WP. 39
ENGLISH ONLY

UNITED NATIONS
ECONOMIC COMMISSION FOR EUROPE

CONFERENCE OF EUROPEAN STATISTICIANS

Work Session on Statistical Data Editing
(Oslo, Norway, 24-26 September 2012)

Topic (vi): New and emerging methods

Automatic Editing with Hard and Soft Edits – Some First Experiences

Prepared by Sander Scholtus and Sevinç Göksen, Statistics Netherlands

I. Introduction

1. Many national statistical institutes (NSIs) nowadays use automatic editing as a partial alternative to manual editing, to increase the efficiency of their editing processes. In most applications, a record is edited automatically in two steps: first, the variables with erroneous or missing values are located (the *error localisation problem*) and subsequently new values are imputed for these variables to obtain a valid record (the *consistent imputation problem*). We shall focus on the error localisation problem here.

2. For the data sets used in official statistics, many *edits* can usually be formulated. Edits describe constraints that should be satisfied by the data. Both manual and automatic editing are guided by the information contained in the edits. In particular, attention is focused on records that do not satisfy – i.e. *fail* – certain edits. In practice, some edits are only failed by erroneous combinations of values (these are known as *hard edits* or fatal edits), while other edits are failed by suspicious combinations of values that are not necessarily incorrect (these are known as *soft edits* or query edits). Since soft edits are an important element of manual editing, it is desirable to use the information contained in these edits also during automatic editing. However, current algorithms for automatic editing treat all edits as hard constraints: each edit failure is attributed to an error. Thus, there are only two ways to handle soft edits during automatic editing: either ignoring them or treating them as if they were hard edits. This is in fact an important practical difference between manual and automatic editing, which is likely to produce systematic differences between data that is edited manually and data that is edited automatically.

3. Recently, a new automatic editing method has been developed at Statistics Netherlands, which can take soft edits into account in a more meaningful way. In this paper, we describe some of the first empirical results obtained with this new method. The remainder of this paper is organised as follows. Section II describes the error localisation problem as formulated by Fellegi and Holt (1976), as well as a generalisation that can distinguish between hard and soft edits. The latter formulation does not specify precisely how soft edit failures should be handled in the error localisation problem. Section III suggests several ways to do this. In Section IV, the suggested approaches are tested in a simulation study with data from the Dutch structural business statistics. Some concluding remarks follow in Section V.

II. Error Localisation Problems

A. An Error Localisation Problem without Soft Edits

4. We will assume throughout that the data consists of numerical variables $(x_1, \ldots x_j, \ldots, x_p)$ and that the edits can be written as linear equalities or inequalities. That is, if we denote the set of edits by Ψ, each edit $\psi_k \in \Psi$ has one of the following two forms:

$$a_{k1}x_1 + \cdots + a_{kp}x_p + b_k = 0 \tag{1}$$

or

$$a_{k1}x_1 + \cdots + a_{kp}x_p + b_k \geq 0, \tag{2}$$

where $a_{k1},\ldots,a_{kp},b_k$ are numerical constants. A variable x_j is said to be *involved* in an edit ψ_k if and only if $a_{kj} \neq 0$. Clearly, the failure or non-failure of an edit is completely determined by the values of the variables involved in that edit.

5. For a given record $(x_1^0,\ldots,x_p^0)$, which might contain both errors and missing values, it is easily assessed which edits are failed and which edits are satisfied. (We make the working assumption that an edit is failed if it involves a variable with a missing value.) If none of the edits are failed, the record is *consistent* with the edits and requires no editing. If the record is not consistent with the edits, then we would like to identify the erroneous values that cause the edit failures, i.e. to solve the error localisation problem. Actually, this problem is impossible to solve with certainty in any realistic application. In practice, we can only solve a less ambitious problem: to identify a subset of the variables which *(i)* can be imputed so that the record becomes consistent with the edits and *(ii)* is 'optimal' according to some chosen criterion. A solution to the latter problem (which is also commonly referred to as 'the error localisation problem') may or may not identify the actual erroneous values in a record. This depends in particular on the optimisation criterion and the set of edits used.

6. Fellegi and Holt (1976) proposed to solve the above-mentioned problem by minimising the number of variables to impute. This optimisation criterion has become widely-used for automatic editing at NSIs. Often a generalised version of the Fellegi-Holt paradigm is used, for which each variable is given a so-called *confidence weight*. These weights provide a means to take into account that some variables naturally contain more errors than others, for instance because they correspond to questions that many respondents find difficult to answer. Higher confidence weights are attached to variables that supposedly contain fewer errors, and vice versa. The objective now becomes to find a set of variables to impute that minimises the following distance measure:

$$D_{FH} = \sum_{j=1}^{p} w_j y_j , \tag{3}$$

where w_j denotes the confidence weight of variable x_j, and y_j is a binary variable with $y_j = 1$ if x_j is imputed and $y_j = 0$ otherwise. The original Fellegi-Holt paradigm is recovered as a special case by choosing all w_j equal, for instance equal to one.

7. Mathematically, the problem of minimising (3) under the condition that the imputed record satisfies a given set of edits of the forms (1) and (2) can be written as a mixed integer linear programming problem (see e.g. Riera-Ledesma and Salazar-González, 2003). As such, it can be solved using commercially available solvers. In addition, some NSIs have developed and implemented specialised algorithms for solving the Fellegi-Holt based error localisation problem; see De Waal et al. (2011) for an overview. One such example is a branch-and-bound algorithm due to De Waal and Quere (2003), which has been implemented at Statistics Netherlands in the automatic editing tool SLICE and, more recently, in the R package `editrules` (De Jonge and Van der Loo, 2011).

B. An Error Localisation Problem with Soft Edits

8. In the above error localisation problem, it is tacitly assumed that all edits are hard edits. Hence, a subset of variables is only considered a feasible solution if it can be imputed to make the record consistent with all edits (cf. condition *(i)* in paragraph 5). If soft edits have been specified for the data at hand, then these have to be either discarded or interpreted as hard edits during automatic editing. As mentioned in the introduction, this rather crude way of handling soft edits may be a source of systematic differences between automatic editing and manual editing.

9. Scholtus (2011) proposed an alternative formulation of the error localisation problem which provides room for a more meaningful use of soft edits. In this formulation, it is assumed that the edit set Ψ is partitioned into two disjoint subsets: $\Psi = \Psi^H \cup \Psi^S$. The set Ψ^H contains the hard edits; the set Ψ^S contains the soft edits. The error localisation problem is now stated as the problem of identifying a subset of the variables which *(i)* can be imputed so that the record becomes consistent with the edits in Ψ^H and *(ii)* minimises the following distance measure:

$$D = \lambda D_{FH} + (1-\lambda) D_{soft}. \tag{4}$$

In this expression, D_{FH} is given by (3), D_{soft} represents the costs associated with failed edits in Ψ^S, and $\lambda \in [0,1]$ is a parameter that balances the contributions of both terms to D.

10. In order to apply expression (4) in practice, a cost function D_{soft} has to be specified. The next section introduces several possible choices for D_{soft}. These choices can be divided into two broad classes: cost functions that only depend on which soft edits are failed, and cost functions that also depend on the amount by which each soft edit is failed. Intuitively, cost functions from the latter class are the most attractive. Human editors interpret soft edit failures as measures of suspicion, and the degree of suspicion becomes higher as the size of the edit failure increases. For example, suppose that the soft edit $x_1 \leq 9$ is failed by two records with $x_1 = 10$ and $x_1 = 100$, respectively. Both records fail the edit and are therefore suspicious, but an editor would usually consider the second record more likely to be in error than the first. By taking the sizes of soft edit failures into account in D_{soft} in (4), we can attempt to mimic this behaviour during automatic editing.

11. An algorithm for solving the error localisation problem of paragraph 9 is presented in Scholtus (2011). This is in fact an extension of the above-mentioned branch-and-bound algorithm of De Waal and Quere (2003). Scholtus (2011) also noted that the problem becomes much more complex to solve for the case that D_{soft} depends on the amounts by which the soft edits are failed. Thus, in spite of the intuitive appeal of using the sizes of edit failures in D_{soft}, it might be impractical to do so in realistic applications. Hence, it is also interesting to look at cost functions that do not depend on the sizes of edit failures, because they can be handled more efficiently.

III. Possible Measures of Soft Edit Failure

A. Preliminary Notation

12. Suppose that Ψ^S consists of the following soft edits: $\psi_1^S, \ldots, \psi_k^S, \ldots, \psi_{K_S}^S$. Let the binary variable z_k be defined so that $z_k = 1$ if edit ψ_k^S is failed and $z_k = 0$ otherwise. Since all edits are assumed to be linear equalities or inequalities, there exists a natural measure of the amount of edit failure. A given record $(x_1^0, \ldots, x_p^0)$ may be said to fail an edit of the form (1) by the amount

$$e_k = \left| a_{k1} x_1^0 + \cdots + a_{kp} x_p^0 + b_k \right| \tag{5}$$

and an edit of the form (2) by the amount

$$e_k = \max\left\{0, -(a_{k1} x_1^0 + \cdots + a_{kp} x_p^0 + b_k)\right\}. \tag{6}$$

In both cases, the value of e_k equals the absolute amount by which the left-hand-side of the expression would have to be shifted in order for the edit to become satisfied. In particular, $e_k = 0$ if the edit is satisfied and $e_k > 0$ otherwise. For the example of paragraph 10, the first record would have $e_k = 1$ and the second record would have $e_k = 91$ according to (6).

13. In Subsection III.B, we will consider some simple variants of D_{soft} that can be computed from the values of $z_1, \ldots, z_{K_S}$ alone. Subsection III.C introduces a trick whereby the amount of edit failure can

be taken into account to some extent while still using only the values of $z_1,\ldots,z_{K_S}$. Finally, Subsection III.D considers variants of D_{soft} that depend directly on $e_1,\ldots,e_{K_S}$. Göksen (2012) described many other variants which we do not mention here. For the most part, we restrict attention to variants that showed promising results in the simulation study to be discussed in Section IV.

B. Cost Functions that Depend on $z_1,\ldots,z_{K_S}$

14. Given the definition of D_{FH} as a sum of confidence weights in (3), it seems natural to use an analogous definition for D_{soft} in (4). That is, we may associate a weight s_k (which we will call a *failure weight*) to each soft edit ψ_k^S, and define D_{soft} as the sum of weights of failed soft edits:

$$D_{soft} = \sum_{k=1}^{K_S} s_k z_k \ . \tag{7}$$

The failure weights in (7) have a similar interpretation to the confidence weights in (3). Just as it may be thought that the variables are not a priori equally reliable, so it may be that some soft edits are considered 'harder' than others. In other words, the probability that a particular soft edit failure is caused by an erroneous combination of values may be different for different soft edits. Note that for hard edits this probability equals 1 by definition.

15. To illustrate the use of expression (4) if D_{soft} is defined by (7), we consider a small example. Suppose that the error localisation algorithm has found the following feasible solutions: *(i)* it is possible to satisfy all edits (hard and soft) by imputing x_1 and x_2; *(ii)* it is possible to satisfy all hard edits and all soft edits except for ψ_1^S by imputing x_3; *(iii)* it is also possible to satisfy all hard edits and all soft edits except for ψ_4^S and ψ_5^S by imputing x_3 (i.e. with a different value). In this example, the value of D equals $\lambda(w_1+w_2)$ for solution *(i)*, $\lambda w_3+(1-\lambda)s_1$ for solution *(ii)*, and $\lambda w_3+(1-\lambda)(s_4+s_5)$ for solution *(iii)*. Hence, the choice of confidence weights w_j, failure weights s_k, and balancing parameter λ determines which of the feasible solutions is the optimal one.

16. As a general point of interest, it should be noted that – since expression (4) is minimised – an edit with a high failure weight is less likely to be failed by the optimal solution to the error localisation problem than an edit with a low failure weight. Thus, soft edits that are relatively 'hard' as mentioned in paragraph 14 should be given a relatively high weight. We will now consider several ways to choose the failure weights in (7). We will distinguish different variants by adding superscripts to s_k.

17. If we have no prior information about the error detecting probability of the soft edits, then we may choose $s_k^A = 1$ for all $k = 1,\ldots,K_S$ by default. In the remainder of this section, we will assume that prior information is available in the form of a reference data set $\Re$ that has already been manually edited. Given this information, one obvious choice for s_k is to take the proportion of records in the edited data set that satisfy soft edit ψ_k^S:

$$s_k^B = \frac{1}{|\Re|}\sum_{i\in\Re}(1-z_{ki}^{clean}) \ , \tag{8}$$

where $|\Re|$ denotes the number of records in $\Re$ and z_{ki}^{clean} denotes the value of z_k for the ith record after manual editing. A high value of $0 \le s_k^B \le 1$ indicates a soft edit that is often satisfied in the edited reference data. In particular, a soft edit with $s_k^B = 1$ is, for all practical purposes, actually a hard edit. Alternatively, s_k^B can be interpreted as an empirically estimated probability that a record satisfies edit ψ_k^S after manual editing. We may write this succinctly as $s_k^B = \hat{P}(z_k^{clean} = 0)$. This probability is estimated from the reference data and subsequently used during automatic editing to predict whether a record should be made to satisfy ψ_k^S or not.

18. Interestingly, the above method of prediction can be improved, at least from a theoretical point of view. Analogous to (8), the following empirically estimated probabilities may be computed from the reference data:

$$\hat{P}(z_k^{raw}=1)=\frac{1}{|\Re|}\sum_{i\in\Re} z_{ki}^{raw},$$

$$\hat{P}(z_k^{raw}=1\wedge z_k^{clean}=0)=\frac{1}{|\Re|}\sum_{i\in\Re} z_{ki}^{raw}(1-z_{ki}^{clean}).$$

The first expression estimates the probability that ψ_k^S is failed prior to editing; note that z_{ki}^{raw} is the counterpart of z_{ki}^{clean} in the unedited version of the reference data set. The second expression represents the estimated probability that a record fails edit ψ_k^S prior to editing *and* satisfies the edit after editing. According to the definition of conditional probability,

$$\hat{P}(z_k^{clean}=0\,|\,z_k^{raw}=1)=\frac{\hat{P}(z_k^{raw}=1\wedge z_k^{clean}=0)}{\hat{P}(z_k^{raw}=1)}.$$

When solving the error localisation problem for a given record, the value of z_k^{raw} is known. Moreover, only the failure weights of the edits with $z_k^{raw}=1$ are relevant for evaluating expression (7). Thus instead of using the above unconditional probability $\hat{P}(z_k^{clean}=0)$ to predict whether ψ_k^S should be satisfied or not, we may also use the conditional probability $\hat{P}(z_k^{clean}=0\,|\,z_k^{raw}=1)$, which might give a more accurate prediction. With this in mind, we define the following failure weight:

$$s_k^C=\hat{P}(z_k^{clean}=0\,|\,z_k^{raw}=1)=\frac{\sum_{i\in\Re} z_{ki}^{raw}(1-z_{ki}^{clean})}{\sum_{i\in\Re} z_{ki}^{raw}}. \tag{9}$$

19. As we have seen, both s_k^B and s_k^C can be interpreted as estimated probabilities. Inserting these probabilities directly into the linear expression (7) might not give the best results. A simple alternative approach is to derive categorical versions of these weights, which we denote by $s_k^{B(cat)}$ and $s_k^{C(cat)}$, respectively. Formally, we map the interval $[0,1]$ onto a finite (in fact rather small) set of values $\{s^{(1)},\ldots,s^{(m)}\}$. For instance, taking $m=3$ we could choose two cut-off values $a_{low}\le a_{high}$ and define

$$s_k^{B(cat)}=\begin{cases}0.5 & \text{if } s_k^B<a_{low}\\ 1 & \text{if } a_{low}\le s_k^B\le a_{high}\\ 1.5 & \text{if } s_k^B>a_{high}\end{cases} \tag{10}$$

with an analogous definition for $s_k^{C(cat)}$. This might give better results than using s_k^B and s_k^C directly if the handling of soft edit failures during manual editing depends on the above estimated probabilities in a strongly non-linear manner.

C. Quantile Edits

20. A drawback of the variants considered so far is that they do not take the sizes of soft edit failures into account. Any record that fails the edit ψ_k^S receives the same contribution to (7), namely s_k. In Subsection III.D, we will discuss variants of D_{soft} that take the sizes of soft edit failures into account directly. The current subsection describes an indirect way to take the edit failure sizes into account.

21. The basic idea is as follows. Consider a soft edit ψ_k^S of the form (2) and suppose that this edit is replaced by the following system of soft edits:

$$\begin{cases} \psi^S_{k(1)}: & a_{k1}x_1 + \cdots + a_{kp}x_p + b_{k(1)} \geq 0 \\ \psi^S_{k(2)}: & a_{k1}x_1 + \cdots + a_{kp}x_p + b_{k(2)} \geq 0 \\ \quad\vdots & \qquad\qquad\qquad\qquad \vdots \\ \psi^S_{k(R)}: & a_{k1}x_1 + \cdots + a_{kp}x_p + b_{k(R)} \geq 0 \end{cases} \tag{11}$$

where the constant terms are chosen so that

$$b_k = b_{k(1)} < b_{k(2)} < \cdots < b_{k(R)}. \tag{12}$$

Suppose in addition that the edits in this system are given failure weights $s_{k(1)}, s_{k(2)}, \ldots, s_{k(R)}$, and that expression (7) is used for D_{soft}. It is not difficult to show that, for $r = 2, \ldots, R$, any record which fails edit $\psi^S_{k(r)}$ automatically also fails the edits $\psi^S_{k(1)}, \ldots, \psi^S_{k(r-1)}$. Hence, a record which fails edit $\psi^S_{k(r)}$ actually receives a contribution of $s_{k(1)} + \cdots + s_{k(r)}$ to cost function (7). In this way, larger edit failures with respect to the original edit ψ^S_k implicitly receive larger contributions to D_{soft}, for the simple reason that they fail more edits in system (11).

22. Assuming again that prior information is available in the form of a reference data set that has been edited manually, an interesting application of the above idea could be as follows. Compute the value of $Q = a_{k1}x_1 + \cdots + a_{kp}x_p$ for each record in the edited data set. Next, consider the univariate distribution of Q and let $Q^{(\alpha)}$ denote the $\alpha \times 100\%$ quantile of this distribution (for $\alpha \in [0,1]$). Let α_1 be the largest percentage point for which $Q^{(\alpha)} \leq -b_k$. Clearly, soft edit (2) is failed by $\alpha_1 \times 100\%$ of the edited records. Now choose specific values $0 \leq \alpha_R < \cdots < \alpha_2 < \alpha_1$ and define $b_{k(r)} = -Q^{(\alpha_r)}$ for $r = 1,2,\ldots,R$. These values satisfy property (12) and therefore may be inserted as constant terms in system (11). Göksen (2012) uses the term *quantile edits* to refer to a system of edits obtained in this manner. An advantage of using quantile edits is that these are soft edits with prescribed failure rates: by construction, the quantile edit $\psi^S_{k(r)}$ is failed by $\alpha_r \times 100\%$ of the edited records.

23. In practice many soft edits are *ratio edits*, i.e. bivariate edits of the form $x_{j_1} / x_{j_2} \geq c$, where both involved variables are constrained to be non-negative by the hard edits. A ratio edit can be linearised as $x_{j_1} - cx_{j_2} \geq 0$, which is an inequality edit of the form (2). For a soft edit ψ^S_k of this kind, a variation on the above idea can be applied by varying the value of coefficient c:

$$\begin{cases} \psi^S_{k(1)}: & x_{j_1} - c_{(1)}x_{j_2} \geq 0 \\ \psi^S_{k(2)}: & x_{j_1} - c_{(2)}x_{j_2} \geq 0 \\ \quad\vdots & \qquad\qquad\qquad \vdots \\ \psi^S_{k(R)}: & x_{j_1} - c_{(R)}x_{j_2} \geq 0 \end{cases}$$

where the coefficients have to be chosen so that $c_{(R)} < \cdots < c_{(2)} < c_{(1)}$. Again it can be shown that $\psi^S_{k(r)}$ is failed only if $\psi^S_{k(1)}, \ldots, \psi^S_{k(r-1)}$ are also failed (for $r = 2, \ldots, R$). As before, we can use this fact to ensure that larger edit failures with respect to the ratio edit receive larger contributions to D_{soft}. In particular, quantile edits can be generated from the univariate distribution of $Q' = x_{j_1} / x_{j_2}$ in a reference data set, by taking $c_{(r)} = Q'^{(\alpha_r)}$ for some choice of percentage points with $0 \leq \alpha_R < \cdots < \alpha_2 < \alpha_1 < 1$. Again, it automatically holds that quantile edit $\psi^S_{k(r)}$ is failed by $\alpha_r \times 100\%$ of the edited records. There exists an analogous construction for ratio edits of the form $x_{j_1} / x_{j_2} \leq c$; see Göksen (2012).

24. A potential drawback of the approach in this subsection is that it increases the number of edits. In general, error localisation algorithms tend to show a dramatic increase in the amount of required computing time and memory as the number of edits becomes larger (for an explanation, see e.g. De Jonge

and Van der Loo, 2011). Therefore, it seems advisable to limit the number of additional edits introduced by this method. In the simulation study to be discussed below we used $R = 3$.

D. Cost Functions that Depend on $e_1, \ldots, e_{K_S}$

25. We now consider two variants of D_{soft} that depend on $e_1, \ldots, e_{K_S}$, the measures of individual edit failure size. First of all, it should be noted that e_k as defined by (5) or (6) is not invariant to arbitrary changes in the formulation of the soft edits. In particular, we can multiply the left-hand-side of (1) or (2) by an arbitrary constant $\lambda > 0$ to obtain an equivalent edit. However, all values of e_k would be inflated by a factor λ under this operation. Moreover, the typical magnitude of e_k may not be directly comparable for different soft edits. Hence, it seems appropriate to standardise the values of $e_1, \ldots, e_{K_S}$ in some way before using them in an expression for D_{soft}.

26. We assume as before that there is a reference data set $\Re$ available. From this data set we may estimate the mean and variance of e_k in manually edited data:

$$\hat{\bar{e}}_k^{clean} = \frac{1}{|\Re_0|} \sum_{i \in \Re_0} e_{ki}^{clean}, \qquad \hat{\sigma}_k^2 = \frac{1}{|\Re_0| - 1} \sum_{i \in \Re_0} (e_{ki}^{clean} - \hat{\bar{e}}_k^{clean})^2,$$

where e_{ki}^{clean} denotes the value of e_k for the edited version of the ith record in the reference data; $\Re_0$ denotes the subset of edited records in $\Re$ that fail at least one (soft) edit. That is, we estimate the above quantities conditional on the fact that a record has at least one non-zero e_k. The rationale behind this is that the expression used for D_{soft} in (4) is only relevant for records that satisfy this condition. Therefore, we are not interested in the distribution of e_k outside $\Re_0$.

27. One possible method to standardise e_k for use in D_{soft} is as follows. Dividing each e_k by its estimated standard deviation to obtain $e_k^s = e_k / \hat{\sigma}_k$, we may define in analogy with (7):

$$D_{soft} = \sum_{k=1}^{K_S} e_k^s . \tag{13}$$

Note that by construction only the failed soft edits have a non-zero contribution to (13), just as in (7).

28. A second possible method computes the well-known Mahalanobis distance between the vector $\mathbf{e} = (e_1, \ldots, e_{K_S})'$ of edit failure sizes and a vector of zeros, corresponding with no soft edit failures:

$$D_{soft} = D_M(\mathbf{e}, \mathbf{0}) = \sqrt{\mathbf{e}' \hat{\boldsymbol{\Sigma}}^{-1} \mathbf{e}}, \tag{14}$$

where $\hat{\boldsymbol{\Sigma}}$ denotes the variance-covariance matrix of e_k in $\Re_0$. This distance measure also takes potential correlations between edit failures into account in the standardisation. Unlike the variants of D_{soft} considered so far, expression (14) cannot be written as a linear combination of $z_1, \ldots, z_{K_S}$ or $e_1, \ldots, e_{K_S}$. Taking the Mahalanobis distance of edit failures was suggested by Hedlin (2003) in a different context.

IV. Simulation Study

29. In this section, we present some results of a simulation study in which the above choices of D_{soft} were tested. For the purpose of this simulation study, a prototype implementation of the error localisation algorithm from Scholtus (2011) was written using the R programming language. This prototype makes extensive use of the `editrules` package (De Jonge and Van der Loo, 2011).

30. We used two data sets in the simulation study. All data was obtained from manually edited records on medium-sized businesses (10-100 employees) from the Dutch wholesale structural business statistics of 2007. The manually edited data was assumed to be error-free. In both cases, we selected half

of the original edited data as reference data. The other half was used to provide test data for automatic editing. For data set 1, we introduced synthetic errors in the edited data by randomly modifying about 20% of the values; see Scholtus (2011) for details on the error mechanism used. For data set 2, we used the unedited data – containing actual errors made by actual respondents – as test data. The number of records to edit equalled 728 for data set 1 and 580 for data set 2.

31. Tables 1 and 2 display the variables and edits for data set 1. As can be seen, most soft edits are linearised ratio edits. For instance, the first soft edit expresses that turnover from wholesale should represent at least 50% of the total net turnover. Data set 2 contains ten numerical variables related to questions on the number of employees. Due to space constraints, we do not present the variables and edits used for data set 2 in detail here; see Göksen (2012). For both data sets, we initially chose all confidence weights equal to 1 and experimented only with the specification of D_{soft}. The results reported here were all obtained with $\lambda = 1/2$ in (4). We refer to Göksen (2012) for results with other values of λ.

Table 1. Variables in data set 1

	variable description
x_1	net turnover from wholesale
x_2	net turnover from other activities
x_3	net total turnover
x_4	net turnover from wholesale[a)]
x_5	internal revenues
x_6	external revenues, type 1
x_7	external revenues, type 2
x_8	total other revenues
x_9	total operating revenues
x_{10}	total operating costs
x_{11}	pre-tax results
x_{12}	number of persons on payroll

[a)] Alternative calculation of net turnover from wholesale as total of a product group breakdown

Table 2. Edits for data set 1

hard edits:	$x_1 + x_2 = x_3$
	$x_1 = x_4$
	$x_5 + x_6 + x_7 = x_8$
	$x_3 + x_8 = x_9$
	$x_9 - x_{10} = x_{11}$
	$x_j \geq 0$ (for all $j \neq 11$)
soft edits:	$x_1 - 0.5x_3 \geq 0$
	$x_3 - 0.9x_9 \geq 0$
	$x_5 + x_6 \geq x_7$
	$x_9 - 50x_{12} \geq 0$
	$5000x_{12} - x_9 \geq 0$
	$0.4x_9 - x_{11} \geq 0$
	$x_{11} + 0.1x_9 \geq 0$
	$x_{12} \geq 1$
	$x_{12} \geq 5$
	$x_{12} \leq 100$

32. Since the distribution of errors was (assumed) known for the data sets in this simulation study, we could directly evaluate the success of each error localisation effort. Consider the following 2×2 contingency table for a given editing approach:

		detected: error	detected: no error
true:	error	*TP*	*FN*
	no error	*FP*	*TN*

From this table, we computed the following quality indicators:

$$\alpha = \frac{FN}{TP + FN}; \qquad \beta = \frac{FP}{FP + TN}; \qquad \gamma = \frac{FN + FP}{TP + FN + FP + TN}.$$

These three indicators evaluate how well an editing approach manages to identify individual values as correct or erroneous: α measures the proportion of true errors that were missed (false negatives); β measures the proportion of correct values mistaken for errors (false positives); and γ measures the overall proportion of wrong decisions. Clearly, a successful editing approach should achieve low values of α, β, and γ. For an alternative evaluation measure, we also calculated the proportion of records for which exactly the right solution was found – that is, the solution that identifies as erroneous all erroneous values and only these. This fourth indicator is denoted by δ. To be successful, an editing approach should achieve a high value of δ.

33. Table 3 displays the evaluation results for a number of error localisation approaches for both data sets. The first two lines present the results of, respectively, not using soft edits and using all soft edits as

if they were hard edits. As noted above, these are the two ways that are available to handle soft edits in the error localisation problem based on (3). For data set 1, it is clear that using the soft edits as hard edits is not a good idea. Although this approach did manage to locate more erroneous values than ignoring the soft edits did ($\alpha = 0.232$ versus $\alpha = 0.364$), the latter approach gave better results according to the other evaluation measures β, γ, and δ. For data set 2, using all soft edits as if they were hard edits was not even possible, because an infeasible problem was encountered if all hard and soft edits had to be satisfied simultaneously by all records. Therefore, we did not evaluate this approach for this data set.

34. The next five lines contain results for the methods introduced in Subsection III.B. For both data sets, it is seen that a significant improvement compared to not using soft edits was already obtained by taking the soft edits into account with all failure weights equal to $s_k^A = 1$. Differentiating the failure weights of soft edits according to the choice s_k^B led to a further improvement for data set 1, but not for data set 2. Contrary to our expectations, the alternative failure weight s_k^C does not appear to be an improvement on s_k^B; in fact, this choice made the results significantly worse for data set 1. We also experimented with several categorised versions $s_k^{B(cat)}$ and $s_k^{C(cat)}$. The results in Table 3 were obtained with expression (10) and its analogue for $s_k^{C(cat)}$, where the optimal choices of a_{low} and a_{high} depended on the data set (see Göksen, 2012). For data set 1, the categorised versions did not perform better than the original weights, but for data set 2 we observed a clear improvement.

Table 3. Evaluation results for both data sets

editing approach	data set 1				data set 2			
	α	β	γ	δ	α	β	γ	δ
no soft edits	0.364	0.047	0.115	40.2%	0.551	0.003	0.131	58.4%
all edits as hard edits	0.232	0.131	0.153	36.8%	-	-	-	-
<u>failure weights</u>								
weights A	0.227	0.060	0.096	47.3%	0.435	0.014	0.121	63.4%
weights B	0.253	0.037	0.083	52.1%	0.508	0.004	0.125	60.9%
weights C	0.332	0.038	0.101	43.3%	0.524	0.003	0.126	60.7%
weights B(cat)	0.224	0.053	0.089	50.0%	0.432	0.009	0.116	64.5%
weights C(cat)	0.333	0.038	0.101	43.1%	0.444	0.007	0.117	64.5%
<u>quantile edits</u>								
weights 1/3, 1/3, 1/3	0.247	0.032	0.078	54.4%	0.481	0.005	0.121	63.4%
weights 0.9, 0.05, 0.05	0.214	0.037	0.075	56.5%	0.466	0.005	0.119	63.8%
<u>amounts of edit failure</u>								
expression (13)	0.273	0.050	0.098	49.2%	-	-	-	-
Mahalanobis distance	0.328	0.049	0.109	46.8%	-	-	-	-

35. The next two lines relate to quantile edits. For both data sets, we replaced each soft edit with $R = 3$ quantile edits, namely those corresponding to the 10%, 5%, and 1% quantile of the underlying distribution (i.e. $\alpha_1 = 0.1$, $\alpha_2 = 0.05$, and $\alpha_3 = 0.01$). The first line shows the results obtained by giving each quantile edit a failure weight of 1/3; the second line shows the results when the 10%, 5%, and 1% quantile edits were given failure weights of 0.9, 0.05, and 0.05, respectively. These choices produced the best results seen so far for data set 1. Moreover, the results for data set 2 also ranked among the best seen so far. Both of the above sets of weights are such that the maximal implicit failure weight for a record that fails the original soft edit equals 1. Göksen (2012) also reported results for other choices of weights, but these did not improve on the results shown here.

36. The final two lines present the results of approaches for which D_{soft} depends explicitly on the sizes of soft edit failures. Due to time constraints, we only evaluated these approaches for the first data set. Somewhat contrary to expectation, the results did not improve on the best results seen so far. In

particular, the results of the Mahalanobis distance were rather poor compared with some of the other approaches for using soft edits.

37. Finally, as noted above, the results presented so far were obtained with all confidence weights equal to 1. If the confidence weights are not all equal to 1, this should be taken into account in the definition of D_{soft}, since otherwise the two terms in (4) are weighted unintentionally. An easy solution, which works for the above approaches that do not depend explicitly on the sizes of soft edit failures, is to multiply each failure weight with an appropriate factor. This factor should inflate D_{soft} to the same magnitude as D_{FH} in (4). Experimental results with the above data sets (not reported here, but see Göksen, 2012) suggested that a good approach is to multiply the failure weight of soft edit ψ_k^S with the harmonic mean of the confidence weights of all variables involved in ψ_k^S. That is, if we denote the set of variables involved in ψ_k^S by $\Im_k$, the failure weight s_k should be multiplied by

$$\frac{|\Im_k|}{\sum_{j\in\Im_k} w_j^{-1}} = \frac{\sum_{j\in\Im_k} 1}{\sum_{j\in\Im_k} w_j^{-1}} = \frac{\sum_{j\in\Im_k} w_j^{-1} w_j}{\sum_{j\in\Im_k} w_j^{-1}} = \sum_{j\in\Im_k} p_j w_j \ , \tag{15}$$

with $p_j = w_j^{-1} \Big/ \sum_{j'\in\Im_k} w_{j'}^{-1}$. The rightmost expression in (15) shows that this factor may be interpreted as follows: it is the expected value of the contribution to D_{FH} when exactly one of the variables involved in ψ_k^S is imputed, under the assumption that, for each variable, the probability to be selected for imputation is proportional to the reciprocal of its confidence weight.

V. Conclusion

38. In this paper, we presented several ways to use the information contained in soft edits during automatic editing. We also reported some empirical results with these methods in a simulation study. These results showed that all methods considered here can be used to improve the quality of automatic editing in comparison with the current method of using only hard edits. Some methods were more successful in this respect than others, however, and the results also depended on the data set at hand. Overall, the method of quantile edits appears to be the best choice of the methods considered here. But there may be room for further improvement. In particular, it seems likely that the approach to let D_{soft} depend explicitly on the sizes of soft edit failures could be used to more effect than we have done here.

39. The description in this paper was restricted to numerical data and edits, but this restriction is not necessary. A similar error localisation problem with hard and soft edits can be formulated and solved for categorical and mixed data; see Scholtus (2011). Future work will include a simulation study that involves mixed data and edits.

VI. Literature

De Jonge, E. and M. van der Loo (2011), Manipulation of Linear Edits and Error Localization with the Editrules Package. Discussion Paper 201120, Statistics Netherlands, The Hague.

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

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

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

Göksen, S. (2012), Automatic Error Localisation with Hard and Soft Constraints at Statistics Netherlands. Master Thesis (in Dutch), Statistics Netherlands, The Hague.

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.

Riera-Ledesma, J. and J.J. Salazar-González (2003), New Algorithms for the Editing and Imputation Problem. Working Paper No. 5, UN/ECE Work Session on Statistical Data Editing, Madrid.

Scholtus, S. (2011), Automatic Editing with Soft Edits. Discussion Paper 201130, Statistics Netherlands, The Hague.