 2, UN/ECE Work Session on Statistical Data Editing, Vienna.
- Latouche, M. and J.M. Berthelot (1992), Use of a Score Function to Prioritize and Limit Recontacts in Editing Business Surveys. *Journal of Official Statistics* 8, pp. 389-400.
- Lawrence, D. and R. McKenzie (2000), The General Application of Significance Editing. *Journal of Official Statistics* 16, pp. 243-253.
- Little, R.J.A. and D.B. Rubin (2002), *Statistical Analysis with Missing Data* (second edition). John Wiley & Sons, New York.
- Pannekoek, J. and T. De Waal (2005), Automatic edit and imputation for business surveys: The Dutch contribution to the EUREDIT project. *Journal of Official Statistics* 21, pp. 257-286.
- Pannekoek, J. and L.-C. Zhang (2011), *Partial (Donor) Imputation with Adjustments*. Working Paper No. 40, UN/ECE Work Session on Statistical Data Editing, Ljubljana.
- Rocke, D.M. and D.L. Woodruff (1996), Identification of Outliers in Multivariate Data. *Journal of the American Statistical Association* 91, pp. 1047-1061.
- Rousseeuw, P.J. and M.L. Leroy (1987), *Robust Regression & Outlier Detection*. John Wiley & Sons, New York.
- Rubin, D.B. (1987), *Multiple Imputation for Non-Response in Surveys*. John Wiley & Sons, New York.
- Schafer, J.L. (1997), *Analysis of Incomplete Multivariate Data*. Chapman & Hall, London.
- Scholtus, S. (2008), *Algorithms for Correcting Some Obvious Inconsistencies and Rounding Errors in Business Survey Data*. Discussion Paper 08015, Statistics Netherlands, The Hague (see also www.cbs.nl)
- Scholtus, S. (2009), *Automatic Detection of Simple Typing Errors in Numerical Data with Balance Edits*. Discussion Paper 09046, Statistics Netherlands, The Hague (see also www.cbs.nl)

Scholtus, S. (2011), Algorithms for Correcting Sign Errors and Rounding Errors in Business Survey Data. *Journal of Official Statistics* 27, pp. 467-490.

Tempelman, D.C.G. (2007), *Imputation of Restricted Data*. Ph.D. Thesis, University of Groningen (see also www.cbs.nl).

Todorov, V., M. Templ and P. Filzmoser (2009), *Outlier Detection in Survey Data using Robust Methods*. Working Paper No. 40, UN/ECE Work Session on Statistical Data Editing, Neuchâtel.

Van der Loo, M.P.J. (2008), *An Analysis of Editing Strategies for Mixed-Mode Establishment Surveys*. Discussion Paper 08004, Statistics Netherlands, The Hague (see also www.cbs.nl).

Van der Loo, M.P.J., E. de Jonge and S. Scholtus (2011), *deducorrect: Deductive correction of simple rounding, typing and sign errors*. R package, available at www.cran.r-project.org/web/packages/deducorrect/.

Wallgren, A. and B. Wallgren (2007), *Register-based Statistics – Administrative Data for Statistical Purposes*. John Wiley & Sons, Chichester.