An overview of statistical data editing

Ton de Waal

Discussion paper (08018)

Statistics Netherlands

The Hague/Heerlen

Explanation of symbols

.	= data not available
*	= provisional figure
x	= publication prohibited (confidential figure)
–	= nil or less than half of unit concerned
–	= (between two figures) inclusive
0 (0,0)	= less than half of unit concerned
blank	= not applicable
2005-2006	= 2005 to 2006 inclusive
2005/2006	= average of 2005 up to and including 2006
2005/’06	= crop year, financial year, school year etc. beginning in 2005 and ending in 2006
2003/’04–2005/’06	= crop year, financial year, etc. 2003/’04 to 2005/’06 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 - Facility Services

Cover
TelDesign, 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

An overview of statistical data editing

Ton De Waal

Summary: This paper gives an overview of statistical data editing. The paper first describes the traditional interactive approach to data editing. It then focuses on modern editing techniques, such as selective editing, automatic editing, and macro-editing. The paper aims to provide an introduction to these topics, and give many references to the literature. A slightly modified version of this paper will be published in the book "Sample Surveys: Theory, Methods and Application", edited by Prof. D. Pfeffermann and Prof. C.R. Rao.

Key words: interactive editing, selective editing, automatic editing, macro-editing, statistical data editing

1 Introduction

Users of statistical information are nowadays demanding high quality data on social, demographic, industrial, economic, financial, political, and cultural aspects of society with a great level of detail and produced within a short span of time. National statistical institutes (NSIs) fulfill a central role in providing such high quality statistical information. Most NSIs face this challenge while their financial budgets are constantly diminishing.

A major complicating factor is that collected data generally contain errors. The data collection stage in particular is a potential source of errors. For instance, a respondent may give a wrong answer (intentionally or not), a respondent may not give an answer (either because he does not know the answer or because he does not want to answer this question), errors can be introduced at the NSI when the data are transferred from the questionnaire to the computer system, etc. The occurrence of errors in the observed data makes it necessary to carry out an extensive process of checking the collected data, and, when necessary, correcting them. This checking and correction process is referred to as statistical data editing (SDE).

To check and correct data two steps have to be carried out. First, the erroneous records and the erroneous fields in these records have to be localized. This is called the error localization step. Second, the localized erroneous fields and the missing fields have to be imputed, i.e. the values of the erroneous fields have to be replaced by better, preferably the correct, values and the values of the missing fields have to be estimated. This is called the imputation step. The error localization step only determines which fields are considered erroneous; the imputation step determines values for these fields as well as for the missing ones.

Although the error localization step and the imputation step are closely related in theory, in practice they are usually treated as two separate steps in the statistical process.

In this paper we will also treat them as two distinct steps. We will use the phrase SDE in the sense of localizing errors, unless stated otherwise.

Traditionally, statistical agencies have always put much effort and resources into SDE, as they considered it a prerequisite for publishing accurate statistics. In traditional survey processing, SDE was mainly an interactive activity where all individual records were checked with the aim to correct all data in every detail. It has long been recognized, however, that it is not necessary to correct all data in every detail. Several studies (see, for example, Granquist, 1984; Granquist, 1997; Granquist and Kovar, 1997; Pannekoek and De Waal, 2005) and many years of practical experience at several NSIs have shown that in general it is not necessary to remove all errors from a data set in order to obtain reliable publication figures. The main products of statistical offices are tables containing aggregate data, which are often based on samples of the population. This implies that small errors in individual records are acceptable. First, because small random errors in individual records generally tend to cancel out, i.e. their sum generally tends to be negligible in comparison to the corresponding publication figure. Second, because if the data are obtained from a sample of the population there will always be a sampling error in the published figures, even when all collected data are completely correct. In this case an error in the results caused by incorrect data is acceptable as long as it is small in comparison to the sampling error. In order to obtain data of sufficiently high quality it is usually sufficient to remove only the most influential errors.

In the past, and often even in the present, too much effort was spent on correcting errors that did not have a noticeable impact on the ultimately published figures. This has been referred to as “over-editing”. Over-editing not only costs budget, but also a considerable amount of time, making the period between data collection and publication unnecessarily long. Sometimes over-editing even becomes “creative editing”: the editing process is then pursued to such an extent that unlikely, but correct, data are “corrected”, or discarded and replaced. Such unjustified alterations can be detrimental for data quality.

To improve the efficiency of the editing process, modern techniques such as selective editing, automatic editing and macro-editing can be applied instead of the traditional micro-editing approach where all records are extensively edited manually. In this paper we discuss these editing techniques. A crucial role in several (versions) of these techniques is often played by so-called edit rules. We describe the use of these edit rules in Section 2. We then continue our discussion with interactive editing in Section 3. We examine the possibility of editing during the data collection phase in Section 4. In the next three sections we examine modern editing techniques: selective editing in Section 5, automatic editing in Section 6, and macro-editing in Section 7. In Section 8 we discuss a strategy for SDE based combining different editing techniques. Section 9 ends the paper with a brief discussion of possible future developments with respect to

SDE.

For more information on SDE in general we refer to Ferguson (1994), and for SDE for business surveys to EDIMBUS (2007). An important international project on SDE and imputation was the EUREDIT project. The EUREDIT project aimed at improving the efficiency and the quality of automatic methods for SDE and imputation at NSIs. For the main findings of this project we refer to EUREDIT Project (2004a and 2004b).

2 The use of edit rules

At NSIs, edit rules, or edits for short, are often used to determine whether a record is consistent or not. An example of an edit is

$$T = P + C, \tag{1}$$

where T is the turnover of an enterprise, P its profit, and C its costs. Edit (1) expresses that the profit and the costs of an enterprise should sum up to its turnover. Such an edit is referred to as a balance edit. Another example is $T \geq 0$, expressing that the turnover of an enterprise should be non-negative. Edits like this are referred to as non-negativity edits. A third example is $P/T \leq 0.5$. Such an edit expressing that the ratio of two variables should be less (or greater) than a certain threshold is referred to as a ratio edit. Examples of edits for categorical (discrete) data are that children of the head of household cannot be married to each other, and that a person can have only one (biological) mother.

In order to construct a set of edits one usually starts with the "hard" (or logical) edits, which hold true for all correctly observed records. Balance edits are usually hard edits. After the hard edits have been specified, one generally uses subject-matter knowledge and statistical analyses to add a number of "soft" edits, which hold true for a high fraction of correctly observed records but not necessarily for all of them. In many cases ratio edits are soft edits. The thresholds of soft ratio edits have to be carefully determined so correct records do not, or only very rarely, violate these edits, while the edits are powerful enough to pinpoint erroneous records. Another example of a soft edit is that a mother must be at least 15 years older than any of her children. In most cases where this edit is violated, the record under consideration is indeed incorrect. Only in extremely rare cases this edit is violated by a correct record.

Records that are inconsistent with respect to the edits, i.e. fail one or more edits, are considered to contain errors if hard edits are violated and are considered to be suspicious if only soft edits are violated. The values in an erroneous record have to be modified in such a way that the resulting record is a better approximation of the true data of the corresponding respondent. Suspicious records are either examined further or are treated as erroneous records.

A consistent record, i.e. a record that satisfies all edits, is not necessarily (considered to be) error-free. For instance, (some of) the values in a consistent record may be outliers with respect to the bulk of the data. Such outlying values are often considered suspicious and are hence checked in the editing process, even if all edits are satisfied.

To avoid over-editing one should in particular be careful not to specify too many soft edits. In general, users tend to apply more soft edits than necessary to the data (see Di Zio, Guarnera and Luzi, 2005a).

3 Interactive editing

The use of computers in the editing process started many years ago. In the early years their role was restricted to checking which edits were violated. For each record all violated edits were listed. Subject-matter specialists then used these lists to correct the records. That is, they retrieved all paper questionnaires that did not pass the edits and corrected these questionnaires, for instance by re-contacting the respondent or by comparing the respondent's data to data from similar respondents. After they had corrected the data, these data were again entered into the computer, and the computer again checked whether the data satisfied all edits. This iterative process continued until (nearly) all records passed the edits.

A major problem with respect to this approach was that during the manual correction process the records were not checked for consistency. As a result, a record that was "corrected" could still fail one or more specified edits. Such a record hence required more correction. It was not exceptional that some records had to be corrected several times. It is therefore not surprising that editing in this way was very costly, both in terms of budget as well as in terms of time (see e.g. Federal Committee on Statistical Methodology, 1990; Granquist and Kovar, 1997).

Subject-matter specialists have extensive knowledge with respect to their area of expertise. This knowledge should be used as well as possible. This aim can be achieved by providing subject-matter specialists with efficient and effective data editing tools. Survey-processing systems such as Blaise (see Blaise Reference Manual, 2002, and Blaise Developer's Guide, 2002) and CSPro (see CSPro User's Guide, 2007, and CSPro Data Entry User's Guide, 2007) are often used to edit data at NSIs. When such systems are used, the specified edits can be checked during or after data entry, and - if necessary - the data may immediately be corrected. This is referred to as interactive or computer-assisted editing. The introduction of systems such as Blaise and CSPro led to a substantial efficiency improvement of the editing process. In this section and the next we use Blaise as an example of a survey-processing system due to the wide use of the system.

When Blaise is used to edit the data, it is no longer necessary to edit the data in several iterations, each consisting of a checking phase and a correction phase. When data are corrected, new error signals due to failed edits, if any, are immediately shown on the computer screen. The error signals, in combination with the data themselves, direct the subject-matter specialist to potential errors in the data. For instance, Blaise can calculate the number of times each field is involved in a failed edit. Fields that are most often involved in failed edits are usually the most likely ones to be in error. In order to correct data, the subject-matter specialists often check the paper questionnaires or scanned versions thereof as this can help them to identify errors in the data.

Data from paper questionnaires can either be entered by fast data entry personnel or by subject-matter specialists. In the former case, the data are keyed in without attempting to edit them at this stage. Later, subject-matter specialists edit the keyed raw data. Alternatively, data can directly be entered by subject-matter specialists. This costs more time than letting the data be keyed in by data entry personnel. However, subject-matter specialists can enter and correct data at the same time. The extra time required to enter the data is often (more than) compensated for by the fact that each record is treated, i.e. entered or edited, only once. A practical drawback of keying in data and editing them at the same time is that the raw, unedited, data are not available for later analyses, for instance analyses with respect to the efficiency and effectiveness of the editing process itself.

An alternative to keying in data is scanning the paper questionnaires in combination with optical character recognition. For paper questionnaires for which the answers mainly consist of numerical data, this often leads to data of similar quality as keyed-in data. Paper questionnaires for which optical character recognition does not give good results are often scanned anyway in order to help the subject-matters specialists during the interactive editing process.

Interactive editing can be used to edit both categorical and numerical data, and is nowadays a standard way to edit data. The number of variables, edits and records may, in principle, be high. Survey managers generally consider data edited in an interactive manner to be of high statistical quality. For more on interactive editing by means of systems like Blaise we refer to Pierzchala (1990).

The fundamental problem of interactive editing is that, even though each record has to be edited only once, still all records have to be edited. We have already mentioned that this can, and often does, lead to over-editing. Instead of editing all records, one could consider editing only the ones with influential errors. This is referred to as selective editing and is discussed in Section 5. In Section 4 we first discuss the most efficient editing technique of all: no editing at all, but instead ensuring that correct data is obtained during the data collection phase.

4 Editing during the data collection phase

Blaise not only applies edits, but also so-called routing rules. Frequently, different questions are posed to different kinds of respondents. For instance, it is not useful to ask a male respondent whether he has ever been pregnant as the answer to this question would not provide any additional information. Blaise ensures that each respondent is asked the questions that are applicable to this kind of respondent. Owing to this functionality Blaise is an excellent system for CAPI (Computer Assisted Personal Interviewing), CATI (Computer Assisted Telephone Interviewing), CASI (Computer Assisted Self Interviewing), and CAWI (Computer Assisted Web Interviewing).

When CAPI is used to collect the data, an interviewer visits the respondent and enters the answers directly into a laptop. When CATI is used to collect the data, the interview is carried out during a telephone call. When CASI or CAWI is used to collect the data, the respondent fills in an electronic questionnaire himself. The difference between these two modes is that for CAWI an electronic questionnaire on the Internet has to be filled in, whereas for CASI an off-line electronic questionnaire has to be filled in. When an invalid value for a question is given or an inconsistency between the answers of two or more questions is noted during any of these data collection modes, this is immediately reported by Blaise. The error can then be resolved by asking the respondent these questions again. For CASI and CAWI generally not all edits that could be specified are actually specified, since the respondent might get annoyed and may refuse to complete the questionnaire when the edits keep on reporting that the answers are inconsistent.

In many cases data collected by means of CAPI, CATI, CASI or CAWI contain fewer errors than data collected by means of paper questionnaires as random errors that affect paper questionnaires can be detected and avoided at collection. For face-to-face interviewing CAPI has in fact become the standard. CAPI, CATI, CASI and CAWI may hence seem to be ideal ways to collect data, but - unfortunately - they too have their disadvantages.

A first disadvantage of CATI and CAPI is that CATI and, especially, CAPI are very expensive. A second disadvantage of CATI and CAPI is that a prerequisite for these two data collection modes is that the respondent is able to answer the questions during the interview. For a survey on persons and households this is often the case. The respondent often knows (good proxies of) the answers to the questions, or is able to retrieve the answers quickly. For a survey on enterprises the situation is quite different. Often it is impossible to retrieve the correct answers quickly, and often the answers are not even known by one person or one department of an enterprise. Finally, even in the exceptional case that one person knew all answers to the questions, the NSI would generally not know the identity of this person. For the above-mentioned reasons many NSIs frequently use CAPI and CATI to collect data on persons and households but only rarely for data on enterprises.

Pilot studies and actual applications have revealed that CASI and CAWI are indeed viable data collection modes, but also that several problems arise when these modes are used. Besides IT problems, such as that the software - and the Internet - should be fast and reliable and the security of the transmitted data should be guaranteed, there are a number of practical and statistical problems. We have already mentioned the practical problem that if the edits keep on reporting that the answers are inconsistent, the respondent may get annoyed and may refuse to fill in the rest of the questionnaire. An example of a statistical problem is that the group of people responding to a web survey may be selective (see, e.g., Bethlehem, 2007). Another important problem for CAWI and CASI is that data collected by of either of these data collection modes may appear to be of higher statistical quality than data collected by means of paper questionnaires, but in fact are not. When data are collected by means of CASI and CAWI one can enforce that the respondents supply data that satisfy build-in edits, or one can avoid balance edits by automatically calculating the totals from their components. As less edits are failed by the collected data, the collected data may appear to be of higher statistical quality. This may not be the case, however, as respondents can be less accurate when filling in the entries in an electronic questionnaire, especially if totals are computed automatically (see Børke, 2008; Hoogland and Smit, 2008).

NSIs seem to be moving towards the use of mixed-mode data collection, where data are collected by a mix of several data collection modes. This obviously has consequences for SDE. Some of the potential consequences have been examined by Børke (2008), Hoogland and Smit (2008), and Van der Loo (2008). For more information on computer assisted data collection in general we refer to Couper et al. (1998).

5 Selective editing

5.1 Introduction to selective editing

Selective (or significance) editing (see, e.g., Hidiroglou and Berthelot, 1986; Latouche and Berthelot, 1992; Lawrence and McDavitt, 1994; Lawrence and McKenzie, 2000; Farwell and Raine, 2000; Hoogland, 2002; Hedlin, 2003) is an umbrella term for several methods for identifying the influential errors in a data set, i.e. the errors that have a substantial impact on the publication figures. The aim of selective editing is to split the data into two streams: a critical and a non-critical stream. The critical stream consists of records that are the most likely ones to contain influential errors; the non-critical stream consists of records that are unlikely to contain influential errors. Only the records in the critical stream are edited interactively. The records in the non-critical stream are either not edited or are edited automatically (see Section 6).

The scope of most techniques for selective editing is limited to (numerical) business data. In these data some respondents can be more important than other respondents, simply because the magnitude of their contributions is higher. Social data are usually count data where respondents contribute more or less the same, namely their raising weight, to estimated population totals. In social data it is therefore difficult to differentiate between respondents. Selective editing has gradually become a popular method for editing business data and increasingly more NSIs use selective editing techniques.

Many selective editing methods are relatively simple ad-hoc methods based on common sense, although also complicated outlier detection techniques have been used in the context of selective editing (see Di Zio, Guarnera and Luzi, 2008). The most often applied basic idea is to use a score function (see e.g. Hidiroglou and Berthelot, 1986; Van de Pol and Molenaar, 1995). We distinguish two important components in order to construct a score function: the influence component and the risk component. The influence component measures the relative influence of a record on a publication figure. The risk component usually measures the deviation of the observed values from "anticipated" values. What suitable anticipated values are depends on the specific data set. For a cross-sectional survey one could, for instance, use means or medians in certain groups of records. For longitudinal surveys one could, for instance, use values from a previous period, possibly multiplied by an estimated trend. For some variables anticipated values may be obtained from available register data.

A score function for an entire record is referred to as a global score function. Such a global score function is often based on local score functions. A local score function is a score function for a single variable within a record. It is usually defined as a distance between observed and anticipated values of a variable y in the record under consideration, taking the influence of this record into account.

An example of a local score function is

$$w_i |y_i - \hat{y}_i|, \tag{2}$$

where y_i denotes the observed value of variable y in record i, $\hat{y}_i$ the corresponding anticipated value, and w_i the raising weight of record i. This local score function can be considered as the product of a risk component, $|y_i - \hat{y}_i|/\hat{y}_i$, which measures the relative deviation of the observed value to the anticipated value, and an influence component, $w_i\hat{y}_i$, which measures the anticipated impact on the publication figure.

A global score function combines the local scores to a measure on the record level, so one can decide whether to edit the record in an interactive manner or not. Local scores can be combined into a global score by, for instance, taking a (weighted) sum of the local scores or by taking the maximum of the local scores (see Subsection 5.2). A record is considered suspicious if the value of the global score function exceeds a certain cut-off value (see Subsection 5.3).

The local score function (2) is suited for simple estimators for population totals. In principle one can also develop local score functions for more complex estimators than simple estimators for totals. This can be done by linearization of the estimators, i.e. by taking the first order Taylor series. For more details we refer to Lawrence and McKenzie (2000).

5.2 Combining local scores into a global score

When combining several local scores into a global score, one first of all needs to take into account that different variables may have a different order of magnitude, or may be measured in different units. This problem can be overcome by scaling the variables. There are several options to scale variables, such as dividing the observed value by the mean value, by the standard error, or by the mean squared error of the variable under consideration (see Lawrence and McKenzie, 2000). From now on, whenever we refer to a local score, we will in fact mean the scaled local score.

The currently most general approach to combine local scores into a global score seems to be the use of the so-called Minkowski metric (see Hedlin, 2008). In our case the Minkowski metric is given by

$$GS_r(\mathbf{LS_r}, \alpha) = \left(\sum_{i=1}^{n} LS_{r,i}^{\alpha} \right)^{1/\alpha}, \tag{3}$$

where GS_r denotes the global score for a record r, $LS_{r,i} \geq 0$ the local score of the i-th variable, n the total number of variables, $\alpha > 0$ a parameter, and $\mathbf{LS_r} = (LS_{r,1}, ..., LS_{r,n})$. The choice of α in (3) determines how the local scores are actually combined into a global score. The influence of large local scores on the global score increases with α. For $\alpha = 1$ the global score is simply given by the sum of the local scores, and for $\alpha = 2$ the global score is the well-known Euclidean metric. For the limit where α goes to infinity, the global score $GS_r(\mathbf{LS_r}, \infty)$ becomes $\max_i LS_{r,i}$, i.e. the maximum of the n local scores.

The cut-off threshold value above which a record is considered to need interactive editing depends on the value of the α parameter and on the data set to be edited. Setting the cut-off threshold value is examined in Subsection 5.3.

The advantage of taking the maximum value of the local scores as the global score is that one is ensured that no influential error on any of the involved variables will slip through. Attempting to avoid the occurrence of influential errors in any of the involved variables may have the drawback that one has to edit many records interactively. Hedlin (2003), however, argues by using a model for the occurrence of influential errors in the data that this may not be the case. It depends on the data set to be edited how valid the assumptions underlying this model are.

The Minkowski metric is a flexible function that encompasses many well-known metrics used for selective editing purposes. By choosing the α parameter one can basically select a metric varying from taking the sum of all local scores to taking the maximum of the local scores. However, more complex metrics cannot be selected. Such a more complex metric may be deemed necessary if there are many variables with many complex interactions between them (see also Subsection 5.5). Presently no good generally applicable technique for combining local scores into a global score using such complex metrics seems available.

5.3 Determining cut-off thresholds

After a method to determine the global score has been decided upon, a cut-off threshold should, in principle, be fixed. All records with a global score above the cut-off threshold are selected for interactive editing, whereas the records with a global score below the cut-off threshold are not edited interactively.

The most common approach to set the cut-off threshold is to perform a simulation study, using a raw, unedited, data set and the corresponding clean, edited, version. These data sets are generally from a previous period. The simulation study consists of calculating the global scores of the records in the raw data set, and prioritizing the records in order of these scores. For several percentages p one then simulates that the first $p\%$ of these records are edited and the remaining records are not. This is done by replacing the first $p\%$ of the records in the prioritized raw data set with the corresponding records in the clean version of the data.

A natural criterion to determine the quality of a selective editing procedure in the simulation study approach is the absolute pseudo-bias (see Latouche and Berthelot, 1992), which measures the absolute deviation between the raw value and the clean value. The difference between the raw value and the clean value is called the pseudo-bias rather than the bias as one cannot be sure that the clean value is indeed the correct value. For the records selected for interactive editing, the corresponding absolute pseudo-bias is zero. Based on the simulation study, the cut-off threshold is selected so that the sum of the absolute pseudo-bias is acceptably low compared to other errors in the data, such as the sampling error and the coverage error.

As Lawrence and McKenzie (2000) argue, the simulation study approach is also a way to check the effectiveness of the edits and the editing process itself. The simulation study allows one, for instance, to check if the records with high global scores indeed contain influential errors.

In some cases the simulation study approach may not be applicable, for instance because data from a previous period are not available. Lawrence and McKenzie (2000) suggest using a model for the editing process in order to determine the cut-off threshold in such cases. Given that model one can then estimate the bias due to not editing some records as a function of the cut-off threshold. By specifying a maximum for the estimated bias, the corresponding cut-off threshold can be determined.

Lawrence and McKenzie (2000) use a relatively simple model for the editing process, but note that it can be extended to more complicated cases. An obvious drawback of the model-based approach is that is dependent on the model assumptions. Lawrence and McKenzie (2000) therefore propose to use the model-based approach only at the beginning of a new editing process to find a first cut-off threshold, and later use the results of the current editing process to improve this threshold by means of a simulation study.

In practice one sometimes does not fix a cut-off threshold before selective editing, but instead only uses the global scores to prioritize the records. One then edits the records in order of priority until budget or time constraints tell one to stop.

5.4 The edit-related approach

Hedlin (2003) proposes an edit-related approach to selective editing rather than the above sketched approach, which he refers to as estimate-related. The underlying idea of the edit-related approach is that influential errors will lead to violated edits. In the edit-related approach one measures how many edits are failed by a record and by how much they fail. For each edit, the amount of failure is measured in some way. For a balance edit one can, for instance, measure the amount of failure as the absolute difference between the observed total and the sum of its observed components. The amount of failure may be of very different orders for different (types of) edits. Hedlin (2003) therefore proposes using the Mahalanobis distance to combine amounts of failure into a global score per record.

The edit-related approach has the advantage that it does not focus on a single target variable. It also has the advantage that, unlike the estimate-related approach, it can be applied to categorical data. The edit-related approach has the drawback that it is dependent on the specified edits. In a study Hedlin (2003) found that the estimate-related approach performed better than the edit-related approach.

Hybrid approaches where an estimate-related approach is combined with an edit-related approach are also possible. For instance, Hoogland (2002) discusses such a hybrid approach that uses anticipated values to estimate the risk of a record and at the same time takes the violations of the edits as well as the number of missing values of that record into account.

5.5 Experimental approaches

At Statistics Netherlands some more advanced - experimental - approaches have been examined that to some extent try to capture complex interactions between different variables. One such approach is based on logistic regression. In this logistic regression approach one either tries to estimate the probability that a record contains influential errors or the probability that a specific variable in a record contains an influential error. In both cases, records or variables that are likely to contain influential errors need interactive editing.

We describe the case where we aim to estimate the probability π that a specific variable contains an influential error. For this, we need a training data set consisting of both the raw, i.e. unedited, data and the clean, i.e. edited, data from a previous period. To each record in the unedited data set we assign a probability π that this record contains an influential error. The assigned probability is high for records for which the edited version differs much from the raw version, and is low for records for which the edited version is close to the raw version. Based on the training data, we then fit a logistic model defined by

$$\log\left(\frac{\pi}{1-\pi}\right) = \beta_0 + \beta_1 x_1 + ... + \beta_p x_p, \tag{4}$$

where the x_1 to x_p are predictor variables, and β_0 to β_p the model parameters.

For a new data set that needs to be edited, one then uses the model given by (4) with model parameters estimated using the training data. For each record in the data set to be edited we hereby obtain an estimate $\hat{\pi}$ that the variable under consideration contains an influential error. If, for a certain record, $w\hat{\pi}$ is above a certain threshold value, where w is the raising weight of this record, the variable under consideration is considered to require to interactive editing for this record.

For the record-level case one can construct a model similar to (4), with the main difference that π denotes the probability that a record with attributes x_1 to x_p contains an influential error, and hence requires interactive editing.

Another approach that has been studied at Statistics Netherlands is the use of classification and regression trees (see Breiman et al., 1984) for selective editing. The idea of this approach is to grow a classification or regression tree that predicts the occurrence of an influential error in a record or a specific variable.

In general, a tree-based model classifies the data in terms of the values of a set of categorical predictor variables. A tree-based model is a binary tree that is generated by successively splitting a training data set into smaller subsets. These subsets are increasingly more homogeneous with respect to a selected response variable. This response variable may be either categorical or numerical. The process of recursively splitting the data set into two subsets continues until a stopping criterion is met. The terminal nodes in this tree form homogeneous clusters.

Homogeneity of a cluster can be measured in many ways. If the response variable is categorical, homogeneity may, for instance, be measured by the so-called Gini index (see Breiman et al., 1984). If the response variable is numerical, homogeneity may, for instance, be measured by ordinary least squares.

In the context of selective editing, there are several options for selecting the response variable. First of all, one has to choose whether one wants to generate a tree, either a classification tree or a regression tree, for a single variable or an entire record. Second, one has to choose between generating a classification tree or a regression tree. If one constructs a categorical response variable that expresses whether one considers a record or variable to need interactive editing or not, a classification tree has to be generated. If one constructs a numerical response variable that expresses the magnitude and impact on the publication figures of the errors in a record or variable, a regression tree has to be generated. By combining the possibilities, one obtains four different options, namely a classification tree for a single variable, respectively for an entire record, and a regression tree for a single variable, respectively an entire record.

To generate a classification or regression tree one again needs a training data set. In our case the training data set consists of a raw, unedited, data set from a previous period, together with information from the editing process applied to this data set. In the cases where the aim is to generate a classification tree we use the information whether a record is considered to require interactive editing, based on the changes that were made to this record during the editing process, as our response variable. In the cases where the aim is to generate a regression tree we use the magnitude and impact on the publication figures of the changes that were made to this record during the editing process as our response variable.

After the generation of a tree, and hence generation of classification rules for constructing homogeneous clusters of records, a data set to be edited is supplied to the tree. The tree is then used to decide whether a variable or record needs to be edited interactively (in the case of a classification tree) or to estimate the magnitude and impact of the error in a single variable or an entire record on the publication figures (in the case of a regression tree).

At Statistics Netherlands, Van Langen (2002) and Sanders (2002) have carried out limited evaluation studies for the logistic regression approach, respectively the tree-based approach. They both used a single data set of the Dutch Structural Business Statistics on the Construction Industry of which four versions were available: raw and clean versions for use as training data and other raw and clean versions for use as evaluation data. The simulation studies showed that in most cases the logistic regression approach and the tree-based approach performed worse than a traditional approach based on an estimate-related global score. An exception was the approach based on a regression tree for an entire record. This approach turned out to be slightly more powerful than a traditional approach based on an estimate-related global score. However, given the complexity of the method and the low transparency of the decision rules generated, the approach based on generating a regression tree for an entire record has thus far not been implemented in editing processes at Statistics Netherlands.

6 Automatic editing

6.1 Introduction to automatic editing

When automatic editing is applied, records are edited by computer without human intervention. In that sense automatic editing is the opposite of the traditional approach to the editing problem, where each record is edited manually. Automatic editing can be applied to both categorical and numerical data. In order to automate the SDE process both the error localization step and the imputation step have to be automated. In this section we focus on discussing the former step.

We can distinguish two kinds of errors: systematic ones and random ones. A systematic error is an error reported consistently by (some of) the respondents. It can be caused by the consistent misunderstanding of a question by (some of) the respondents. Examples are when gross values are reported instead of net values, and particularly when values are reported in units instead of, for instance, the requested thousands of units (so-called “thousand-errors”). Random errors are not caused by a systematic deficiency, but by accident. An example is an observed value where a respondent by mistake typed in a digit too many.

Systematic errors, such as thousand-errors, can often be detected by comparing a respondent’s present values with those from previous years, by comparing the responses to questionnaire variables with values of register variables, or by using subject-matter knowledge. Other systematic errors, such as transpositions of returns and costs and redundant minus signs, can be detected and corrected by systematically exploring all possible transpositions and inclusions/omissions of minus signs. Rounding errors - a class of systematic errors where balance edits are violated because the values of the involved variables have been rounded - can be detected by testing whether failed balance edits can be satisfied by slightly changing the values of the involved variables. Once detected, a systematic error is often simple to correct. We treat systematic errors in more detail in Subsection 6.2.

Generally speaking, we can subdivide the methods for automatic error localization of random errors into methods based on statistical models, methods based on deterministic checking rules, and methods based on solving a mathematical optimization problem. Methods based on statistical models, such as outlier detection techniques and neural networks (see Nordbotten, 1995, for one of the first attempts to apply neural networks in the context of SDE) are extensively discussed in the literature. We therefore do not discuss these techniques in this paper.

Deterministic checking rules state which variables are considered erroneous when the edits in a certain record are violated. An example of such a rule is: if component variables do not sum up to the corresponding total variable, the total variable is considered to be erroneous. Advantages of this approach are its transparency and its simplicity. A drawback of this approach is that many detailed checking rules have to be specified, which can be time and resources consuming to do. Another drawback is that maintaining and checking the validity of a high number of detailed checking rules can be complex. Moreover, in some cases it may be impossible to develop deterministic checking rules that are powerful enough to identify errors in a reliable manner. A final disadvantage is the fact that bias may be introduced as one aims to correct random errors in a systematic manner.

The automatic error localization problem for random errors can be formulated as a mathematical optimization problem in several ways. Freund and Hartley (1967) were among the first to propose such a formulation. It is based on minimizing the sum of the distance between the observed data and the “corrected” data and a measure for the violation of the edits. After “correction” some of the edits may still be failed. A second formulation, based on minimizing a quadratic function measuring the distance between the observed data and the “corrected” data subject to the constraint that the “corrected” data satisfy all edits, has later been proposed by Casado Valera et al. (1996).

A third approach for the automatic error localization problem for random errors is based on first imputing missing values and potentially erroneous values for an inconsistent record by means of hot-deck donor imputation, using a number of donor records. Subsequently, an imputed record that satisfies all edits and that is “closest” to the original record according to some distance function is selected. The values in the original record that differ from the corresponding values in the selected imputed record are considered to be erroneous. This paradigm forms the basis of NIM (Nearest-neighbour Imputation Methodology; see Bankier et al., 2000). Since the hot-deck donor imputation approach underlying NIM is much more suited for social data than for economic data, NIM has thus far mainly been used for demographic data. In some cases NIM has been used in combination with methodology based on the Fellegi-Holt paradigm, which is discussed below (see Manzari, 2004).

The most often used approach for the automatic error localization problem for random errors is based on the paradigm of Fellegi and Holt (see Fellegi and Holt, 1976). This paradigm is, in fact, only one of three principles for automatic edit and imputation proposed by Fellegi and Holt in 1976. These three principles are:

(1) the data in each record should be made to satisfy all edits by changing the fewest possible items of data (fields);

(2) as far as possible the frequency structure of the data file should be maintained;

(3) imputation rules should be derived from the corresponding edit rules without explicit specification.

In the context of error localization the first one of these principles is referred to as the "Fellegi-Holt paradigm". With regards to error localization it is the most important principle of the three. The other two principles relate to imputation after errors have been localized. In their second principle, which was originally formulated in the context of categorical data only, Fellegi and Holt basically note that imputation should result in the preservation of the distribution of the true data, and in their third principle that error localization and imputation should be applied in combination and not as completely separate processes.

In due course the Fellegi-Holt paradigm has been generalized to: the data of a record should be made to satisfy all edits by changing the values of the variables with the smallest possible sum of reliability weights. That is, for each record $(x_1, ..., x_n)$ we wish to ensure the existence of, and often want to find, a synthetic record $(\hat{x}_1, ..., \hat{x}_n)$ such that $(\hat{x}_1, ..., \hat{x}_n)$ satisfies all edits, and

$$\sum_{k=1}^{n} w_k \delta\left(x_k, \hat{x}_k\right) \tag{5}$$

is minimized, where $\delta\left(x_k, \hat{x}_k\right)$ equals 1 if x_k is missing or differs from $\hat{x}_k$ and 0 otherwise, and $w_k \geq 0$ is the so-called reliability weight of the k-th variable $(k = 1, ..., n)$. A reliability weight of a variable expresses how reliable one considers the values of this variable to be. A high reliability weight corresponds to a variable of which the values are considered trustworthy, a low reliability weight to a variable of which the values are considered not so trustworthy. The (generalized) Fellegi-Holt paradigm can be applied to numerical data as well as to categorical data.

A variable k $(k = 1, ..., n)$ for which x_k is missing or differs from $\hat{x}_k$ is considered to be erroneous. Such a variable has to be imputed later using a suitable imputation method. The existence of a synthetic record satisfying all edits ensures that the variables considered erroneous can indeed be imputed consistently, i.e. such that all edits can be satisfied. The synthetic record is generally not used as the "corrected" record.

For business surveys, an overview of algorithms for solving the error localization problem based on the Fellegi-Holt paradigm has been given in De Waal and Coutinho (2005). Algorithms for categorical data have been proposed by Fellegi and Holt (1976), Garfinkel, Kunnathur and Liepins (1986), Winkler (1998), Bruni, Reale and Torelli (2001), and Bruni and Sassano (2001). The algorithms described in the first two articles have been examined in detail by Boskovitz (2008). Algorithms for solving the error localization problem in a mix of categorical and continuous data have been proposed by Sande (1978), Schaffer (1987), De Waal (2003a, 2003b and 2005), and De Waal and Quere (2003). This latter algorithm is sketched and illustrated in Subsections 6.3 and 6.4.

Software packages for automatic editing of continuous data include Banff (see Banff Support Team, 2003) and its predecessor GEIS (see Kovar and Whitridge, 1990), SPEER (Winkler and Draper, 1997), AGGIES (see Todaro, 1999), a SAS program developed by the Central Statistical Office of Ireland (see Central Statistical Office, 2000), and CherryPi (see De Waal, 1996).

Software packages for automatic editing of categorical data include DIA (see Garcia Rubio and Criado, 1990), SCIA (see Barcaroli et al., 1995) and DISCRETE (see Winkler and Petkunas, 1997).

CANCEIS (see Bankier, 2003) and SLICE (see De Waal, 2001) can automatically edit a mix of categorical and numerical data, although the focus of CANCEIS is on categorical data and that of SLICE on numerical data.

Assuming that only few errors are made, the Fellegi-Holt paradigm obviously is a sensible one. Provided that the set of edits used is sufficiently powerful, application of this paradigm generally results in data of higher statistical quality, especially when used in combination with other editing techniques. This is confirmed by various evaluation studies such as Hoogland and Van der Pijll (2003).

A drawback of using the Fellegi-Holt paradigm is that the class of errors that can safely be treated is limited to random errors. A second drawback is that the class of edits that can be handled is restricted to "hard" edits. "Soft" edits cannot be handled as such, and - if specified - are treated as hard edits. Especially in the case of automatic editing one should be careful not to specify too many soft edits in order to avoid over-editing (see Di Zio, Guarnera and Luzi, 2005a).

6.2 Approaches to automatic editing of systematic errors

As already mentioned, a well-known class of systematic errors consists of so-called thousand-errors. These are cases where a respondent replied in units rather than in the requested thousands of units. The usual way to detect such errors is by considering "anticipated" values, which could, for instance, be values of the same variable from a previous period or values available from a register. One then calculates the ratio of the observed value to the anticipated one. If this ratio is higher than a certain threshold value, say 300, it is assumed that the observed value is a factor 1,000 too high. The observed value is then corrected by dividing it by 1,000. A minor practical problem occurs when the anticipated value equals zero. Usually this problem can easily be solved in practice.

Al-Hamad, Lewis and Silva (2008) note more important problems with this standard procedure. The main problem they note is that the anticipated value itself has to be of sufficiently high quality. If this value is incorrect, the procedure may not detect a thousand-error in the observed value if it is present.

They propose an alternative procedure, which simply consists of comparing the number of digits of the observed value to the anticipated value. In this way thousand-errors (and larger errors) may be identified as those records for which the difference between the number of digits of the observed and anticipated value is 3 or more. In a study they found that this alternative rule slightly outperformed the standard rule based on taking ratios.

A more complex approach for detecting and correcting thousand-errors, or more generally unity measure errors, i.e. any error due to the erroneous choice by some respondents of the unity measure in reporting the amount of a certain variable, has been proposed by Di Zio, Guarnera and Luzi (2005b). That approach uses model-based cluster analysis to pinpoint various kinds of unity measure errors. The model applied consists of a finite mixture of multivariate normal distributions.

A second kind of systematic error that can relatively easily be corrected occurs when a respondent adds a minus sign to a value that is subtracted. The questionnaire of the Dutch Structural Business Survey (SBS) contains a number of combinations of items where costs have to be subtracted from returns to obtain a balance. If a respondent adds a minus sign to the reported costs, the value becomes wrongfully negative after data processing. Such an error where a respondent by mistake adds or deletes a minus sign is called a sign error. An obvious way to correct a sign error is by taking the absolute value of the reported value.

The situation becomes more complicated when a respondent may also have interchanged returns and costs on the questionnaire. Scholtus (2008a and 2008b) examines this situation. Part of the Dutch SBS is the so-called results block. In this block of related questions a respondent has to fill in a number of balance amounts. We denote the balance variables by x_0, x_1, ..., x_{n-1}. The so-called pre-tax result is denoted by x_n and equals the sum of x_0 to x_{n-1}, i.e.

$$x_0 + x_1 + ... + x_{n-1} = x_n \tag{6}$$

Some of these balance variables are equal to the difference between a returns variable and a costs variable. That is,

$$x_{k,r} - x_{k,c} = x_k, \tag{7}$$

where $x_{k,r}$ denotes the returns variable and $x_{k,c}$ the corresponding costs variable of the k-th balance restriction.

We give a simple example of sign errors and interchanged returns and costs. To this end, we consider a record with the following values: $x_{0,r} = 3,250$, $x_{0,c} = 3,550$, $x_0 = 300$, $x_{1,r} = 110$, $x_{1,c} = 10$, $x_1 = 100$, $x_{2,r} = 50$, $x_{2,c} = 90$, $x_2 = 40$, $x_{3,r} = 30$, $x_{3,c} = 10$, $x_3 = 20$, and $x_4 = -140$. This record has to satisfy (6) with $n = 4$ and (7) for $k = 0, 1, 2, 3$. The record can be made to satisfy all edits by changing the value of x_0

from 300 to -300 and interchanging the values of $x_{2,r}$ and $x_{2,c}$. These are likely to be the correct values as this is the only way to make the record satisfy all edits by means of such simple and natural modifications.

Assuming that if an inconsistent record can be made to satisfy all balance edits (6) and (7) by adding/deleting minus signs and interchanging returns and costs this is indeed the way the record should be corrected, Scholtus (2008a and 2008b) provides a formulation for correcting a record as a binary linear programming problem. Well-known operations research techniques can be applied to find a solution to this problem.

Assuming that the variables $x_{0,r}$ and $x_{0,c}$, which incidentally are the so-called operating returns, respectively operating costs in the case of the Dutch SBS, are not interchanged, Scholtus (2008b) proves that if a solution is found it is the unique solution under some mild additional conditions.

Balance edits are often violated by the smallest possible difference. That is, the absolute difference between the total and the sum of its components is equal to 1 or 2. Such inconsistencies are often caused by rounding. An example is when the terms of the balance edit $x_1 + x_2 = x_3$ with $x_1 = 2.7$, $x_2 = 7.6$ and $x_3 = 10.3$ are rounded to integers. If conventional rounding is used, x_1 is rounded to 3, x_2 to 8 and x_3 to 10, and the balance edit becomes violated.

From a purely statistical point of view, rounding errors are rather unimportant as by their nature they have virtually no influence on publication figures. Rounding errors may be important, however, when we look at them from the point of view of the SDE *process*. Some statistical offices apply automatic editing procedures for random errors, such as automatic editing procedures based on the Fellegi-Holt paradigm. Such automatic editing procedures are computationally very demanding. The complexity of the automatic error localization problem increases rapidly as the number of violated edit rules becomes larger, irrespective of the magnitude of these violations. A record containing many rounding errors may hence be too complicated to solve for an automatic editing procedure for random errors, even if the number of random errors is actually low. From the point of view of the SDE process it may therefore be advantageous to resolve rounding errors at the beginning of the editing process.

Scholtus (2008a and 2008b) describes a heuristic method for resolving rounding errors. The method does not lead to solutions that are "optimal" according to some criterion, such as that the number of changed variables or the total change in value is minimized. Instead the method just leads to a good solution. Given that the statistical impact of resolving rounding errors is small, a time-consuming and complex algorithm aimed at optimizing some target function is not necessary anyway. The heuristic method is referred to as the "scapegoat algorithm", because for each record assumed to contain rounding errors a number of variables, the "scapegoats", are selected beforehand and the rounding errors are resolved by changing only the values of the selected variables.

Under certain very mild conditions, the algorithm guarantees that exactly one choice of values exists for the selected variables such that the balance edits become satisfied. Different variables are selected for each record to minimize the effect of the adaptations on published aggregates.

In general, the obtained solution might contain fractional values, whereas most business survey variables are restricted to be integer-valued. If this is the case, a controlled rounding algorithm could be applied to the values to obtain an integer-valued solution (see e.g. Salazar-González et al., 2004). Under certain additional mild conditions, which appear to be satisfied by most data sets arising in practice, the problem of fractional values does not occur, however. For details we refer to Scholtus (2008a and 2008b).

Rounding errors often occur in combination with other "obvious" systematic errors. For instance, a sign error might be obscured by the presence of a rounding error. Scholtus (2008a and 2008b) provides a single mathematical model for detecting sign errors and rounding errors simultaneously.

6.3 Example of a Fellegi-Holt based algorithm

In this subsection we sketch an algorithm based on the Fellegi-Holt paradigm in order to illustrate how such algorithms work. We will first describe the algorithm for numerical data, and later describe how the algorithm can be adapted to categorical data. In Subsection 6.4 the algorithm is illustrated by means of an example.

The basic idea of the algorithm we describe in this section is that for each record a binary tree is constructed. In our case, we use a binary tree to split up the process of searching for solutions to the error localization problem. We need some terminology with respect to binary trees before we can explain our algorithm. Following Cormen, Leiserson and Rivest (1990), we recursively define a binary tree as a structure on a finite set of nodes that either contains no nodes, or comprises three disjoint sets of nodes: a root node, a left (binary) subtree and a right (binary) subtree. If the left subtree is non-empty, its root node is called the left child node of the root node of the entire tree, which is then called the parent node of the left child node. Similarly, if the right subtree is non-empty, its root node is called the right child node of the root node of the entire tree, which is then called the parent node of the right child node. All nodes except the root node in a binary tree have exactly one parent node. Each node in a binary tree can have at most two (non-empty) child nodes. A node in a binary tree that has only empty subtrees as its child nodes is called a terminal node, or also a leaf. A non-leaf node is called an internal node. In each internal node of the binary tree generated by our algorithm a variable is selected that has not yet been selected in any predecessor node. If all variables have already been selected in a predecessor node, we have reached a terminal node of the tree.

We first assume that no values are missing. After the selection of a variable two branches, i.e. subtrees, are constructed: in one branch we assume that the observed value of the selected variable is correct, in the other branch we assume that the observed value is incorrect. By constructing a binary tree we can, in principle, examine all possible error patterns and search for the best solution to the error localization problem.

In the branch in which we assume that the observed value is correct, the variable is fixed to its original value in the set of edits. In the branch in which we assume that the observed value is incorrect, the selected variable is eliminated from the set of edits. A variable that has either been fixed or eliminated is said to have been treated (for the corresponding branch of the tree). To each node in the tree we have an associated set of edits for the variables that have not yet been treated in that node. The set of edits corresponding to the root node of our tree is the original set of edits.

Eliminating a variable is non-trivial as removing a variable from a set of edits may imply additional edits for the remaining variables. To illustrate why edits may need to be generated, we give a very simple example. Suppose we have three variables x_1, x_2 and x_3, and two edits: $x_1 \leq x_2$ and $x_2 \leq x_3$. If we want to eliminate variable x_2 from these edits, we cannot simply delete this variable and the two edits, but have to generate the new edit $x_1 \leq x_3$ implied by the two old ones for else we could have that $x_1 > x_3$ and the original set of edits cannot be satisfied.

To ensure that the original set of edits can be satisfied Fourier-Motzkin elimination is used. For inequalities Fourier-Motzkin elimination basically consists of using the variable to be eliminated to combine these inequalities pairwise (if possible), as we did in the above example. If the variable to be eliminated is involved in a balance edit, we use this equation to express this variable in terms of the other variables and then use this expression to eliminate the variable from the other edits.

In each branch of the tree the set of current edits is updated. Updating the set of current edits is the most important aspect of the algorithm. How the set of edits has to be updated depends on whether the selected variable is fixed or eliminated. Fixing a variable to its original value is done by substituting this value in all current edits, failing as well as non-failing ones. Conditional on fixing the selected variable to its original value, the new set of current edits is a set of implied edits for the remaining variables in the tree. That is, conditional on the fact that the selected variable has been fixed to its original value, the remaining variables have to satisfy the new set of edits. As a result of fixing the selected variable to its original value some edits may become tautologies, i.e. may become satisfied by definition. An example of a tautology is "$1 \geq 0$". Such a tautology may, for instance, arise if a variable x has to satisfy the edit $x \geq 0$, the original value of x equals 1, and x is fixed to its original value. These tautologies may be discarded from the new set of edits. Conversely, some edits may become self-contradicting relations. An example of a self-contradicting relation is:

"$0 \geq 1$". If self-contradicting relations are generated, this particular branch of the binary tree cannot result in a solution to the error localization problem. Eliminating a variable by means of Fourier-Motzkin elimination amounts to generating a set of implied edits that do not involve this variable. This set of implied edits has to be satisfied by the remaining variables. In the generation process we need to consider all edits, both the failing edits as well as the non-failing ones, in the set of current edits pairwise. The generated set of implied edits plus the edits not involving the eliminated variable become the set of edits corresponding to the new node of the tree.

If values are missing in the original record, the corresponding variables only have to be eliminated from the set of edits (and not fixed).

After all variables have been treated we are left with a set of relations involving no unknowns. If and only if this set of relations contains no self-contradicting relations, the variables that have been eliminated in order to reach the corresponding terminal node of the tree can be imputed consistently, i.e. such that all original edits can be satisfied (cf. Theorems 1 and 2 in De Waal and Quere, 2003). The set of relations involving no unknowns may be the empty set, in which case it obviously does not contain any self-contradicting relations. In the algorithm we check for each terminal node of the tree whether the variables that have been eliminated in order to reach this node can be imputed consistently. Of all the sets of variables that can be imputed consistently we select the ones with the lowest sum of reliability weights. In this way we find all optimal solutions to the error localization problem (cf. Theorem 3 in De Waal and Quere, 2003).

For categorical data, the algorithm is essentially the same as the above-described algorithm. The only difference between the algorithms for the two data types is the way in which variables are eliminated, and hence the way in which implied edits are generated. As we have mentioned above, when a variable in numerical data is eliminated, we pairwise apply Fourier-Motzkin elimination. For the case of categorical data, when a variable is to be eliminated we apply the method originally proposed by Fellegi and Holt (1976) to generate implied edits, using the variable to be eliminated as the so-called generating field.

We again denote the number of variables by n. Furthermore, we denote the domain, i.e. the set of all allowed values of a variable i by D_i. In the case of categorical data, an edit j is usually written in so-called *normal form*, i.e. as a collection of sets F_i^j $(i = 1, 2, ..., n)$:

$$(F_1^j, F_2^j, ..., F_n^j), \tag{8}$$

meaning that if for a record with values $(v_1, v_2, ..., v_n)$ we have $v_i \in F_i^j$ for all $i = 1, 2, ..., n$, then the record fails edit j, otherwise the record satisfies edit j. For instance, suppose we have three variables: *Marital status*, *Age* and *Relation to head of household*. The possible values of *Marital status* are Married, Unmarried, Divorced and Widowed, of *Age* "< 16 years" and "≥ 16 years", and of *Relation to head of household* Spouse, Child, and Other. The edit that someone who is less than 16 years cannot be married, can be written in normal form as

$$(\{\text{Married}\}, \{< 16 \text{ years}\}, \{\text{Spouse, Child, Other}\}) \tag{9}$$

The edit that someone who is not married cannot be the spouse of the head of household can be written as

$$(\{\text{Unmarried, Divorced, Widowed}\}, \{< 16 \text{ years}, \ \geq 16 \text{ years}\}, \{\text{Spouse}\}). \tag{10}$$

Note that whereas for numerical data an edit is *satisfied* if a certain condition is fulfilled, for categorical data an edit is *violated* if a certain condition is fulfilled. We assume that all edits are written in the form (8). We call a categorical variable i involved in an edit given by (8) if $F_i^j \neq D_i$.

Now, if we eliminate a variable v_r we start by determining all index sets S such that

$$\bigcup_{j \in S} F_r^j = D_r \tag{11}$$

and

$$\bigcap_{j \in S} F_i^j \neq \emptyset \qquad \text{for all } i = 1, ..., r-1, r+1, ..., n. \tag{12}$$

From these index sets we select the *minimal* ones, i.e. the index sets S that obey (11) and (12), but none of their proper subsets obey (11).

Given such a minimal index set S we construct the implied edit

$$(\bigcap_{j \in S} F_1^j, ..., \bigcap_{j \in S} F_{r-1}^j, D_r, \bigcap_{j \in S} F_{r+1}^j, ..., \bigcap_{j \in S} F_n^j). \tag{13}$$

For example, if we eliminate variable *Marital status* from the edits (9) and (10), we obtain the implied edit

$$(\{\text{Married, Unmarried, Divorced, Widowed}\}, \{< 16 \text{ years}\}, \{\text{Spouse}\}),$$

which expresses that someone who is less than 16 years of age cannot be the spouse of the head of household.

Note that variable v_r is not involved in the edit (13). By adding the implied edits resulting from all minimal sets S to the set of edits, and removing all edits involving the eliminated variable, one obtains the updated set of current edits.

The algorithms for numerical and categorical data can be combined into a single algorithm for numerical and categorical data (see De Waal and Quere, 2003). That algorithm can be further extended to deal with integer-valued data in a heuristic manner (see De Waal, 2005).

6.4 Illustration of the Fellegi-Holt based algorithm

In this subsection we illustrate the algorithm for numerical data described in Subsection 6.3 by means of an example. Suppose the explicit edits are given by

$$T = P + C \tag{14}$$

$$P \leq 0.5T \tag{15}$$

$$-0.1T \leq P \tag{16}$$

$$T \geq 0 \tag{17}$$

$$T \leq 550N \tag{18}$$

where T denotes the turnover of an enterprise, P its profit, C its costs, and N the number of employees. Let us consider a specific erroneous record with values $T = 100$, $P = 40,000$, $C = 60,000$ and $N = 5$. Edits (16) to (18) are satisfied, whereas edits (14) and (15) are violated. The reliability weights of the variables T, P and C equal 1, and the reliability weight of variable N equals 2. As edits (14) and (15) are violated, the record contains errors.

We select a variable, say T, and construct two branches: one where T is eliminated and one where T is fixed to its original value. We consider the first branch, and eliminate T from the set of edits. We obtain the following edits.

$$P \leq 0.5(P + C) \tag{19}$$

$$-0.1(P + C) \leq P \tag{20}$$

$$P + C \geq 0 \tag{21}$$

$$P + C \leq 550N \tag{22}$$

Edits (19) to (21) are satisfied, edit (22) is violated. Because edit (22) is violated, changing T is not a solution to the error localization problem. If we were to continue examining the branch where T is eliminated by eliminating and fixing more variables, we would find that the best solution in this branch has an objective value (5) equal to 3. We now consider the other branch where T is fixed to its original value. We fill in the original value of T in edits (14) to (18), and obtain (after removing any tautology that might arise) the following edits:

$$100 = P + C \tag{23}$$

$$P \leq 50 \tag{24}$$

$$-10 \leq P \tag{25}$$

$$100 \leq 550N \tag{26}$$

Edits (25) and (26) are satisfied, edits (23) and (24) are violated. We select another variable, say P, and again construct two branches: one where P is eliminated and one where P is fixed to its original value. Here, we only examine the former branch, and obtain the following edits (again after removing any tautology that might arise):

$$100 - C \leq 50$$

$$-10 \leq 100 - C \tag{27}$$

$$100 \leq 550N \tag{28}$$

Only edit (27) is violated. We select variable C and again construct two branches: one where C is eliminated and another one where C is fixed to its original value. We only examine the former branch, and obtain edit (28) as the only implied edit. As this edit is satisfied by the original value of N, changing P and C is a solution to the error localization problem. By examining all branches of the tree, including the ones that we have skipped here, we find that this is the only optimal solution to this record.

7 Macro-editing

7.1 Introduction to macro-editing

Thus far, we have examined micro-editing methods, i.e. methods that use the data of a single record and related auxiliary information to check and correct it. In this section we examine macro-editing methods. Macro-editing techniques often examine the potential impact on survey estimates to identify suspicious data in individual records. Macro-editing can lead to the detection of errors that would go unnoticed with selective editing or automatic editing. Micro-editing and macro-editing are complementary. Errors that are apparent from one point of view may not be apparent from the other. For instance, micro-editing may reveal more errors than macro-editing, but macro-editing may trace bigger, more influential errors.

Macro-editing can be seen as a form of selective editing. A major difference between macro-editing and selective editing is the moment at which they are applied in the SDE process. Whereas selective editing can be used early in the SDE process while a substantial part of the data to be edited may still be collected, macro-editing is used at the end of the SDE process when (most of) the data have already been collected. This allows a different approach. Whereas selective editing basically treats each record to be edited separately, macro-editing treats the data set to be edited as a whole. Selective editing checks whether each record to be edited is plausible; macro-editing checks whether the data set as a whole is plausible.

We distinguish between two forms of macro-editing. The first form is called the aggregation method (see, e.g., Granquist, 1990 and 1995). 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, with quantities based on register data, or with related quantities from other sources. Examples of this form of macro-editing are the foreign trade surveys of the Netherlands (see Van de Pol and Diederen, 1996) and Canada (see Laflamme et al., 1996). Only if an unusual quantity is observed, a micro-editing procedure is applied to the individual records and fields contributing to this quantity. An unusual quantity may, for instance, be detected by checking whether

$$\left|\frac{Y-\hat{Y}}{\hat{Y}}\right| > p/100, \tag{29}$$

where Y denotes a publication figure to be checked, $\hat{Y}$ an “anticipated” value for this publication figure, and p a certain percentage. If (29) holds true, i.e. if Y deviates more than $p\%$ from its anticipated value, the microdata underlying the publication figure Y are subjected to a micro-editing procedure.

Generally in software packages for macro-editing (for some software packages, see, for instance, Houston and Bruce, 1993; Esposito et al, 1994; Esposito, Lin and Tidemann, 1997; Weir, Emery and Walker, 1997; Engström, and Ängsved, 1997; De Waal, Renssen and Van de Pol, 2000), the influence of individual observations on population figures is estimated. Starting from the most influential observation, individual data can be interactively checked and corrected, raising weights can be adjusted, or records can be removed all together. The interactive editing process terminates when further corrections have a negligible effect on the estimated population figures. The impact of such changes to the data on estimates of publication figures can be monitored by re-estimating the publication figures each time a change has been made.

A second form of macro-editing is the distribution method. Here the available data, either the data set to be edited or a reference data set, are used to characterize the distribution of the variables. Next, all individual values are compared with this 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 correction. The distribution method is examined in more detail in Subsection 7.2.

7.2 Exploratory Data Analysis and related techniques

There is an area in statistics providing all kinds of techniques for analyzing the distribution of variables, namely Exploratory Data Analysis (EDA) (see, e.g., Tukey, 1977). Many EDA techniques can be applied in macro-editing. Advocates of EDA stress the importance of the use of graphical techniques (see, e.g. Chambers, 1983). These techniques can provide much more insight in the behavior of variables than numerical techniques do. Graphs of the distribution of the data show a lot of information, and often are capable of showing unexpected properties that would not have been discovered if just numerical quantities were computed.

The application of EDA techniques for data editing has been the subject of a number of papers. DesJardins (1997) gives a description of how several EDA techniques can be used during the data editing stage. The techniques range from traditional EDA techniques, such as boxplots and scatterplots, to more advanced techniques, such as so-called 6D-plots and "industry plots". Industry plots have been devised by DesJardins to present a comprehensive overview of an entire survey in a single graph. An industry plot attempts to depict the multivariate relation between the key variables of each individual company. In such an industry plot the "normal" companies are clustered around the center point of the plot, whereas outlying companies lie far from the center point. Once an industry plot has been designed it is a powerful tool to quickly detect outliers. However, designing an industry plot seems a non-trivial task. Another drawback of industry plots is that they have to be redesigned for each type of survey.

Bienas et al. (1997) describe the application of graphical EDA techniques to identify potentially incorrect data in two different surveys. The EDA techniques that were applied were boxplots, scatterplots and bivariate fitting. Transformations, such as taking logarithms, were applied to the data to discern patterns more easily. The fitting methods that were applied were ordinary least squares and resistant regression, which reduces the influence of outlying cases on the fit of the regression model. Ordinary least squares fitting proved very useful when there are only a few unusual records that can be easily distinguished from the usual records. In the case that there are relatively many outlying records, resistant fitting proved to be more useful. Bienas et al. (1997) mention that the EDA approach can be combined with batch-type micro-editing.

Frequently used techniques in software packages for macro-editing are so-called anomaly plots, time series analysis, outlier detection methods, and the already mentioned EDA techniques such as boxplots and scatterplots. Anomaly plots are graphical overviews of the important estimates, where unusual estimates are highlighted. Once suspicious data have been detected on a macro-level in such an anomaly plot one can usually drill-down to sub-populations and individual records. Outlying records can often be identified by means of graphical EDA techniques. In particular, scatterplots comparing the data in the current period to the corresponding data in a previous period can often be used. Also, the concept of linked plots, where an outlier in one plot is automatically also highlighted as outlier in other plots, helps the analyst to study the behavior of an outlier in one plot in other plots. Besides graphical EDA techniques, software packages for macro-editing sometimes also offer a mathematical (multivariate) outlier detection algorithm for identifying outlying records.

7.3 Possibilities and pitfalls of macro-editing

One may wonder whether the application of macro-editing approaches will result in micro-data of less statistical quality than would have been obtained after exhaustive micro-editing. Data users who consider their applications more “micro” than the usual publication figures of totals and means often have to be convinced that macro-editing approaches are not harmful, especially for multivariate micro analysis. A reassuring point for these users is that so-called micro level analysis does not actually involve the inspection of individual records. Multivariate analysis brings along the estimation of parameters which always are some sort of an aggregate. For instance, the estimation of output elasticities for energy, labor and material from annual construction survey data, turned out to differ less than one standard deviation when comparing results after no data editing, selective data editing and exhaustive data editing (see Van de Pol and Bethlehem, 1997).

Another point to convince skeptic data users that application of macro-editing techniques does not lead to a loss of quality is that all methods of data editing, including the traditional exhaustive micro-editing approach, will leave some errors unnoticed and not corrected, because not all errors are apparent. In case of over-editing, some other fields will be changed without good justification. Data editors are human, which means that they make errors and miss errors from time to time. This will occur less often when they have good tools to navigate in the data and to distinguish between important and unimportant errors. Because multivariate methods often are sensitive to outliers, data editing methods that trace these outliers, such as macro-editing techniques, should be welcomed.

Despite these points in favor of macro-editing, changing from an exhaustive micro-editing approach to a much less exhaustive macro-editing approach is a big step for many potential users. They have to be convinced that the application of a macro-editing approach can result in data of sufficiently high quality. In the words of DesJardins (1997): “Introducing graphical EDA can be a tough nut”.

Graphical macro-editing certainly offers a lot of possibilities, but unfortunately there are some problems and pitfalls one should be aware of when applying this approach, and before deciding to develop a software tool for macro-editing.

A limitation of macro-editing, at least in the applications known to us, is that it is much more suited for editing of economic data than of social data. A drawback of macro-editing is that the time and resources required for editing are hard to predict. A further drawback is that one needs to wait with the macro-editing process until all or most of the data have arrived and are ready for processing.

Persons can interpret data that are depicted in several scatterplots simultaneously, so graphical macro-editing allows one to edit relatively large amounts of data simultaneously. There is also a limit, however. It is impossible for (most) human beings to interpret, say, 10 scatterplots at the same time. For a data set with many important key variables graphical macro-editing is usually not the most suitable editing method, unless applied in combination with other SDE methods.

A very important methodological drawback of the aggregation method is that this approach involves the risk that records contributing to publication figures that are considered non-suspicious still do contain influential errors, errors that were not detected and corrected. This will lead to biased publication figures. Relying fully on macro-editing may also prevent publication of unexpected, but true changes in trend. Outliers in one direction may be removed until outliers in the opposite direction cancel out the unexpected trend.

For more on the possibilities and pitfalls of macro-editing, we refer to De Waal, Renssen and Van de Pol (2000).

7.4 Macro-editing versus micro-editing

An advantage of macro-editing in comparison to micro-editing is that micro-editing, either automatically or interactively, requires edits. Specifying edits, for instance the bounds of ratio-edits, can be difficult and time-consuming. Of course, one does not want to specify edits that are too lenient in the sense that influential incorrect data are not detected. On the other hand, one also does not want to specify edits that are too severe in the sense that many correct, or only slightly incorrect, records are considered suspicious because this would result in over-editing. So, not having to specify edits clearly has its benefits.

Although not having to specify edits is one of the advantages of macro-editing, it is at the same time also a bit dangerous. When edits are specified and the micro-editing approach is used, it is clear when a record will be considered suspicious. When the macro-editing approach is used and edits are not specified, it is for a substantial part left to the subject-matter specialists who do the editing to decide which records are suspicious and which are not. That is, it will depend on the partly subjective judgment of the subject-matter specialists how the records are divided into suspicious and non-suspicious records.

Automatic editing and imputation can be implemented in such a way that the results can be reproduced, i.e. if the same data set is edited and imputed again the same results are obtained. This is not the case for interactive editing and macro-editing. The results of interactive editing and macro-editing are partly subjective, i.e. they partly depend on the specific subject-matter specialist who edits the data. Different subject-matter specialists, or even the same subject-matter specialist at different moments in time, may obtain different results.

When the incoming raw data contain many errors, i.e. when almost every record needs correction, micro-editing is more efficient than macro-editing. In that case the extra effort to trace erroneous records from a macro point of view should be postponed until the data set has a reasonably good quality due to micro-editing.

An argument for maintaining some sort of micro-editing is that this is the only way to make sure that records are internally consistent, i.e. that they satisfy the edits. Also, automatic correction of "obvious" systematic errors should always be done before macro-editing in our opinion. This form of micro-editing is not costly and can improve the estimates of aggregates and distributions used in the macro-editing phase.

8 A strategy for statistical data editing

In this section we propose a strategy for SDE in which we combine the editing techniques described in the previous sections. We assume that a data set to be edited has already been collected. Our proposed strategy depends on whether the data are numerical or categorical. We start with our strategy for numerical data. For these data we advocate an SDE approach that consists of the following phases:

(1) correction of "obvious" (systematic) errors, such as thousand-errors, sign errors, interchanged returns and costs, and rounding errors;

(2) application of selective editing to split the records in a critical stream and a non-critical stream;

(3) editing of the data: the records in the critical stream are edited interactively, the records in the non-critical stream are edited automatically;

(4) validation of the publication figures by means of macro-editing.

The above steps are used at Statistics Netherlands in the production process for structural business statistics (see De Jong, 2002). The goal of the first phase is to treat errors that are obviously errors and that once detected are also easy to correct. Typically, "obvious" systematic errors, such as thousand-errors, are dealt with in this phase. The main goal of the second phase is to select the influential errors. In the third phase these errors are treated interactively. Most influential errors will be resolved by the subject-matter specialists; in some cases the respondents will be re-contacted. In the third phase also non-influential errors are treated. As these errors often occur in a high number of records, they have to be detected and corrected as efficiently as possible, both in terms of budget and time. Automatic editing is hence the most often used way to handle non-influential errors. The fourth phase, the validation phase, is performed by subject-matter specialists, who use macro-editing to compare the publication figures based on the edited data to publication figures from a previous year, for instance. In this final step the focus is more on the overall results than on the correctness of individual records. An additional goal of macro-editing is to check whether the SDE process itself has functioned well.

One could argue that with selective editing the automatic editing step is superfluous. Personally, we advocate the use of automatic editing, even when selective editing is used. We mention three reasons. First, the sum of the errors of the records in the non-critical stream may have an influential effect on the publication figures, even though each error itself may be non-influential. This can in particular be the case if the data contain systematic errors as then a substantial part of the data may be biased in the same direction. The correction of "obvious" systematic errors evidently leads to data of higher statistical quality. In addition, provided that the set of edits used is sufficiently powerful, application of the Fellegi-Holt paradigm also generally results in data of higher statistical quality. Second, many non-critical records will be internally inconsistent, i.e. they will fail specified edits, if they are not edited, which may lead to problems when publication figures are calculated or when micro-data are released to external researchers. Finally, automatic editing provides a mechanism to check the quality of the selective editing procedures. If selective editing is well-designed and well-implemented, the records that are not selected for interactive editing need no or only slight adjustments. Records that are substantially changed during the automatic editing step therefore possibly point to an incorrect design or implementation of the selective editing step.

Phases 2 and 4 of our strategy for numerical data do not, or hardly, apply to categorical data. For those data our proposed strategy simply consists of checking and correcting errors, first obvious ones as in phase 1 and later more complex ones as in phase 3, as much as possible automatically.

We feel that a combined approach using - if applicable - selective editing, interactive editing, automatic editing and macro-editing can improve the efficiency of the traditional interactive SDE process while at the same time maintaining or even enhancing the statistical quality of the produced data.

9 Discussion

In this paper we have focused on identifying errors in the data as this has traditionally been considered the most important aim of SDE in practice. In fact, however, this is only one of the goals of SDE. Granquist (1995) identifies the following main goals of SDE:

(1) identify error sources in order to provide feedback on the entire survey process;

(2) provide information about the quality of the incoming and outgoing data;

(3) identify and treat influential errors and outliers in individual data;

(4) when needed, provide complete and consistent individual data.

During the last few years, the first two goals - providing feedback on the other survey phases, such as the data collection phase, and providing information on the quality of the collected data and the final results - have gained in importance. The feedback on other survey phases can be used to improve those phases and reduce the amount of errors arising in these phases. SDE forms part of the entire statistical process at NSIs. A direction for potential future research is hence the relation between SDE and other steps of the statistical process, such as data collection (see, e.g., Børke, 2008) and statistical disclosure control (see Shlomo and De Waal, 2008). In the next few years the first two goals of SDE are likely to become even more important.

From our discussion of SDE the reader may have gotten the feeling that the basic problems of SDE are fixed, and will never change. This is definitely not the case! The world is rapidly changing, and this certainly holds true for SDE. The traditional way of producing data, by sending out questionnaires to selected respondents or interviewing selected respondents, and subsequently processing and analyzing the observed data, is for a substantial part being replaced by making use of already available register data. This presents us with new problems related to SDE.

First, differences in definitions of the variables and the population units between the available register data and the desired information have to be resolved before register data can be used. This can be seen as a special form of SDE. Second, the external register data may have to be edited themselves. Major differences between editing self-collected survey data and external register data are that in the former case one knows, in principle, all the details regarding the data collection process whereas in the latter case one does not, and that in the former case one can re-contact respondents as a last resort whereas in the latter case this is generally impossible. Another difference is that the use of register data requires co-operation with other agencies, for instance tax offices. An increased use of register data seems to be the way of the future for most NSIs. The main challenge for the near future for SDE is to adapt itself so we can handle these data efficiently and effectively.

SDE and imputation are more closely related than space restrictions allow us to describe in this book. In practice, one often needs a well-balanced selected mix of SDE and imputation techniques to edit and impute a data set (see, e.g., Pannekoek and De Waal, 2005). Often in survey and census practice, imputation is carried out to deal with edit failures. Apart from briefly sketching the basic idea underlying NIM, where imputations are used to identify erroneous fields, in Section 6, we have not examined the relation between SDE and imputation any further.

In this paper we have also not examined methods that deal simultaneously with outliers and missing data in multivariate settings. There is a growing literature on these methods; we refer the interested reader to Béguin, and Hulliger (2004 and 2008), Ghosh-Dastidar and Schafer (2006), Elliott and Stettler (2007), and in particular to the paper by Little and Smith (1987). For some recent work on outliers in the context of data editing we refer to Di Zio, Guarnera and Luzi (2008).

In this book we also do not examine how to impute missing data in such a way that all specified edits are satisfied. Some recent work has been carried out in this area. For imputation of categorical data subject to edits we refer to Winkler (2003) and for imputation of numerical data subject to edits to Drechsler and Raghunathan (2008), Pannekoek, Shlomo and De Waal (2008) and, in particular, Tempelman (2007).

References

Al-Hamad A., Lewis, D. and Silva, P.L.N. (2008), *Assessing the Performance of the Thousand Pounds Automatic Editing Procedure at the ONS and the Need for an Alternative Approach*. UN/ECE Work Session on Statistical Data Editing, Vienna.

Banff Support Team (2003), *Functional Description of the Banff System for Edit and Imputation*. Technical report, Statistics Canada.

Bankier, M. (2003), *Current and Future Applications of CANCEIS at Statistics Canada*. UN/ECE Work Session on Statistical Data Editing,Madrid.

Bankier, M., Poirier, P., Lachance, M. and Mason, P. (2000), A Generic Implementation of the Nearest-Neighbour Imputation Methodology (NIM). *Proceedings of the Second International Conference on Establishment Surveys*, Buffalo, pp. 571-578.

Barcaroli, G., Ceccarelli, C., Luzi, O., Manzari, A., Riccini, E. and Silvestri, F. (1995), *The Methodology of Editing and Imputation of Qualitative Variables Implemented in SCIA*. Internal Report, ISTAT, Rome.

Béguin, C. and Hulliger, B. (2004), Multivariate Outlier Detection in Incomplete Survey Data: The Epidemic Algorithm and Transformed Rank Correlations. *Journal of the Royal Statistical Society, Series A*, 167, pp. 275-294.

Béguin, C. and Hulliger, B. (2008), The BACON-EEM Algorithm for Multivariate Outlier Detection in Incomplete Survey Data. *Survey Methodology*, 34, pp. 91-103.

Bethlehem, J. (2007), *Reducing the Bias of Web Survey Based Estimates*. Discussion paper 07001, Statistics Netherlands, Voorburg.

Bienas, J.L., Lassman, D.M., Scheleur, S.A. and Hogan, H. (1997), Improving Outlier Detection in Two Establishment Surveys. *Statistical Data Editing (Volume 2); Methods and Techniques*, United Nations, Geneva.

Blaise Reference Manual (2002), Methods and Informatics Department, Statistics Netherlands, Heerlen.

Blaise Developer's Guide (2002), Methods and Informatics Department, Statistics Netherlands, Heerlen.

Børke, S. (2008), *Using "Traditional" Control (Editing) Systems to Reveal Changes when Introducing New Data Collection Instruments*. UN/ECE Work Session on Statistical Data Editing, Vienna.

Boskovitz, A. (2008), *Data Editing and Logic: The Covering Set Method from the Perspective of Logic*. Ph.D. Thesis, Australian National University, Canberra.

Breiman, L., Friedman, J.H., Olsen, R.A. and Stone, C.J. (1984), *Classification and Regression Trees*. Wadsworth International Group, Belmont.

Bruni, R., Reale, A. and Torelli, R. (2001), Optimization Techniques for Edit Validation and Data Imputation. *Proceedings of Statistics Canada Symposium 2001 "Achieving Data Quality in a Statistical Agency: a Methodological Perspective" XVIII-th International Symposium on Methodological Issues*.

Bruni, R. and Sassano, A. (2001), *Logic and Optimization Techniques for an Error Free Data Collecting*. Report, University of Rome "La Sapienza".

Casado Valero, C., Del Castillo Cuervo-Arango, F., Mateo Ayerra, J. and De Santos Ballesteros, A. (1996), *Quantitative Data Editing: Quadratic Programming Method*. Presented at the COMPSTAT 1996 Conference, Barcelona.

Central Statistical Office (2000), *Editing and Calibration in Survey Processing*. Report SMD-37, Ireland.

Chambers, J.M., Cleveland, W.S., Kleiner, B. and Tukey, P.A. (1983), *Graphical Methods for Data Analysis*. Duxburry Press, Boston.

Couper, M.P., Baker, R.P., Bethlehem, J., Clark, C.Z.F., Martin, J., Nichols, W.L. II and O'Reilly, J.M. (eds.) (1998), *Computer Assisted Survey Information Collection*. John Wiley & Sons, New York.

Cormen, T.H., Leiserson, C.E. and Rivest, R.L. (1990), *Introduction to Algorithms*. The MIT Press/McGraw-Hill Book Company, Cambridge MA.

CSPro User's Guide (2007), International Programs Center, U.S. Census Bureau, Washington D.C.

CSPro Data Entry User's Guide (2007), International Programs Center, U.S. Census Bureau, Washington D.C.

De Jong, A. (2002), *Uni-Edit: Standardized Processing of Structural Business Statistics in the Netherlands*. UN/ECE Work Session on Statistical Data Editing, Helsinki.

DesJardins, D. (1997), *Experiences with Introducing New Graphical Techniques for the Analysis of Census Data*. UN/ECE Work Session on Statistical Data Editing, Prague.

De Waal, T. (1996), *CherryPi: A Computer Program for Automatic Edit and Imputation*. UN/ECE Work Session on Statistical Data Editing, Voorburg.

De Waal, T. (2001), SLICE: Generalised Software for Statistical Data Editing. In: *Proceedings in Computational Statistics* (eds. J.G. Bethlehem and P.G.M. Van der Heijden), Physica-Verlag, New York, pp. 277-282.

De Waal, T. (2003a), *Processing of Erroneous and Unsafe Data*. Ph.D. Thesis, Erasmus University, Rotterdam.

De Waal, T. (2003b), Solving the Error Localization Problem by Means of Vertex Generation. *Survey Methodology*, 29, pp. 71-79.

De Waal, T. (2005), Automatic Error Localisation for Categorical, Continuous and Integer Data. *Statistics and Operations Research Transactions*, 29, pp. 57-99.

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

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

De Waal, T., Renssen, R. and Van de Pol, F. (2000), Graphical Macro-Editing: Possibilities and Pitfalls. *Proceedings of the Second International Conference on Establishment Surveys*, Buffalo, pp. 579-588.

Di Zio, M., Guarnera, U. and Luzi, O. (2005a), *Improving the Effectiveness of a Probabilistic Editing Strategy for Business Data*. ISTAT, Rome.

Di Zio, M., Guarnera, U. and Luzi, O. (2005b), Editing Systematic Unity Measure Errors through Mixture Modelling. *Survey Methodology*, 31, pp. 53-63.

Di Zio, M., Guarnera, U. and Luzi, O. (2008), *Contamination Models for the Detection of Outliers and Influential Errors in Continuous Multivariate Data*. UN/ECE Work Session on Statistical Data Editing, Vienna.

Drechsler, J. and Raghunathan, T.E. (2008), *Evaluating Different Approaches for Multiple Imputation under Linear Constraints*. UN/ECE Work Session on Statistical Data Editing, Vienna.

EDIMBUS (2007), *Recommended Practices for Editing and Imputation in Cross-Sectional Business Surveys*. (see `http://edimbus.istat.it/dokeos/document/document.php`).

Elliott, M.R. and Stettler, N. (2007), Using a Mixture Model for Multiple Imputation in the Presence of Outliers: the ‘Healthy for Life’ Project. *Applied Statistics*, 56, pp. 63-78.

Engström, P. and Ängsved, C. (1997), A Description of a Graphical Macro-Editing Application. *Statistical Data Editing (Volume 2); Methods and Techniques*, United Nations, Geneva.

Esposito, R., Fox, J.K., Lin, D. and Tidemann, K. (1994), ARIES: A Visual Path in the Investigation of Statistical Data. *Journal of Computational and Graphical Statistics*, 3, pp. 113-125.

Esposito, R., Lin, D. and Tidemann, K. (1997), The ARIES Review System in the BLS Current Employment Statistics Program. *Statistical Data Editing (Volume 2); Methods and Techniques*, United Nations, Geneva.

EUREDIT Project (2004a), *Towards Effective Statistical Editing and Imputation Strategies – Findings of the Euredit project, Volume 1.* (see `http://www.cs.york.ac.uk/euredit/results/results.html`).

EUREDIT Project (2004b), *Methods and Experimental Results from the Euredit Project, Volume 2.* (see `http://www.cs.york.ac.uk/euredit/results/results.html`).

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

Federal Committee on Statistical Methodology (1990), *Data Editing in Federal Statistical Agencies*. Statistical Policy Working Paper 18, U.S. Office of Management and Budget, Washington D.C.

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

Ferguson, D.P. (1994), An Introduction to the Data Editing Process. *Statistical Data Editing (Volume 1); Methods and Techniques*, United Nations, Geneva.

Freund, R.J. and Hartley, H.O. (1967), A Procedure for Automatic Data Editing. *Journal of the American Statistical Association*, 62, pp. 341-352.

Garcia Rubio, E. and Criado, I.V. (1990), DIA System: Software for the Automatic Imputation of Qualitative Data. *Proceedings of the US Census Bureau 6th Annual Research Conference*.

Garfinkel, R.S., Kunnathur, A.S. and Liepins, G.E. (1986), Optimal Imputation of Erroneous Data: Categorical Data, General Edits. *Operations Research*, 34, pp. 744-751.

Ghosh-Dastidar, B. and Schafer, J.L. (2006), Outlier Detection and Editing Procedures for Continuous Multivariate Data. *Journal of Official Statistics*, 22, pp. 487-506.

Granquist, L. (1984), *Data Editing and its Impact on the Further Processing of Statistical Data*. Workshop on Statistical Computing, Budapest.

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* (eds. Cox, Binder, Chinnappa, Christianson 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 Kovar, J. (1997), Editing of Survey Data: How Much is Enough?. In: *Survey Measurement and Process Quality* (eds. Lyberg, Biemer, Collins, De Leeuw, Dippo, Schwartz and Trewin), John Wiley & Sons, New York, pp. 415-435.

Hedlin, D. (2003), Score Functions to Reduce Business Survey Editing at the U.K. Office for National Statistics. *Journal of Official Statistics*, 19, pp. 177-199.

Hedlin, D. (2008), *Local and Global Score Functions in Selective Editing*. UN/ECE Work Session on Statistical Data Editing, Vienna.

Hidiroglou, M.A. and Berthelot, J.-M. (1986), Statistical Editing and Imputation for Periodic Business Surveys. *Survey Methodology*, 12, pp. 73-83.

Hoogland, J. (2002), *Selective Editing by Means of Plausibility Indicators*. UN/ECE Work Session on Statistical Data Editing, Helsinki.

Hoogland, J. and Smit, R. (2008), *Selective Automatic Editing of Mixed Mode Questionnaires for Structural Business Statistics*. UN/ECE Work Session on Statistical Data Editing, Vienna.

Hoogland, J. and Van der Pijll, E. (2003), *Summary of the Evaluation of Automatic versus Manual Editing of the Production Statistics 2000 Trade and Transport*. UN/ECE Work Session on Statistical Data Editing, Madrid.

Houston, G. and Bruce, A.G. (1993), Gred: Interactive Graphical Editing for Business Surveys. *Journal of Official Statistics*, 9, pp. 81-90.

Kovar, J. and Whitridge, P. (1990), Generalized Edit and Imputation System; Overview and Applications. *Revista Brasileira de Estadistica*, 51, pp. 85-100.

Laflamme, F., Barrett, C., Johnson, W. and Ramsay, L. (1996), Experiences in Re-Engineering the Approach to Editing and Imputing Canadian Imports Data. *Proceedings of the Bureau of the Census Annual Research Conference and Technology Interchange*, pp. 1025-1037.

Latouche, M. and Berthelot, J.-M. (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 McDavitt, C. (1994), Significance Editing in the Australian Survey of Average Weekly Earning. *Journal of Official Statistics*, 10, pp. 437-447.

Lawrence, D. and McKenzie, R. (2000), The General Application of Significance Editing. *Journal of Official Statistics*, 16, pp. 243-253.

Little, R.J.A. and Smith, P.J. (1987), Editing and Imputation of Quantitative Survey Data. *Journal of the American Statistical Association*, 82, pp. 58-68.

Manzari, A. (2004), Combining Editing and Imputation Methods: An Experimental Application on Population Census Data. *Journal of the Royal Statistical Society A*, 167, pp. 295-307.

Nordbotten, S. (1995), Editing Statistical Records by Neural Networks. *Journal of Official Statistics*, 11, pp. 391-411.

Pannekoek, J. and De Waal, T. (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., Shlomo, N. and De Waal, T. (2008), *Calibrated Imputation of Numerical Data under Linear Edit Restrictions*. UN/ECE Work Session on Statistical Data Editing, Vienna.

Pierzchala, M. (1990), A Review of the State of the Art in Automated Data Editing and Imputation. *Journal of Official Statistics*, 6, pp. 355-377.

Salazar-González, J.J., Lowthian, P., Young, C., Merola, G., Bond, S. and Brown, D. (2004), Getting the Best Results in Controlled Rounding with the Least Effort. In: *Privacy in Statistical Databases* (eds. Domingo-Ferrer and Torra), Springer-Verlag, Berlin, pp. 58-71.

Sande, G. (1978), *An Algorithm for the Fields to Impute Problems of Numerical and Coded Data*. Technical report, Statistics Canada.

Sanders, S. (2002), *Selectief Gaafmaken m.b.v. Classificatie- en Regressiebomen* (in Dutch). Statistics Netherlands, Voorburg.

Schaffer, J. (1987), Procedure for Solving the Data-Editing Problem with Both Continuous and Discrete Data Types. *Naval Research Logistics*, 34, pp. 879-890.

Scholtus, S. (2008a), *Algorithms for Detecting and Resolving Obvious Inconsistencies in Business Survey Data*. UN/ECE Work Session on Statistical Data Editing, Vienna.

Scholtus, S. (2008b), *Algorithms for Correcting Some Obvious Inconsistencies and Rounding Errors in Business Survey Data*. Discussion paper, Statistics Netherlands, Voorburg.

Shlomo, N. and De Waal, T. (2008), Protection of Micro-Data Subject to Edit Constraints Against Statistical Disclosure. *Journal of Official Statistics*, 24, pp. 229-253.

Tempelman, C. (2007), *Imputation of Restricted Data*. Ph.D. Thesis, University of Groningen.

Todaro, T.A. (1999), *Overview and Evaluation of the AGGIES Automated Edit and Imputation System*. UN/ECE Work Session on Statistical Data Editing, Rome.

Tukey, J.W. (1977), *Exploratory Data Analysis*. Addison-Wesley, London.

Van Langen, S. (2002), *Selectief Gaafmaken met Logistische Regressie* (in Dutch). Statistics Netherlands, Voorburg.

Van de Pol, F. and Bethlehem, J. (1997), Data Editing Perspectives. *Statistical Journal of the United Nations ECE*, 14, pp. 153-171.

Van de Pol, F. and Diederen, B. (1996), A Priority Index for Macro-Editing the Netherlands Foreign Trade Survey. *Proceedings of the Data Editing Workshop and Exposition*, Washington D.C.

Van de Pol, F. and Molenaar, W. (1995), Selective and Automatic Editing with CADI-Applications. In: *Essays on Blaise 1995, Proceedings of the Third International Blaise Users' Conference* (ed. Kuusela), Statistics Finland, Helsinki, pp. 159-168.

Van der Loo, M.P.J. (2008), *An Analysis of Editing Strategies for Mixed-Mode Establishment Surveys*. Discussion paper 08004, Statistics Netherlands.

Weir, P., Emery, R. and Walker, J. (1997), The Graphical Editing Analysis Query System. *Statistical Data Editing (Volume 2); Methods and Techniques*, United Nations, Geneva.

Winkler, W.E. (1998), *Set-Covering and Editing Discrete Data*. Statistical Research Division Report 98/01, U.S. Bureau of the Census, Washington D.C.

Winkler, W.E. (2003), *Contingency-Table Model for Imputing Data Satisfying Analytic Constraints*. U.S. Bureau of the Census, Washington D.C.

Winkler, W.E. and Draper, L.A. (1997), The SPEER Edit System. *Statistical Data Editing (Volume 2); Methods and Techniques*, United Nations, Geneva.

Winkler, W.E. and Petkunas, T.F. (1997), The DISCRETE Edit System. *Statistical Data Editing (Volume 2); Methods and Techniques*. United Nations, Geneva.