Working Paper.
ENGLISH ONLY

UNITED NATIONS
ECONOMIC COMMISSION FOR EUROPE

CONFERENCE OF EUROPEAN STATISTICIANS

Work Session on Statistical Data Editing
(Paris, France, 28-30 April 2014)

Topic (ii): New and emerging methods

A generalised Fellegi-Holt paradigm for automatic editing

Prepared by Sander Scholtus (Statistics Netherlands)

I. Introduction

1. Traditionally, data editing has been a manual task, performed by human editors with extensive subject-matter knowledge. To improve the efficiency and timeliness of editing, many statistical institutes have attempted to automate parts of this process. This has resulted in, on the one hand, deductive correction methods for systematic errors (e.g., unit of measurement errors) and, on the other hand, error localisation algorithms for random errors (De Waal et al., 2011). In this paper, I will focus on automatic editing for random errors. In this step, each record of data is minimally adjusted, according to some optimisation criterion, so that it becomes consistent with a predefined set of constraints known as *edit rules* (or *edits* for short). Depending on the effectiveness of the optimisation criterion and the strength of the edit rules, automatic editing may be used as a partial alternative to manual editing.

2. Most automatic editing methods that are currently used in official statistics are based on the seminal paradigm of Fellegi and Holt (1976). According to this paradigm, the error localisation problem is solved by finding, for each record, the smallest subset of variables that can be imputed so that the record becomes consistent with the edits. A slight generalisation is obtained by assigning so-called *confidence weights* to the variables and minimising the total weight of the imputed variables; variables with higher confidence weights are then assumed less likely to contain errors. Having obtained a solution to the error localisation problem, one needs to find suitable imputations for the variables that have been identified as erroneous. This is a separate problem, known as the consistent imputation problem [see, e.g., De Waal et al. (2011) and their references]. In this paper, I will focus on the error localisation problem.

3. In practice, automatic editing is applied nearly always in combination with some form of selective editing. Hence, the most influential errors are still treated manually. Most statisticians consider manually edited data to be of higher quality than data that have been edited automatically. In fact, the outcome of manual editing is usually taken as the "gold standard" for assessing the quality of automatic editing. A critical evaluation of this assumption is beyond the scope of the present paper. Here I simply note that, by improving the ability of automatic editing methods to mimic the results of manual editing, their usefulness in practice may be increased. In turn, this means that the share of automatic editing may be increased to improve the efficiency of the data editing process (Pannekoek et al., 2013).

4. Some years ago, Statistics Netherlands conducted a series of evaluation studies in which data sets from the Dutch Structural Business Statistics (SBS) were edited both automatically and manually. When the results of the two editing efforts were compared, a number of systematic differences were found. Many of these differences could be explained by the fact that human editors performed certain types of adjustments that do not fit well within the constraints of the Fellegi-Holt paradigm. For instance, editors sometimes interchanged the value of *costs of type A* with that of *revenues of type A*. This type of adjustment corresponds to one underlying error: the respondent mixed up the answers to two related

questions. However, under the Fellegi-Holt paradigm, this adjustment requires two independent imputations. In addition, the error of interchanging costs and revenues of the same type often caused an edit failure that could also be solved by adjusting the value of *balance of type A*, the relevant edit being:

balance of type A = revenues of type A – costs of type A.

During automatic editing, this alternative solution was usually preferred because it requires only one imputation. Subject-matter specialists preferred interchanging the costs and revenues, based on their knowledge of respondent behaviour.

5. As another example, editors sometimes transferred (parts of) reported amounts between variables; for instance, they would transfer a part of the reported *turnover from retail trade* to *turnover from wholesale* when the original amounts did not match the reported stocks of retail goods and wholesale goods. Again, this is a complex type of adjustment involving at least two variables that nonetheless is considered as a single correction by the editors.

6. To some extent, systematic differences between automatic and manual editing can be prevented by a clever choice of confidence weights. In general, however, it is difficult to predict the effects that a certain modification of the confidence weights will have on the results of automatic editing. Moreover, if the editors apply a number of different complex adjustments, it might be impossible to model all of them under the Fellegi-Holt paradigm using a single set of confidence weights. Another option is to try to catch errors for which the Fellegi-Holt paradigm is known to provide an unsatisfactory solution at an earlier stage in the data editing process, i.e., during deductive editing of systematic errors. Ideally, this would ensure that Fellegi-Holt-based editing is only applied to those errors for which it is suited. There are some practical limitations to this approach, however. Properly designing a large collection of automatic correction rules, and maintaining such a collection over time, can be a difficult task. For instance, the same set of rules applied to the same record may produce a different outcome depending on the way the rules are ordered. Moreover, it is not self-evident that appropriate correction rules can be found for all errors that do not fit within the Fellegi-Holt paradigm.

7. In this paper, a different approach is suggested. A new definition of the error localisation problem is proposed that allows the possibility that errors affect more than one variable at a time. It is shown that this new problem contains error localisation under the original Fellegi-Holt paradigm as a special case. Informally speaking, under the new paradigm errors are located by minimising the number of so-called *edit operations*. Imputing a new value for one variable at a time is an example of an edit operation. However, more general edit operations can also be allowed that involve changes to multiple variables. For now, I restrict attention to numerical data and linear edits.

8. The remainder of this paper is organised as follows. In Section II, the concept of an edit operation as it will be used here is formally introduced and illustrated. The new error localisation problem is defined in Section III and illustrated by means of a small example in Section IV. Some features of the new error localisation problem are discussed in Section V. Finally, some conclusions and questions for further research follow in Section VI.

II. Edit operations

9. Let $\boldsymbol{x} = (x_1, \ldots, x_p)' \in \mathbb{R}^p$ be a record of p numerical variables. Suppose that this record has to satisfy k edit rules, in the form of the following system of linear (in)equalities:

$$\mathbf{A}\boldsymbol{x} + \boldsymbol{b} \odot \mathbf{0}, \tag{1}$$

where $\mathbf{A}$ is a $k \times p$ matrix of coefficients and $\boldsymbol{b}$ is a k vector of constants. Throughout this paper, $\mathbf{0}$ will be used to represent a vector of zeros of appropriate length; similarly, $\odot$ will represent a symbolic vector of operators from the set $\{\geq, =\}$.

10. I define an edit operation g to be a linear function of $\mathbb{R}^p$ to itself, having the general form

$$g(\boldsymbol{x}) = \mathbf{T}\boldsymbol{x} + \boldsymbol{h}, \tag{2}$$

where $\mathbf{T}$ denotes a known $p \times p$ coefficient matrix and $\boldsymbol{h}$ denotes a p vector. Importantly, the elements h_i may be either known constants or linear functions of free parameters. The unique free parameters (if any) that occur in $\boldsymbol{h}$ are denoted by $\alpha_1, \ldots, \alpha_m$, or by α if there is only one. In some cases, it may be useful to impose one or several linear constraints on these parameters:

$$\mathbf{R}\boldsymbol{\alpha} + \boldsymbol{d} \odot \mathbf{0}, \tag{3}$$

with $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_m)'$, $\mathbf{R}$ a known coefficient matrix, and $\boldsymbol{d}$ a known vector of constants.

11. In automatic editing based on the Fellegi-Holt paradigm, a central role is attached to the replacement of one original value by an arbitrary new value (imputation). This is in fact a particular edit operation of the form (2) which I will call an *FH operation*. To find the FH operation that imputes the variable x_j, one takes $\mathbf{T}$ to be a diagonal matrix with $t_{jj} = 0$ and all other diagonal elements equal to one; in addition, all elements of $\boldsymbol{h}$ except h_j are zero, while $h_j = \alpha$, with α an unrestricted parameter. The resulting FH operation yields:

$$g\left((x_1, \ldots, x_{j-1}, x_j, x_{j+1}, \ldots, x_p)'\right) = (x_1, \ldots, x_{j-1}, \alpha, x_{j+1}, \ldots, x_p)', \tag{4}$$

with $\alpha \in \mathbb{R}$ representing the imputed value. It should be noted that for a record of p variables, p distinct FH operations of the form (4) can be defined. No restrictions of the form (3) are imposed directly on α; however, in the context of the error localisation problem, an FH operation is of interest only if there exists a value for α that can help to make the record consistent with the edits (1).

12. To illustrate the concept of an edit operation, some further examples will now be given. For notational convenience, I restrict attention to the case $p = 3$.

(a) An edit operation that changes the sign of one of the variables:

$$g\left(\begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix}\right) = \begin{pmatrix} -1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \cdot \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} + \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix} = \begin{pmatrix} -x_1 \\ x_2 \\ x_3 \end{pmatrix}.$$

(b) An edit operation that interchanges the values of two adjacent items:

$$g\left(\begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix}\right) = \begin{pmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix} \cdot \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} + \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix} = \begin{pmatrix} x_2 \\ x_1 \\ x_3 \end{pmatrix}.$$

(c) An edit operation that transfers an amount between two items, where the amount transferred may equal at most K units in either direction:

$$g\left(\begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix}\right) = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \cdot \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} + \begin{pmatrix} \alpha \\ 0 \\ -\alpha \end{pmatrix} = \begin{pmatrix} x_1 + \alpha \\ x_2 \\ x_3 - \alpha \end{pmatrix},$$

with the constraint that $-K \leq \alpha \leq K$.

(d) An edit operation that computes the value of a total from the values of its parts:

$$g\left(\begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix}\right) = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 1 & 1 & 0 \end{pmatrix} \cdot \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} + \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix} = \begin{pmatrix} x_1 \\ x_2 \\ x_1 + x_2 \end{pmatrix}.$$

(e) An edit operation that imputes two variables simultaneously using a fixed ratio:

$$g\left(\begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix}\right) = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix} \cdot \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} + \begin{pmatrix} \alpha_1 \\ \alpha_2 \\ 0 \end{pmatrix} = \begin{pmatrix} \alpha_1 \\ \alpha_2 \\ x_3 \end{pmatrix},$$

with the constraint that $\boldsymbol{\alpha} = (\alpha_1, \alpha_2)'$ satisfies $10\alpha_1 - \alpha_2 = 0$.

13. These examples illustrate in particular that free parameters can be useful to avoid the need to construct a separate edit operation for every possible adjustment to the data. For instance, an FH operation does not specify the exact value that is imputed. Similarly, in the third example above, the exact amount that is transferred between x_1 and x_3 is not specified.

14. Intuitively, an edit operation is supposed to "reverse the effects" of a particular type of error[1] that may have occurred in the observed data. That is to say, if the error associated with edit operation g actually occurred in the observed record $\boldsymbol{x}$, then $g(\boldsymbol{x})$ is the record that would have been observed if that error had not occurred.

15. It should be clear that one could construct an abundance of edit operations of the form (2). In any particular application, only a small subset of these potential operations would have a substantively meaningful interpretation (in the sense that the associated types of errors are known to occur in that application). In what follows, I assume that a finite set of specific edit operations of the form (2) has been identified as relevant for a particular application. This will be called the set of *allowed edit operations* for that application. Some suggestions on how to construct this set will be given in Section VI.

III. A generalised error localisation problem

16. Consider a set of allowed edit operations for a given application of automatic editing. Informally, I propose to generalise the error localisation problem of Fellegi and Holt (1976) by replacing "the smallest subset of variables that can be imputed to make the record consistent" with "the shortest sequence of allowed edit operations that can be applied to make the record consistent". To give a formal definition of this generalised error localisation problem, some new notation and concepts need to be introduced first.

17. Let $\mathcal{G}$ be a finite set of allowed edit operations. To each edit operation $g \in \mathcal{G}$, a weight $w_g > 0$ can be associated that expresses the costs of applying edit operation g. These weights may be seen as a generalisation of the confidence weights that were mentioned in Section I, by identifying the weight of the FH operation that imputes a new value for x_j with the confidence weight of that variable.

18. Consider a sequence of points $\boldsymbol{x} = \boldsymbol{x}_0, \boldsymbol{x}_1, \dots, \boldsymbol{x}_t = \boldsymbol{y}$ in $\mathbb{R}^p$. A *path* from $\boldsymbol{x}$ to $\boldsymbol{y}$ is defined as a sequence of *distinct* edit operations $g_1, \dots, g_t \in \mathcal{G}$ such that $\boldsymbol{x}_n = g_n(\boldsymbol{x}_{n-1})$ for all $n \in \{1, \dots, t\}$. Note that it is not allowed to use the same edit operation twice on the same path, not even with a different choice of parameter(s). A path is denoted by $P = [g_1, \dots, g_t]$. The set of all possible paths from $\boldsymbol{x}$ to $\boldsymbol{y}$ is denoted by $\mathcal{P}(\boldsymbol{x}, \boldsymbol{y})$. This set may be empty.

19. The *length* of a path $P = [g_1, \dots, g_t]$ is defined as the sum of the weights of its constituent edit operations:

$$\ell(P) = \sum_{n=1}^{t} w_{g_n}, \tag{5}$$

where, by convention, the empty path has length zero. Note that two paths have the same length if they consist of the same subset of edit operations $G \subseteq \mathcal{G}$, regardless of the order. Next, the *distance* from $\boldsymbol{x}$ to $\boldsymbol{y}$ is defined as the length of the shortest path that connects $\boldsymbol{x}$ to $\boldsymbol{y}$. If no such path exists, then this distance is considered to be infinitely large:

$$d(\boldsymbol{x}, \boldsymbol{y}) = \begin{cases} \min\{\ell(P) | P \in \mathcal{P}(\boldsymbol{x}, \boldsymbol{y})\} & \text{if } \mathcal{P}(\boldsymbol{x}, \boldsymbol{y}) \neq \emptyset, \\ \infty & \text{otherwise.} \end{cases}$$

[1] Note that an error is defined in this paper as any type of disturbance that may occur in the observed data. This disturbance may be multivariate; for instance, interchanging the values of two variables can be considered an error. Hence, an error is not the same thing as an erroneous value; it is a more general concept. It is also important to distinguish between errors and edit failures. The latter can be *caused* by errors – and hence used to recognise that errors occurred in the observed data – but they are not errors as such.

In general, $d(\boldsymbol{x}, \boldsymbol{y})$ satisfies the standard axioms of a metric *except* that it need not be symmetric in $\boldsymbol{x}$ and $\boldsymbol{y}$; it is a so-called "quasimetric" (Scholtus, 2014). Accordingly, I will refer to $d(\boldsymbol{x}, \boldsymbol{y})$ as "the distance from $\boldsymbol{x}$ to $\boldsymbol{y}$" rather than "the distance between $\boldsymbol{x}$ and $\boldsymbol{y}$".

20. Let D be a closed, non-empty subset of $\mathbb{R}^p$. The distance from any point $\boldsymbol{x}$ to D is defined to be the distance from $\boldsymbol{x}$ to the nearest point $\boldsymbol{y} \in D$:

$$d(\boldsymbol{x}, D) = \min\{d(\boldsymbol{x}, \boldsymbol{y}) | \boldsymbol{y} \in D\}.$$

One subset of $\mathbb{R}^p$ that is of particular interest here is the set of all points that satisfy the edits (1); denote this set by D_0. Provided that the system of edits is feasible (as it should be in practice), it holds that $D_0 \neq \emptyset$. Moreover, D_0 is closed because (1) does not contain any strict inequalities.

21. I can now formulate the generalised error localisation problem. Consider a given set of consistent records D_0 [defined by a system of linear edits (1)], a given set of allowed edit operations $\mathcal{G}$, and a given record $\boldsymbol{x}$. If $d(\boldsymbol{x}, D_0) = \infty$, then the error localisation problem for $\boldsymbol{x}$ is infeasible. Otherwise, any record $\boldsymbol{y} \in D_0$ such that $d(\boldsymbol{x}, \boldsymbol{y}) < \infty$ is called a *feasible solution* to the error localisation problem for $\boldsymbol{x}$. A feasible solution $\boldsymbol{x}^*$ is called *optimal* if it holds that

$$d(\boldsymbol{x}, \boldsymbol{x}^*) = d(\boldsymbol{x}, D_0). \tag{6}$$

Formally, then, the generalised error localisation problem consists of finding an $\boldsymbol{x}^* \in D_0$ that satisfies expression (6).

22. It should be noted that if $\boldsymbol{x}^*$ is an optimal solution to the error localisation problem for $\boldsymbol{x}$, any other record in D_0 that can be reached by the same path of edit operations is also an optimal solution. To solve the error localisation problem, it is sufficient to find an optimal path of edit operations. Constructing an associated record $\boldsymbol{x}^* \in D_0$ may then be regarded as a generalisation of the consistent imputation problem. Moreover, in the special case that $\boldsymbol{x} \in D_0$, the unique optimal solution is given by $\boldsymbol{x}^* = \boldsymbol{x}$: the error localisation problem is trivial for records that are already consistent with the edits.

23. So far, no assumption has been made about the set of allowed edit operations $\mathcal{G}$, other than that this set should be finite. In general, the above error localisation problem might be infeasible for some records $\boldsymbol{x}$. This would happen whenever $\boldsymbol{x}$ cannot be mapped onto D_0 by any combination of distinct edit operations in $\mathcal{G}$. To avoid this situation, $\mathcal{G}$ should be sufficiently large so that $d(\boldsymbol{x}, D_0) < \infty$ for all $\boldsymbol{x} \in \mathbb{R}^p$. It can be shown that this property holds in particular for any $\mathcal{G}$ that contains all p distinct FH operations. In fact, it is possible to connect any point in $\mathbb{R}^p$ to any other point in $\mathbb{R}^p$ by a path that concatenates the FH operations associated with the coordinates on which the two points differ. For simplicity, Scholtus (2014) assumes that $\mathcal{G}$ contains all FH operations. In principle, though, one could also construct other sets of allowed edit operations for which the error localisation problem is always feasible. The special case that $\mathcal{G}$ consists *only* of the p distinct FH operations is of some interest. It is not difficult to see that problem (6) then reduces to the original error localisation problem of Fellegi and Holt (1976), albeit with confidence weights.

IV. Example

24. Consider the following system of linear edits in two numerical variables x_1 and x_3:

$$x_1 + x_3 = 19, \tag{7}$$
$$x_1 \geq 4, \tag{8}$$
$$-x_1 \geq -7, \tag{9}$$
$$-x_1 + x_3 \geq 5, \tag{10}$$
$$x_1 - x_3 \geq -10, \tag{11}$$
$$x_3 \geq 0. \tag{12}$$

Scholtus (2014) describes a more elaborate version of this example that also includes the variable x_2. I refer to that paper for detailed derivations of results that are stated without proof here.

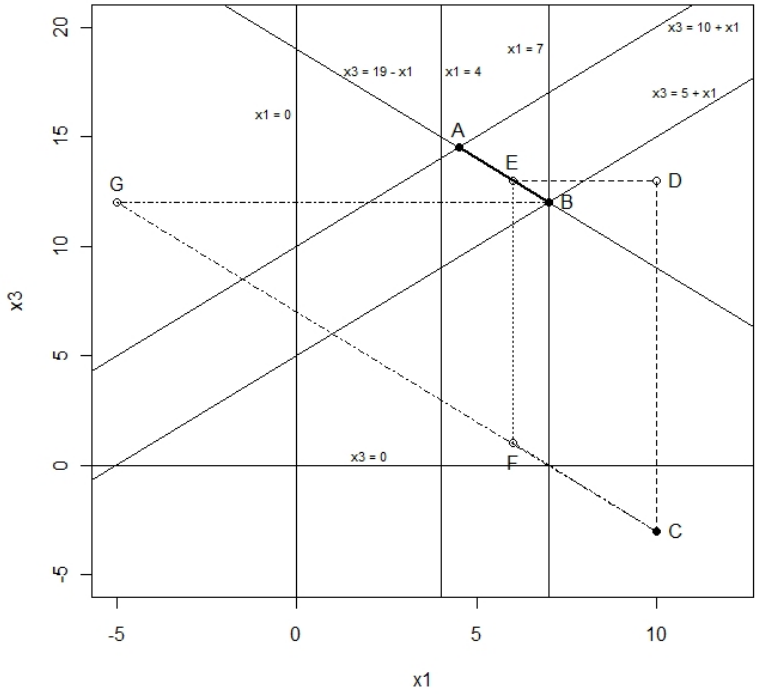

Figure 1. Illustration of error localisation results for a two-dimensional example.

25. It can be shown that a record is consistent with the edits (7)–(12) if, and only if, it has the following form: $(x_1, x_3) = (7 - \beta, 12 + \beta)$, with $0 \leq \beta \leq 5/2$. Figure 1 illustrates this graphically. The boundary of the region defined by each edit from (7)–(12) is plotted as a solid line in the (x_1,x_3) plane. The feasible region defined jointly by these edits is shown as the bold line segment AB, with $A = \left(4\frac{1}{2}, 14\frac{1}{2}\right)$ and $B = (7,12)$; note that AB contains precisely all points of the above-mentioned form.

26. The record $(x_1, x_3) = (10, -3)$, plotted as point C in Figure 1, requires editing as it fails some of the edits in (7)–(12). Suppose that the following edit operations of the form (2) are available:

name	description	weight
g_1	FH operation for variable x_1	$w_{g_1} = 1$
g_2	FH operation for variable x_3	$w_{g_2} = 3$
g_3	change the sign of variable x_1	$w_{g_3} = 0.5$
g_4	transfer an amount of at most $K = 15$ units between x_1 and x_3 (in either direction)	$w_{g_4} = 1$

Formal expressions for these edit operations can be derived from the examples in Section II. A choice of weights for these operations is given in the last column of the table.

27. Under the paradigm of Fellegi and Holt (1976), the only edit operations that are allowed are g_1 and g_2. For the record $(x_1, x_3) = (10, -3)$, both of these edit operations are needed to obtain a feasible solution to the error localisation problem; i.e., both variables have to be imputed. Any point E on the line segment AB (in fact, any point in $\mathbb{R}^2$) can be obtained in this way, by varying the imputed values. One potential path, shown in Figure 1, consists of the line segment CD (i.e., an imputation for x_3) followed by DE (i.e., an imputation for x_1). The path length associated with this solution is: $w_{g_1} + w_{g_2} = 4$.

28. Now suppose that all four of the above edit operations are allowed. It can be shown that the error localisation problem of Section III has two more feasible solutions in addition to the previous one. The first solution uses the edit operations g_2 and g_4 to impute x_3 and transfer an amount between the two variables. It can be shown that, again, any point on AB can be reached from C using these operations. A possible path is shown in Figure 1 as CF (i.e., a transferred amount from x_1 to x_3) followed by FE (an

imputation for x_3). The associated path length is: $w_{g_2} + w_{g_4} = 4$. The second, optimal solution in this example is given by the edit operations g_1 and g_4. The associated path length is: $w_{g_1} + w_{g_4} = 2$. This solution is more restrictive than the previous two: the only point on AB that can be reached from C by applying these operations is the point B. The corresponding path is displayed in Figure 1 as CG (a transferred amount from x_1 to x_3; note that the maximal allowed amount $K = 15$ is used here) followed by GB (an imputation for x_1). In terms of distances, it holds that $d(C, B) = 2$ and $d(C, E) = 4$ for all points $E \neq B$ on AB. Apparently, for the weights chosen above, it is considered better to adjust C towards B than towards any other point on AB in this example.

V. Some interesting features of the new error localisation problem

A. Implied edits under linear edit operations

29. Scholtus (2014) outlines an algorithm that could be used to solve the error localisation problem of Section III. Here, I will focus on one element of this algorithm: a method to determine whether a given path of edit operations can yield a record that is consistent with the edits.

30. In the special case of error localisation under the Fellegi-Holt paradigm, a similar question arises: whether a given combination of variables can be imputed to obtain a consistent record. To answer this question, many existing error localisation algorithms use a technique called *Fourier-Motzkin elimination* (FM elimination). FM elimination transforms a system of linear constraints having p variables into a system of *implied* linear constraints having at most $p - 1$ variables; thus, at least one of the original variables is eliminated from the constraints. For mathematical details, see, e.g., De Waal et al. (2011) or Williams (1986).

31. FM elimination has the following fundamental property: the system of implied constraints is satisfied by the values of the non-eliminated variables if, and only if, there exists a value for the eliminated variable that, together with the other values, satisfies the original system of constraints. By repeatedly applying this fundamental property, it is possible to verify whether any particular combination of variables can be imputed to obtain a consistent record, given the original values of the other variables (De Waal et al., 2011). A clear illustration of the use of FM elimination for error localisation under the Fellegi-Holt paradigm is provided by the error localisation algorithm of De Waal and Quere (2003).

32. Turning to the generalised error localisation problem of Section III, I first consider an edit operation g of the form (2) that does *not* contain any free parameters. Let $\boldsymbol{x}$ be any record and let $\boldsymbol{y}$ be the record that is obtained by applying g to $\boldsymbol{x}$; that is, $\boldsymbol{y} = g(\boldsymbol{x}) = \mathbf{T}\boldsymbol{x} + \boldsymbol{h}$. By definition, $\boldsymbol{y}$ satisfies the edits (1) if, and only if, $\mathbf{A}(\mathbf{T}\boldsymbol{x} + \boldsymbol{h}) + \boldsymbol{b} \odot \mathbf{0}$, which is equivalent to

$$(\mathbf{AT})\boldsymbol{x} + (\mathbf{A}\boldsymbol{h} + \boldsymbol{b}) \odot \mathbf{0}. \tag{13}$$

Expression (13) may be interpreted as follows: the record $\boldsymbol{y} = g(\boldsymbol{x})$ is consistent with the original edits (1) if, and only if, the record $\boldsymbol{x}$ satisfies a similar system of linear edits, with $\mathbf{AT}$ as coefficient matrix and $\mathbf{A}\boldsymbol{h} + \boldsymbol{b}$ as vector of constants. That is to say, applying the edit operation g to $\boldsymbol{x}$ yields a consistent record if, and only if, $\boldsymbol{x}$ satisfies (13).

33. As an example, consider the following edits for (x_1, x_2):

$$x_1 \geq 0, \tag{14}$$
$$x_2 \geq 0, \tag{15}$$
$$x_1 + x_2 \leq 5. \tag{16}$$

Let g be the edit operation that changes the sign of x_1: $g\big((x_1, x_2)'\big) = (-x_1, x_2)'$. Under this edit operation, the above edits are transformed into the following system:

$$-x_1 \geq 0, \tag{17}$$
$$x_2 \geq 0, \tag{18}$$
$$-x_1 + x_2 \leq 5. \tag{19}$$

The example record $(x_1, x_2)' = (-2,3)'$ is inconsistent with the original edit rules (14)–(16). On the other hand, it does satisfy the transformed edit rules (17)–(19). This implies that the record can be made consistent with the original edits by changing the sign of x_1. It is easily verified that the resulting record $(x_1, x_2)' = (2,3)'$ indeed satisfies (14)–(16).

34. Next, consider the case that g involves at least one free parameter $\boldsymbol{\alpha}$. One still obtains (13) but now as a system of constraints on the original record $\boldsymbol{x}$ *and* the parameters in $\boldsymbol{\alpha}$. By the same reasoning as above, a consistent record can be obtained by applying g to $\boldsymbol{x}$ with a certain choice of $\boldsymbol{\alpha}$ if, and only if, $\boldsymbol{x}$ and $\boldsymbol{\alpha}$ satisfy (13) and (if relevant) the additional restrictions (3). Interestingly, (13) and (3) constitute a system of *linear* restrictions of the form (1) for the extended record $(\boldsymbol{x}', \boldsymbol{\alpha}')'$. Therefore, FM elimination may be used to remove all free parameters from (13) and (3). This yields a system of implied linear restrictions for $\boldsymbol{x}$. Moreover, the fundamental property of FM elimination states that $\boldsymbol{x}$ satisfies this system of implied edits if, and only if, there exist parameter values for $\boldsymbol{\alpha}$ that, together with $\boldsymbol{x}$, satisfy (13) and (3). Hence, it follows that applying the edit operation g to $\boldsymbol{x}$ can lead to a consistent record (for some choice of parameter values) if, and only if, $\boldsymbol{x}$ satisfies the system of implied edits obtained by eliminating $\boldsymbol{\alpha}$ from (13) [and, if relevant, (3)].

35. The extension of this result to paths of more than one edit operation is straightforward; see Scholtus (2014) for more details. It should be noted that, for the special case that g is an FH operation, the above result is consistent with the traditional use of FM elimination. In fact, for the FH operation that imputes the variable x_j, the transformed system of edits (13) is obtained by replacing every occurrence of x_j in the original edits by an unrestricted parameter α. Eliminating α from (13) is therefore equivalent to eliminating x_j directly from the original edits. In this sense, the above result generalises the fundamental property of FM elimination to all edit operations of the form (2).

36. One aspect in which error localisation with general edit operations of the form (2) differs from Fellegi-Holt based editing is that the order in which edit operations are applied may be important. In fact, when only FH operations are used, the order in which variables are imputed does not matter. With general edit operations of the form (2), it can happen that, for instance, the path $P = [g_1, g_2]$ yields a consistent record while the path $P' = [g_2, g_1]$ does not. Scholtus (2014) describes an automated procedure to test whether this order effect occurs. For the paths of edit operations that were used in the example of Section IV, it can be shown that the order does not matter.

B. A statistical interpretation of the error localisation problem

37. In motivating their paradigm for automatic error localisation, Fellegi and Holt (1976) did not provide any formal statistical argument. Their reasoning was more intuitive:

> *"The data in each record should be made to satisfy all edits by changing the fewest possible items of data (fields). This we believe to be in agreement with the idea of keeping the maximum amount of original data unchanged, subject to the constraints of the edits, and so manufacturing as little data as possible. At the same time, if errors are comparatively rare, it seems more likely that we will identify the truly erroneous fields."* (Fellegi and Holt, 1976, p. 18)

In fact, error localisation under the Fellegi-Holt paradigm is often regarded as a "mechanical" approach without a clear statistical interpretation. Alternative error localisation procedures of a more statistical nature have been proposed by, e.g., Little and Smith (1987) and Ghosh-Dastidar and Schafer (2006). These procedures use outlier detection techniques and are based on an explicit model for the true data. Unfortunately, they cannot handle edit rules such as (1) in a straightforward manner, which makes them unsuitable for most applications in official statistics.

38. Scholtus (2014) argues that, under certain conditions, the optimal solution to the generalised error localisation problem of Section III may be interpreted as an approximate maximum likelihood estimator. The derivation of this result is based on Kruskal (1983, pp. 38–39), who gave a similar argument to justify the use of the so-called Levenshtein distance in string comparisons.

39. Heuristically, the argument proceeds as follows; see Scholtus (2014) for more details. Suppose that errors occur stochastically and independently of each other, and that each edit operation in $\mathcal{G}$ acts as a "corrector" for exactly one potential error. Moreover, suppose that $w_g = -\log p_g$, with p_g the probability that the error associated with edit operation g occurs. Assuming that errors are rare ($p_g \ll 1$), the probability of observing the record $\boldsymbol{x}$ given the associated error-free record $\boldsymbol{y}$ is approximately

$$\Pr(\boldsymbol{x}|\boldsymbol{y}) \approx \sum_{P \in \mathcal{P}(\boldsymbol{x},\boldsymbol{y})} \exp\{-\ell(P)\},$$

with $\ell(P)$ given by expression (5). The largest contribution to this sum comes from the shortest path of edit operations connecting $\boldsymbol{x}$ to $\boldsymbol{y}$, say P^*, with $\ell(P^*) = d(\boldsymbol{x}, \boldsymbol{y})$. If it may be assumed that the above sum is dominated by its largest term, it holds approximately that

$$\Pr(\boldsymbol{x}|\boldsymbol{y}) \approx \exp\{-d(\boldsymbol{x}, \boldsymbol{y})\}.$$

Let $\log L(\boldsymbol{y}|\boldsymbol{x})$ denote the loglikelihood function of $\boldsymbol{y}$, given the observed record $\boldsymbol{x}$. Assuming that the edits are hard edits, it holds that $\log L(\boldsymbol{y}|\boldsymbol{x}) = \log 0 = -\infty$ for all $\boldsymbol{y} \notin D_0$. For $\boldsymbol{y} \in D_0$, one obtains that $\log L(\boldsymbol{y}|\boldsymbol{x}) = \log \Pr(\boldsymbol{x}|\boldsymbol{y}) \approx -d(\boldsymbol{x}, \boldsymbol{y})$. This shows that minimising $d(\boldsymbol{x}, \boldsymbol{y})$ over all $\boldsymbol{y} \in D_0$ is approximately equivalent to maximising the loglikelihood of $\boldsymbol{y}$ given $\boldsymbol{x}$. In this sense, the optimal solution to error localisation problem (6) can be justified as an approximate maximum likelihood estimator.

VI. Discussion and conclusion

40. In this paper, a new formulation was proposed of the error localisation problem in automatic editing. It was suggested to find the (weighted) minimal number of edit operations needed to make an observed record consistent with the edits. The new error localisation problem can be seen as a generalisation of the problem proposed by Fellegi and Holt (1976), because the operation that imputes a new value for one variable at a time is an important special case of an edit operation. The discussion in this paper was restricted to numerical data and linear edits. The original Fellegi-Holt paradigm has been applied also to categorical and mixed data (cf. De Waal et al., 2011); in principle, it should be possible to extend the approach of this paper to that context, using appropriately defined edit operations, but this remains to be investigated.

41. Scholtus (2014) describes some theoretical aspects of the new error localisation problem and outlines a possible error localisation algorithm. As discussed in Section V.A, the same elimination technique that is often used to solve the Fellegi-Holt based error localisation problem can be applied also in the context of the new problem. Nevertheless, the task of solving the new error localisation problem is challenging from a computational point of view, at least for the numbers of variables, edits, and edit operations that would be encountered in practical applications.

42. An obvious candidate for applying the new error localisation method in practice would be the SBS. As mentioned in the introduction, the automatic editing process that is currently used in the Dutch SBS is known to produce data that deviate in some respects from data that are edited manually. The method described in this paper has the potential to reduce these systematic differences between automatic and manual editing, by improving the flexibility of automatic editing in terms of the types of amendments that can be made to the data. A reduction of the differences between automatic and manual editing would mean in turn that the efficiency of the editing process could be improved by increasing the fraction of data that is edited automatically.

43. However, more research is needed before the method described in this paper can be applied in practice. To apply the method in a particular context, it is necessary first to specify the relevant edit operations. Ideally, each edit operation should correspond to a combination of amendments to the data that human editors consider to be a correction for one particular error. In addition, a suitable set of weights w_g has to be determined for these edit operations. This would require information about the relative frequencies of the most common types of amendments made during manual editing. Both aspects could be investigated based on historical data before and after manual editing (including paradata logged

during regular production), editing instructions and other documentation used by the editors, and interviews with editors and/or supervisors of editing.

44. On a more fundamental level, a question of demarcation arises between deductive correction methods and automatic editing under the new error localisation problem. In principle, many known types of error could be resolved either by automatic correction rules or by error localisation using edit operations. Both approaches have their advantages and disadvantages (Scholtus, 2014). It is likely that some compromise will yield the best results, with some errors handled deductively and others by edit operations. However, it is not obvious how to make this division in practice.

45. Ultimately, the aim of the new methodology proposed in this paper is to improve the usefulness of automatic editing in practice. Whether the new error localisation problem can be successful in this respect remains to be seen.

References

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.

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

Kruskal, J. B. (1983). An Overview of Sequence Comparison. In D. Sankoff and J. B. Kruskal (eds.), *Time Warps, String Edits, and Macromolecules: The Theory and Practice of Sequence Comparison*, pp. 1–44. Addison-Wesley.

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

Pannekoek, J., S. Scholtus, and M. van der Loo (2013). Automated and Manual Data Editing: A View on Process Design and Methodology. *Journal of Official Statistics* **29**, pp. 511–537.

Scholtus, S. (2014). Error Localisation using General Edit Operations. Discussion paper, Statistics Netherlands, The Hague. Available at www.cbs.nl.

Williams, H. P. (1986). Fourier's Method of Linear Programming and its Dual. *The American Mathematical Monthly* **93**, pp. 681–695.