Discussion paper

A MIP approach for a generalised data editing problem

Jacco Daalmans
Sander Scholtus

July 2018

Content

Summary

Data editing is the problem of identifying and correcting erroneous or missing values in questionnaires that are returned by individual respondents. The error localisation problem amounts to finding the erroneous fields in a record. For automatic error localisation, the paradigm of Fellegi and Holt (1976) is often applied, stating that a minimum number of values should be selected for correction. Recently, Scholtus (2014 and 2016) generalised the Fellegi-Holt (FH) paradigm. He suggested to find the (weighted) minimal number of *edit operations* that are needed to correct a record. Edit operations are typical corrections that are often observed in manual data editing, for instance interchanging the values of two fields. Scholtus (2014 and 2016) presented an algorithm to solve the generalised problem, but noted that the computational performance is too low for a typical practical application.

The current paper further operationalises the generalised paradigm. Two additional assumptions are presented to simplify the problem. Also, a mixed-integer programming (MIP) formulation of the problem is presented, which enables the use of an efficient MIP solver. Based on this MIP formulation, a prototype implementation of an error localisation tool was created in R. This paper also presents results of an application to real-life and simulated data. The results demonstrate the feasibility of the approach for a production setting, and shows that a wider range of errors are identified with this approach than with the traditional, Fellegi-Holt-based method.

Keywords

Data editing; automatic editing; mixed-integer programming; error localisation; Fellegi-Holt.

1. Introduction

Collected microdata usually contain errors, e.g., pregnant men, an average salary of 5 million euro and components of a total that do not add up to that total. Correction of such errors is often necessary to prevent flaws and inconsistencies in statistics to be published. Traditionally, the intention was often to correct all data in every detail, a costly and time-consuming process. Currently, only records with potentially influential errors are manually edited, all other records are automatically processed, a process that is known as selective editing (Granquist and Kovar, 1997; De Waal *et al.*, 2011).

When automating the data editing process, one often strives for results that closely resemble the results that would have been obtained in a manual editing process. Usually, three steps are distinguished in an automated data editing process:

- correction of systematic errors;
- localisation of random errors;
- imputation of new values to correct for random errors.

Systematic errors are errors with a known cause. A typical example is the 'thousand error', the problem that respondents provide values that are a factor 1000 larger than they actually mean. The purpose of error localisation is to designate a set of values, whose correction can produce a record that obeys all edit rules. Edit rules are formal relationships between variables that should be satisfied by the data. A simple example is that profit has to be equal to the difference between turnover and cost. Imputation is the problem of finding appropriate values for the erroneous variables. An imputed record should satisfy all edit rules.

The paradigm of Fellegi and Holt (1976) is often used for error localisation. It states that a minimum number of variables should be amended to satisfy all rules. A slight extension is the generalised Fellegi-Holt paradigm, which minimises the sum of the confidence weights of the variables to be adjusted. These weights take differences in reliability into account. Reliably measured variables can be assigned a lower probability of being designated erroneous than unreliably measured variables.

Recently, Scholtus (2014 and 2016) further generalised the Fellegi-Holt (FH) paradigm. He suggested to find the (weighted) minimal number of *edit operations* that are needed to make an observed record consistent with the edit rules. Edit operations are typical corrections that are often observed in manual data editing. Four examples have been mentioned in Scholtus (2016):

1. Replace a single value by an arbitrary value;
2. Change the sign of a value, for instance replace $x = 100$ by $x = -100$;
3. Interchange the values of two variables, e.g., replace $(x, y) = (100,200)$ by $(x, y) = (200,100)$;
4. Transfer a certain amount from one variable to another, e.g., replace $(x, y) = (100,200)$ by $(x, y) = (140,160)$.

The first operation is the same as the one used in a FH-based error localisation problem. The other operations are specific to the generalised error localisation paradigm. Contrary to the classical FH paradigm, the new paradigm allows more than one variable to be involved with one 'correction'. Another innovative aspect is a blurred distinction between correction of systematic errors, error localisation and imputation. The new approach allows for the correction of systematic errors that may not be detected by existing methods. The extended FH approach takes all available edit rules into account, while most existing methods do not use edit rules for correcting systematic errors. The new paradigm does not only detect incorrect values, like FH does, but might also provide the most suitable correction. An example of this is the change of the sign of a variable.

The solutions that can be obtained with the new paradigm can also be obtained in a traditional FH-based process. For example, an interchange of values can also be the result of a particular FH solution in which the two variables with interchanged values are designated to be erroneous. However, special corrections that are often applied in manual editing may be dominated by other corrections in a FH-based approach, because this approach aims to minimise the (weighted) number of changed values. The new paradigm is meant to ensure that often occurring manual corrections are more likely to occur in an automatic process. Hence, one could argue that the new paradigm enhances the application possibilities of an automated process.

Scholtus (2016) developed an algorithm, based on Fourier-Motzkin elimination, to solve the extended error localisation problem. He demonstrated the feasibility of this technique for a relatively small example, consisting of five variables. The application to larger problems was mentioned to be challenging from a computational point of view. One problem is that a huge number of so-called implied edits need to be generated; new rules that are implied by a combination of existing rules. For that reason, it was mentioned that there is a need for more efficient algorithms (Scholtus, 2016, p.14). Another problem is that the outcome of applying several edit operations to a record may depend on the order in which they are applied. This greatly increases the number of potential solutions to the error localisation problem.

The current paper's aim is to present an algorithm based on Mixed Integer Programming (MIP) for numerical data editing according to the new paradigm. A formulation as a MIP problem of the standard data editing problem was already given in De Waal *et al.* (2011). Because of the excellent performance of a MIP approach for standard FH-based data editing (De Jonge and Van der Loo, 2014), the approach might also be expected to be suitable for data editing according to the extended paradigm.

The remainder of this paper is organised as follows. In Section 2, we review an existing MIP formulation of error localisation under the original FH paradigm. In Section 3, we extend this MIP formulation to the generalised paradigm. We introduce two additional assumptions to simplify the problem. Sections 4 and 5 discuss the results of applying the new MIP formulation to a real dataset and a realistic dataset with simulated errors, respectively. In Section 6, we briefly discuss consistent

imputation after the error localisation problem has been solved. Some concluding remarks follow in Section 7.

Acknowledgement
The authors would like to thank Jeroen Pannekoek for his helpful comments on an earlier version of this paper.

2. A MIP formulation for Fellegi-Holt-based editing

This section shows how the standard Fellegi-Holt data editing problem can be formulated as a MIP problem. Let $\boldsymbol{x} = (x_1, \dots, x_n)'$ be a vector of n numerical variables. This record has to satisfy k edit rules that are given by

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

where $\mathbf{A} = (a_{rc})$ is a $k \times n$ matrix of coefficients and $\boldsymbol{b} = (b_1, \dots, b_k)'$ is a vector of constants. Further, $\mathbf{0}$ represents a k-vector of zeros and $\odot$ denotes a symbolic vector of operators from the set $\{\leq, =, \geq\}$. Examples of edit rules of this form that often occur in practice are balance edits (e.g., $x_1 + x_2 = x_3$) and non-negativity edits (e.g., $x_1 \geq 0$).

Let $\boldsymbol{p} = (p_1, \dots, p_n)'$ denote an originally observed record (where the *p* stands for 'preliminary'). The values after imputation are denoted by $\boldsymbol{x} = (x_1, \dots, x_n)'$. For each preliminary value there are two options: either the value remains the same (i.e., $p_i = x_i$), or the value is imputed, meaning that x_i can attain any value. We use the binary variables δ_i^{FH} $(i = 1, \dots, n)$ to indicate which variables in the original record are selected for imputation (1 indicates imputation, 0 indicates no imputation). The results for δ_i^{FH} are the solution of an error localisation problem. According to the original Fellegi-Holt paradigm (with confidence weights), the error localisation problem amounts to finding the minimal value of

$$\sum_{i=1}^{n} \delta_i^{FH} w_i^{FH},$$

Here, w_i^{FH} $(w_i^{FH} > 0)$ is the confidence weight of p_i. The outcomes of δ_i^{FH} have to be determined in such a way that at least one "imputed" record $\boldsymbol{x} = (x_1, \dots, x_n)'$ exists that satisfies the constraints in (1). Moreover, for all variables that are not imputed (i.e., $\delta_i^{FH} = 0$), x_i has to be the same as p_i. If it happens that the provisional values $\boldsymbol{p}$ already satisfy all constraints in (1), all δ_i^{FH} will be zero in the error localisation solution, meaning that no adjustment is made. Otherwise, at least one δ_i^{FH} will be one.

Different algorithms have been developed for solving the Fellegi-Holt-based error localisation problem; see De Waal et al. (2011) for an overview. In this paper, we focus on the following formulation of error localisation as a MIP problem (cf. De Jonge and Van der Loo, 2014):[1]

$$\text{Minimise}_{\delta_i^{FH}, \boldsymbol{x}} \quad \sum_{i=1}^{n} \delta_i^{FH} w_i^{FH}, \tag{2}$$

[1] It may be noted that the minimisation in (2) is over both δ_i^{FH} and $\boldsymbol{x}$, but only the former occur in the objective function. In general, there may be infinitely many feasible records $\boldsymbol{x}$ that can be obtained under the same optimal solution for δ_i^{FH}. Some MIP solvers will return one feasible $\boldsymbol{x}$ as part of their output. However, in the context of error localisation, we are only interested in the optimal δ_i^{FH}; the construction of a corresponding imputed record $\boldsymbol{x}$ is done in a later step of the data editing process.

subject to
$\mathbf{A}\boldsymbol{x} + \boldsymbol{b} \odot \mathbf{0}$,
$p_i - C_i^{FH} \leq x_i \leq p_i + C_i^{FH} \quad (i = 1, \ldots, n)$,
$C_i^{FH} = M\delta_i^{FH} \quad (i = 1, \ldots, n)$,
$\delta_i^{FH} \in \{0,1\} \quad (i = 1, \ldots, n)$,
$\boldsymbol{x} \in \mathbb{R}^n$.

Here, M is an arbitrary large number, an order of magnitude larger than the largest absolute value for any of the p_i. As noted in the introduction, the Fellegi-Holt-based error localisation problem is based on the operation "replace a single original value by an arbitrary value". The symbol C_i^{FH} is used to denote the correction that is made by the Fellegi-Holt operation.

The objective function minimises a weighted number of corrections. The first line of constraints reflects that erroneous values should be selected, such that a corrected record exists that satisfies all rules. From the second and third lines it follows that if $\delta_i^{FH} = 0$, x_i necessarily has to be the same as the preliminary value p_i. In other words: x_i is not imputed. Otherwise, if $\delta_i^{FH} = 1$, x_i can attain any value between $p_i - M$ and $p_i + M$. The use of a large constant M is common practice in MIP problems and is often referred to as the big M-method. An important advantage of using the above formulation is that efficient solvers are available for MIP problems.

As an example, consider three variables x_1, x_2 and x_3 with one constraint, stating that $x_1 + x_2 = x_3$. Suppose that all weights w_i^{FH} $(i = 1,2,3)$ are one. Consider the following record of preliminary values: $(p_1, p_2, p_3) = (5, 15, 10)$. Then, for this example, the optimisation problem in (2) becomes:

$\text{Minimise}_{\delta_i^{FH}, \boldsymbol{x}} \quad \delta_1^{FH} + \delta_2^{FH} + \delta_3^{FH}$,
subject to
$x_1 + x_2 = x_3$,
$5 - M\delta_1^{FH} \leq x_1 \leq 5 + M\delta_1^{FH}$,
$15 - M\delta_2^{FH} \leq x_2 \leq 15 + M\delta_2^{FH}$,
$10 - M\delta_3^{FH} \leq x_3 \leq 10 + M\delta_3^{FH}$,
$\delta_1^{FH}, \delta_2^{FH}, \delta_3^{FH} \in \{0,1\}$,
$\boldsymbol{x} \in \mathbb{R}^3$.

The formulation in (2) can be extended to take account of conditional rules (IF-THEN edits), where the variables in the IF and THEN clauses are subject to change. An example of such a conditional rule is: $\text{IF } x_1 > 0 \text{ THEN } x_2 > 0$. It has been shown in Daalmans (2018) that conditional rules can be written as

$$\mathbf{A}\boldsymbol{x} + \mathbf{B}\boldsymbol{y} + \boldsymbol{b} \odot \mathbf{0}, \tag{3}$$

where $\boldsymbol{y}$ is a vector of binary variables. All MIP problems that will be considered in this paper can easily be extended to allow for conditional rules by using (3) instead of (1); cf. De Jonge and Van der Loo (2014). For notational simplicity, we will work with formulas based on (1) in the remainder of this paper.

3. A MIP formulation for the new paradigm

First, we give a generic formulation of the new data editing paradigm as a MIP problem. Then, we illustrate how three specific examples of edit operations – change of sign, transfer of an amount, and interchange of two values – fit within this MIP formulation.

3.1 Generic framework

Scholtus (2016) defined the error localisation problem in terms of general paths of edit operations, where the same variable may be affected by multiple operations. In general, the order in which edit operations occur on such a path can be important. For instance, one could imagine that an interchange of two values followed by a change of sign for one of these two variables gives a different result than conducting these operations in reverse order.

For simplicity, we now add the following restrictions to the formulation of Scholtus (2016):

1. In the optimal solution all special edit operations occur as if they are applied to the original observed record.
2. In the optimal solution Fellegi-Holt operations are always applied *after* special edit operations.

In particular, these restrictions imply that each variable may be affected by at most one special edit operation *in addition to* its Fellegi-Holt operation. In this way, problems with the order-dependency of edit operations are avoided. Firstly, the second assumption fixes the ordering of special edit operations and Fellegi-Holt operations on the optimal path. It seems natural to apply special edit operations before Fellegi-Holt operations. This amounts to first finding as many errors as possible where the error mechanism is understood, and then correcting any remaining inconsistencies by means of single-value imputations (which is always possible). Secondly, the first assumption implies that the internal ordering of the special edit operations does not matter. The internal ordering of the Fellegi-Holt operations does not matter anyway, as it can be shown that these operations are commutative amongst themselves (Scholtus, 2014, pp. 36-37); in fact this explains why order-dependency was not an issue in the original error localisation problem of Section 2.

Furthermore, by making these assumptions, it will be seen below that the error localisation problem can be written as a MIP problem that generalises formula (2). Our formulation is still flexible enough to encompass many typical corrections that are often observed in practice.

Similar to Scholtus (2016), we consider edit operations of the following form:

$$g(\boldsymbol{p}, \boldsymbol{\alpha}) = \mathbf{S}\boldsymbol{\alpha} + \mathbf{T}\boldsymbol{p} + \boldsymbol{c}. \tag{4}$$

The result of $g(\boldsymbol{p}, \boldsymbol{\alpha})$ in (4) is an n-dimensional vector that includes the values of a record. The correction in (4) contains a variable part, based on the additional variables $\boldsymbol{\alpha}$, and a fixed part, $\mathbf{T}\boldsymbol{p} + \boldsymbol{c}$, that replaces existing values by a linear combination of the preliminary values $\boldsymbol{p}$ and a constant vector $\boldsymbol{c}$. The coefficients for $\boldsymbol{\alpha}$ and $\boldsymbol{p}$ are contained in the matrices $\mathbf{S} = (s_{ir})$ and $\mathbf{T} = (t_{ij})$, respectively. It should be noted that the variable part in (4) is optional: some edit operations do not require additional variables $\boldsymbol{\alpha}$. In that case the edit operation is given by $g(\boldsymbol{p}) = \mathbf{T}\boldsymbol{p} + \boldsymbol{c}$.

The special edit operations mentioned in the introduction (sign changes, interchanged values, transferred amounts) can all be written as special cases of the general form (4), as will be shown below. The standard Fellegi-Holt operation that imputes a new value in place of p_i can also be written as an operation of the form (4), with $\mathbf{S} = \boldsymbol{e}_i$ (the i-th standard basis vector in $\mathbb{R}^n$), $\mathbf{T} = \mathbf{I}_n - \boldsymbol{e}_i\boldsymbol{e}_i'$ and $\boldsymbol{c} = \mathbf{0}$. In this case, α is a single scalar that denotes the value that will be imputed.

It should be noted that in the original expression in Scholtus (2016) the correction was allowed to depend on the values of an arbitrary record $\boldsymbol{x}$, rather than on the preliminary values $\boldsymbol{p}$: $g(\boldsymbol{x}, \boldsymbol{\alpha}) = \mathbf{S}\boldsymbol{\alpha} + \mathbf{T}\boldsymbol{x} + \boldsymbol{c}$. Our more restrictive expression (4) arises from the first assumption made above. This assumption is not as restrictive as it may seem. In principle, one is free to specify as many edit operations as desired. By introducing new edit operations that combine two or more original edit operations – *in a particular order* –, one can still use these combinations in the error localisation problem. Thus, one could say that this additional assumption helps to clarify the role of edit operations in the generalised error localisation problem.

Suppose that one wants to use – in addition to the standard Fellegi-Holt operations – a set of K edit operations of this form, $g_1, \ldots, g_K$. Each special edit operation affects one or multiple values. For example, a "change of sign" affects only one variable and an "interchange of values" affects two variables. To keep track of which edit operations influence which variable(s), we define the following indicators:

$$I_{ki}^{EO} = \begin{cases} 1 & \text{if } i \in \mathcal{I}_k, \\ 0 & \text{if } i \notin \mathcal{I}_k, \end{cases} \tag{5}$$

where the superscript *EO* stands for 'edit operation' (other than Fellegi-Holt) and $\mathcal{I}_k$ is a set of variables that can be changed by operation k. The value of the indicator is one if p_i may be adjusted by operation k, otherwise its value is zero.

We obtain the following formula for g_{ki}, the function that provides the value of x_i after conducting operation g_k, given the original record $\boldsymbol{p}$ and (if relevant) the additional variables $\boldsymbol{\alpha}_k = (\alpha_{k1}, \ldots, \alpha_{km_k})'$ that occur in this operation:

$$\begin{aligned} g_{ki}\colon (\boldsymbol{p}, \boldsymbol{\alpha}_k) &\mapsto \sum_{j=1}^{n} t_{kij} p_j + \sum_{r=1}^{m_k} s_{kir}\alpha_{kr} + c_{ki}, && \text{if } i \in \mathcal{I}_k, \\ g_{ki}\colon (\boldsymbol{p}, \boldsymbol{\alpha}_k) &\mapsto p_i, && \text{if } i \notin \mathcal{I}_k. \end{aligned} \tag{6}$$

where some of the constants t_{kij}, s_{kir} and c_{ki} may be zero.

A generic MIP formulation for the extended error localisation problem is given by expression (7) below. To illustrate this formulation, special cases of this expression will be derived in Subsection 3.2 for three specific types of edit operations. Examining these special cases may be helpful to understand the contents of expression (7).

$$\text{Minimise}_{\delta_i^{FH},\delta_k^{EO},\boldsymbol{x},\boldsymbol{\alpha}_k} \quad \sum_{i=1}^{n} \delta_i^{FH} w_i^{FH} + \sum_{k=1}^{K} \delta_k^{EO} w_k^{EO}$$

subject to

$$\begin{aligned}
&\mathbf{A}\boldsymbol{x} + \boldsymbol{b} \odot \mathbf{0}, \\
&p_i - C_i^{FH} + C_i^{EO} \le x_i \le p_i + C_i^{FH} + C_i^{EO} \quad (i = 1, \dots, n), \\
&C_i^{FH} = M\delta_i^{FH} \quad (i = 1, \dots, n), \\
&C_i^{EO} = \textstyle\sum_{k=1}^{K} I_{ki}^{EO}(C_{ki}^{D} + C_{ki}^{V}) \quad (i = 1, \dots, n), \\
&C_{ki}^{D} = \delta_k^{EO}\left(\textstyle\sum_{j=1}^{n} t_{kij} p_j + c_{ki} - p_i\right) \quad (k = 1, \dots, K;\ i = 1, \dots, n), \\
&C_{ki}^{V} = \textstyle\sum_{r=1}^{m_k} s_{kir}\alpha_{kr} \quad (k = 1, \dots, K;\ i = 1, \dots, n), \\
&-M\delta_k^{EO} \le \alpha_{kr} \le M\delta_k^{EO} \quad (k = 1, \dots, K;\ r = 1, \dots, m_k), \\
&\textstyle\sum_{k=1}^{K} \delta_k^{EO} I_{ki}^{EO} \le 1 \quad (i = 1, \dots, n), \\
&\delta_i^{FH}, \delta_k^{EO} \in \{0,1\} \quad (k = 1, \dots, K;\ i = 1, \dots, n), \\
&\boldsymbol{x} \in \mathbb{R}^n, \boldsymbol{\alpha}_k \in \mathbb{R}^{m_k} \quad (k = 1, \dots, K).
\end{aligned} \tag{7}$$

The first three lines of the restrictions, without C_i^{EO}, are the same as for the original FH-problem in (2). The symbol C_i^{EO} expresses the corrections made to variable i by special editing operations. The fourth line shows that C_i^{EO} combines fixed and variable corrections, as expressed by C_{ki}^{D} and C_{ki}^{V}, respectively. Lines 5–7 are used to define the corrections. The case $\delta_k^{EO} = 1$ means that special correction k is selected to occur. Then, C_{ki}^{D} and C_{ki}^{V} attain the values as defined in (6); see the next paragraph for more details. Alternatively, if $\delta_k^{EO} = 0$, the corrections resulting from operation k are zero. The eighth line ensures that each variable is adjusted by at most one 'special' edit operation, in addition to the Fellegi-Holt operation, in line with the assumptions made above. Finally, lines 9–10 define the domains of the variables.

Regarding the correction of variable x_i, four cases can be distinguished:
Case 1: If $\delta_i^{FH} = 0$ and if $\delta_k^{EO} = 0$ for all special operations k with $I_{ki}^{EO} = 1$, variable x_i is fixed to its original value p_i, since it follows that $C_i^{FH} = C_i^{EO} = 0$. Namely, $C_i^{FH} = 0$ because of the third line of restrictions in (7) and $C_i^{EO} = 0$ because of lines 4–7 of restrictions in (7): for all k with $I_{ki}^{EO} = 1$, the fifth line of restrictions implies that $C_{ki}^{D} = 0$ and the sixth and seventh lines of restrictions imply that $C_{ki}^{V} = 0$.
Case 2: If $\delta_i^{FH} = 1$ and if $\delta_k^{EO} = 0$ for all special operations k with $I_{ki}^{EO} = 1$, variable x_i is adjusted by its Fellegi-Holt operation. In this case, it follows from the restrictions in (7) that $C_i^{FH} = M$ and $C_i^{EO} = 0$, so x_i may be imputed freely within the interval $p_i - M \le x_i \le p_i + M$.
Case 3: If $\delta_i^{FH} = 0$ and $\delta_k^{EO} = 1$ for a certain k with $I_{ki}^{EO} = 1$, then variable x_i is adjusted by edit operation g_k. In this case, $C_i^{FH} = 0$ and C_i^{EO} equals $C_{ki}^{D} + C_{ki}^{V} = \sum_{j=1}^{n} t_{kij} p_j + c_{ki} + \sum_{r=1}^{m_k} s_{kir}\alpha_{kr} - p_i$, where the additional variables α_{kr} (if present) may assume any value in the interval $-M \le \alpha_{kr} \le M$. Hence, x_i is restricted to a new value that follows from the definition of the edit operation in (6): $x_i = \sum_{j=1}^{n} t_{kij} p_j + c_{ki} + \sum_{r=1}^{m_k} s_{kir}\alpha_{kr}$. If the edit operation does not involve additional variables ($m_k = 0$), this value is uniquely specified. Otherwise, it depends on the outcomes for α_{kr}.

Case 4: If $\delta_i^{FH} = 1$ and $\delta_k^{EO} = 1$ for a certain k with $I_{ki}^{EO} = 1$, then x_i is both adjusted by edit operation g_k and imputed by its Fellegi-Holt operation. Assuming that M is sufficiently large, the restrictions on x_i will be equivalent to the ordinary Fellegi-Holt restrictions $p_i - M \leq x_i \leq p_i + M$ from Case 2. Since the Fellegi-Holt operation can impute any value for x_i, a solution of this form may be optimal only when the edit operation g_k simultaneously adjusts other variables in addition to x_i, in such a way that the Fellegi-Holt operation is not required for those other variables. Otherwise, the aforementioned solution will be dominated by a solution that only applies the Fellegi-Holt operation (Case 2).

To ensure that all edit operations can actually be obtained as a solution of the optimisation problem, a necessary condition is that the weights w_k^{EO} and w_j^{FH} are chosen such that

$$w_k^{EO} < \sum_{j \in \mathcal{J}_k} w_j^{FH}, \qquad k = 1, \dots, K. \tag{8}$$

If $w_k^{EO} \geq \sum_{j \in \mathcal{J}_k} w_j^{FH}$, the effects of edit operation g_k could be obtained at the same or lower costs by applying Fellegi-Holt operations to all individual variables in $\mathcal{J}_k$, and therefore g_k would never be used.

Other restrictions on the weights may also be necessary to ensure that each edit operation can be used. Suppose that any record that can be obtained by applying edit operation g_k could also be obtained by applying edit operation g_l. In this case, one could say that edit operation g_l "generalises" edit operation g_k. Clearly, we should then choose $w_k^{EO} < w_l^{EO}$ in order to give g_k a proper chance of being selected.

More generally, suppose that any record that can be obtained by applying a combination of edit operations $g_{k_1}, \dots, g_{k_R}$ and Fellegi-Holt operations for variables $x_{i_1}, \dots, x_{i_S}$ could also be obtained by applying a combination of edit operations $g_{l_1}, \dots, g_{l_T}$ and Fellegi-Holt operations for variables $x_{j_1}, \dots, x_{j_U}$. We should then choose weights such that

$$\sum_{r=1}^{R} w_{k_r}^{EO} + \sum_{s=1}^{S} w_{i_s}^{FH} < \sum_{t=1}^{T} w_{l_t}^{EO} + \sum_{u=1}^{U} w_{j_u}^{FH}. \tag{9}$$

The two previous weight conditions follow as special cases, with

$$\{k_1, \dots, k_R\} = \{k\}, \qquad \{i_1, \dots, i_S\} = \emptyset, \qquad \{l_1, \dots, l_T\} = \emptyset, \qquad \{j_1, \dots, j_U\} = \mathcal{J}_k,$$

and

$$\{k_1, \dots, k_R\} = \{k\}, \qquad \{i_1, \dots, i_S\} = \emptyset, \qquad \{l_1, \dots, l_T\} = \{l\}, \qquad \{j_1, \dots, j_U\} = \emptyset,$$

respectively. In practice, if the number of edit operations is large and many operations affect multiple variables, it may be difficult to find all relevant conditions on the weights.

As a final remark, we note that Scholtus (2016) also considered possible restrictions on the vector $\boldsymbol{\alpha}_k$ of additional variables that occur in edit operation g_k, in the form of a set of linear (in)equalities:

$$\mathbf{R}_k \boldsymbol{\alpha}_k + \boldsymbol{d}_k \odot \mathbf{0}, \tag{10}$$

where again $\odot$ contains operators from the set $\{\leq, =, \geq\}$. In principle, these restrictions could be incorporated into the MIP problem (7). For ease of presentation we will not consider these additional constraints in the remainder of this paper.

3.2 Specific editing operations

This subsection considers the special operations that were mentioned in the introduction:

- Change Sign (CS);
- Interchange Values (CV);
- Transfer Amount (TA).

Below we explain how these operations can be formulated within the framework described in (7).

Change Sign

The edit operation that changes the sign of variable x_i is given by

$$g_i^{CS}: p_i \longmapsto -p_i. \tag{11}$$

Assuming for simplicity that every variable can be affected by a sign change, we have $K = n$ additional operations with $\mathcal{I}_i^{CS} = \{i\}$ for $i = 1, \dots, n$, so each variable has its own operation. The subscript k will be dropped for ease of presentation.

It follows that $t_{ii} = -1$, $t_{jj} = 1$ if $j \neq i$, and all other elements of $\mathbf{T}$ are zero, while $\boldsymbol{c} = \mathbf{0}$. Additional variables $\boldsymbol{\alpha}$ are not required for a change of sign. The expression $\sum_{j=1}^{n} t_{ij} p_j + c_i - p_i$ for the correction in (7) reduces to $-2p_i$. That is, $-2p_i$ is added to preliminary value p_i, which replaces p_i by $-p_i$.

If sign changes are the only special edit operations, the MIP in (7) simplifies to:

$$\begin{aligned}
&\text{Minimise}_{\delta_i^{FH}, \delta_k^{CS}, \boldsymbol{x}} \quad \sum_{i=1}^{n} \delta_i^{FH} w_i^{FH} + \sum_{i=1}^{n} \delta_i^{CS} w_i^{CS} \\
&\text{subject to} \\
&\mathbf{A}\boldsymbol{x} + \boldsymbol{b} \odot \mathbf{0}, \\
&p_i - C_i^{FH} + C_i^{CS} \leq x_i \leq p_i + C_i^{FH} + C_i^{CS} \quad (i = 1, \dots, n), \\
&C_i^{FH} = M\delta_i^{FH} \quad (i = 1, \dots, n), \\
&C_i^{CS} = -2p_i\delta_i^{CS} \quad (i = 1, \dots, n), \\
&\delta_i^{FH}, \delta_i^{CS} \in \{0,1\} \quad (i = 1, \dots, n), \\
&\boldsymbol{x} \in \mathbb{R}^n.
\end{aligned} \tag{12}$$

The superscript *CS* is used for change sign. If the 'change sign' operation applies to x_i, i.e., if $\delta_i^{CS} = 1$ (and $\delta_i^{FH} = 0$), we obtain $C_i^{FH} = 0$ and $C_i^{CS} = -2p_i$ and it follows that the lower and upper bounds for x_i are both set to $-p_i$.

According to the weight condition formula (8), we should choose $w_i^{CS} < w_i^{FH}$ for all $i = 1, \dots, n$. Otherwise, a change of sign operator will never be optimal in a solution of (12), as it could alternatively be obtained by a FH-correction. If, in a practical application, not all variables are eligible for a change of sign, only the relevant subset of all δ_i^{CS} variables is included in the above problem.

As a further illustration, consider again the three-variable example from Section 2 and suppose that we now also allow for a 'change sign' operation for variable x_2. The weight of this special operation is chosen to be $1/2$. The MIP formulation in (12) for this example can be written as:

$\text{Minimise}_{\delta_i^{FH},\delta_2^{CS},\boldsymbol{x}}\ \ \delta_1^{FH}+\delta_2^{FH}+\delta_3^{FH}+\delta_2^{CS}/2$,
subject to
$x_1+x_2=x_3$,
$5-M\delta_1^{FH}\le x_1\le 5+M\delta_1^{FH}$,
$15-M\delta_2^{FH}-30\delta_2^{CS}\le x_2\le 15+M\delta_2^{FH}-30\delta_2^{CS}$,
$10-M\delta_3^{FH}\le x_3\le 10+M\delta_3^{FH}$,
$\delta_1^{FH},\delta_2^{FH},\delta_3^{FH},\delta_2^{CS}\in\{0,1\}$,
$\boldsymbol{x}\in\mathbb{R}^3$.

In case $\delta_2^{CS}=1$ (and $\delta_2^{FH}=0$), x_2 is bounded by -15, the negative of its provisional value.

Interchange Values

The edit operation that interchanges the values of x_i and x_j is given by

$$g_{ij}^{CV}:(p_i,p_j)\longmapsto(p_j,p_i). \tag{13}$$

For ease of notation, an interchange of i and j will be denoted by $\mathcal{I}_{ij}^{CV}$, where *CV* stands for (inter)change values. Because an interchange of i and j has the same effect as an interchange of j and i, we can restrict our attention to pairs with $i<j$. Assuming for simplicity that every pair of variables can be interchanged, we have $K=n(n-1)/2$ additional operations with $\mathcal{I}_{ij}^{CV}=\{i,j\}$ for all $i<j$.

The CV operation g_{ij}^{CV} is obtained by choosing $t_{ij}=1$, $t_{ji}=1$, $t_{ii}=t_{jj}=0$, $t_{ll}=1$ if $l\notin\{i,j\}$, setting all other elements of $\mathbf{T}$ equal to zero, and choosing $\boldsymbol{c}=\mathbf{0}$ in (6). Additional variables $\boldsymbol{\alpha}$ are, again, not needed. The expression $\sum_{j=1}^{n}t_{ij}p_j+c_i-p_i$ for the correction in (7) reduces to p_j-p_i for variable x_i, and p_i-p_j for variable x_j.

The MIP formulation in (7) for a problem with FH and CV operations is given by:

$$\begin{aligned}
&\text{Minimise}_{\delta_i^{FH},\delta_{ij}^{CV},\boldsymbol{x}}\ \ \sum_{i=1}^{n}\delta_i^{FH}w_i^{FH}+\sum_{i=1}^{n-1}\sum_{j=i+1}^{n}\delta_{ij}^{CV}w_{ij}^{CV}\\
&\text{subject to}\\
&\mathbf{A}\boldsymbol{x}+\boldsymbol{b}\odot\mathbf{0},\\
&p_i-C_i^{FH}+C_i^{CV}\le x_i\le p_i+C_i^{FH}+C_i^{CV}\quad(i=1,\dots,n),\\
&C_i^{FH}=M\delta_i^{FH}\quad(i=1,\dots,n),\\
&C_i^{CV}=\textstyle\sum_{j=i+1}^{n}\delta_{ij}^{CV}\left(p_j-p_i\right)+\sum_{h=1}^{i-1}\delta_{hi}^{CV}(p_h-p_i)\quad(i=1,\dots,n),\\
&\textstyle\sum_{j=i+1}^{n}\delta_{ij}^{CV}+\sum_{h=1}^{i-1}\delta_{hi}^{CV}\le 1\quad(i=1,\dots,n),\\
&\delta_i^{FH},\delta_{hj}^{CV}\in\{0,1\}\quad(i=1,\dots,n;h=1,\dots,n-1;j=h+1,\dots,n),\\
&\boldsymbol{x}\in\mathbb{R}^n.
\end{aligned}\tag{14}$$

The first three lines of constraints are similar as before. The fourth line implies that $\delta_{ij}^{CV}=1$ means that preliminary values p_i and p_j are interchanged. The fifth line ensures that each variable is subject to at most one CV operation at the same time.

According to the weight condition formula (8), we should choose $w_{ij}^{CV}<w_i^{FH}+w_j^{FH}$ for all relevant i and j, otherwise a CV operation will never be optimal. All pairs of variables can be interchanged in formulation (14). If, in practice, only certain pairs of variables are likely to have interchanging values between them, only the relevant pairs should be included in the model.

Consider the same three-variable example as before and suppose that we now allow for an operation that interchanges the variables x_2 and x_3, with weight one. The MIP formulation in (14) for this example yields:

$$
\begin{aligned}
&\text{Minimise}_{\delta_i^{FH},\delta_{23}^{CV},\boldsymbol{x}} \quad \delta_1^{FH}+\delta_2^{FH}+\delta_3^{FH}+\delta_{23}^{CV}\ ,\\
&\text{subject to}\\
&x_1+x_2=x_3,\\
&5-M\delta_1^{FH}\le x_1\le 5+M\delta_1^{FH},\\
&15-M\delta_2^{FH}-5\delta_{23}^{CV}\le x_2\le 15+M\delta_2^{FH}-5\delta_{23}^{CV},\\
&10-M\delta_3^{FH}+5\delta_{23}^{CV}\le x_3\le 10+M\delta_3^{FH}+5\delta_{23}^{CV},\\
&\delta_1^{FH},\delta_2^{FH},\delta_3^{FH},\delta_{23}^{CV}\in\{0,1\},\\
&\boldsymbol{x}\in\mathbb{R}^3.
\end{aligned}
$$

If $\delta_{23}^{CV}=1$ (and $\delta_2^{FH}=\ \delta_3^{FH}=0$), x_2 and x_3 interchange values.

Transfer Amount

The edit operation that transfers an arbitrary amount between x_i and x_j is given by

$$g_{ij}^{TA}\colon (p_i,p_j)\mapsto (p_i-\alpha_{ij}^{TA},p_j+\alpha_{ij}^{TA}). \tag{15}$$

where superscript *TA* stands for transfer amount. Assuming for simplicity that transfers are possible between every pair of variables, we have $K=n(n-1)/2$ additional operations with $\mathcal{I}_{ij}^{TA}=\{i,j\}$ for all $i<j$. Again, we can restrict attention to pairs with $i<j$ since the operation is symmetric; a transfer from i to j can also be written as transferring a negative amount from j to i.

Contrary to the previous two operations, the operation that transfers an arbitrary amount involves a variable part, as the transferred amount is an endogenous variable α_{ij}^{TA}. The matrix $\mathbf{T}$ is the identity matrix and $\boldsymbol{c}=\mathbf{0}$. The coefficient s_{i1} for α_{ij}^{TA} is -1, because the amount α_{ij}^{TA} is subtracted from p_i. Similarly, the coefficient s_{j1} is $+1$, since α_{ij}^{TA} is added to p_j.

The MIP formulation in (7) for a problem with FH and TA operations simplifies to:

$$
\begin{aligned}
&\text{Minimise}_{\delta_i^{FH},\delta_{ij}^{TA},\boldsymbol{x},\alpha_{ij}^{TA}} \quad \sum_{i=1}^{n}\delta_i^{FH}w_i^{FH}+\sum_{i=1}^{n-1}\sum_{j=i+1}^{n}\delta_{ij}^{TA}w_{ij}^{TA}\\
&\text{subject to}\\
&\mathbf{A}\boldsymbol{x}+\boldsymbol{b}\odot\mathbf{0},\\
&p_i-C_i^{FH}+C_i^{TA}\le x_i\le p_i+C_i^{FH}+C_i^{TA}\quad (i=1,\dots,n),\\
&C_i^{FH}=M\delta_i^{FH}\quad (i=1,\dots,n),\\
&C_i^{TA}=-\textstyle\sum_{j=i+1}^{n}\alpha_{ij}^{TA}+\sum_{h=1}^{i-1}\alpha_{hi}^{TA}\quad (i=1,\dots,n),\\
&-M\delta_{ij}^{TA}\le\alpha_{ij}^{TA}\le M\delta_{ij}^{TA}\quad (i=1,\dots,n-1;\ j=i+1,\dots,n),\\
&\textstyle\sum_{j=i+1}^{n}\delta_{ij}^{TA}+\sum_{h=1}^{i-1}\delta_{hi}^{TA}\le 1\quad (i=1,\dots,n),\\
&\delta_i^{FH},\delta_{hj}^{TA}\in\{0,1\}\quad (i=1,\dots,n;h=1,\dots,n-1;j=h+1,\dots,n),\\
&\boldsymbol{x}\in\mathbb{R}^n,\alpha_{ij}^{TA}\in\mathbb{R}\quad (i=1,\dots,n-1;\ j=i+1,\dots,n).
\end{aligned}
\tag{16}
$$

The first three lines of the constraints are similar as before. The TA operation is expressed in the fourth and fifth lines. It follows that if $\delta_{ij}^{TA}=1$, the transferred amount α_{ij}^{TA} can attain any value between $-M$ and $+M$. The fourth line shows that α_{ij}^{TA} is subtracted from p_i and added to p_j. It also follows that if $\delta_{ij}^{TA}=0$, the transferred amount α_{ij}^{TA} necessarily has to be zero, since its lower and upper bounds

in the fifth line are both zero. The sixth line is used to ensure that at most one TA operation can be applied to each variable.

According to the weight condition formula (8), we should choose $w_{ij}^{TA} < w_i^{FH} + w_j^{FH}$ for relevant i and j, otherwise TA operations cannot appear in an optimal solution. If the interchanging values operations were also added to the problem, the weights should also satisfy $w_{ij}^{CV} < w_{ij}^{TA}$ for all relevant i and j, because transferring an amount between two variables "generalises" interchanging the values of two variables. That is, choosing $\alpha_{ij}^{TA} = p_i - p_j$ reduces a transfer operation to an interchange operation.

Again, it should be noted that in practice, often only certain pairs of variables are likely to have transferred values between them. In that case, only the relevant subset of all δ_{ij}^{TA}s needs to be included in the above problem.

As a final illustration, suppose that in the previous three-variable example we now allow for an operation that transfers an amount between the variables x_1 and x_2, with weight one. The MIP formulation in (16) then yields:

$$\text{Minimise}_{\delta_i^{FH},\delta_{12}^{TA},\boldsymbol{x},\alpha_{ij}^{TA}} \quad \delta_1^{FH} + \delta_2^{FH} + \delta_3^{FH} + \delta_{12}^{TA},$$

subject to

$$x_1 + x_2 = x_3,$$
$$5 - M\delta_1^{FH} - \alpha_{12}^{TA} \le x_1 \le 5 + M\delta_1^{FH} - \alpha_{12}^{TA},$$
$$15 - M\delta_2^{FH} + \alpha_{12}^{TA} \le x_2 \le 15 + M\delta_2^{FH} + \alpha_{12}^{TA},$$
$$10 - M\delta_3^{FH} \le x_3 \le 10 + M\delta_3^{FH},$$
$$-M\delta_{12}^{TA} \le \alpha_{12}^{TA} \le M\delta_{12}^{TA},$$
$$\delta_1^{FH}, \delta_2^{FH}, \delta_3^{FH}, \delta_{12}^{TA} \in \{0,1\},$$
$$\boldsymbol{x} \in \mathbb{R}^3, \alpha_{12}^{TA} \in \mathbb{R}.$$

If $\delta_{12}^{TA} = 1$ an amount α_{12}^{TA} is transferred from p_1 to p_2.

4. Application

4.1 Aim and setup

A simulation was conducted to test the feasibility and usefulness of the new data editing paradigm. In particular, it was tested whether the new approach is feasible, using a standard desktop computer and a freely available MIP solver. Further, it was verified whether the results of the new paradigm approximate the manually edited data better than the outcomes of a FH-editing process.

A prototype implementation of the algorithms in Section 3 was created in R. This prototype made use of the existing functionality for Fellegi-Holt-based automatic data editing, as available in the editrules package (Van der Loo and De Jonge, 2012; De Jonge and Van der Loo, 2014). To solve the MIP problem, this package uses the lpSolveAPI package (Konis, 2014). All applications were conducted on a desktop PC with a 2.8 GHz CPU under Windows 7. The data that were used for the application come from a Dutch structural business survey (SBS), the 2012 survey on trade and transport. From this survey 377 records were selected for which both manually and automatically edited data are available. The data come from 16 different (but similar) questionnaires, for each of which a set of edits is available from the production process. The total number of edit rules ranges from 98 to 130 and a total number of 140 unique variables appear in all questionnaires. Typically, 20 percent of edit rules are conditional (IF-THEN). Confidence weights w_i^{FH} are available for all variables; these are the weights that are used in production.

A selection of 15 special operations was determined on the basis of occurrence in the manually edited data: two CS operations, four CV operations and nine TA operations. Table 1 lists these edit operations with a description of the variables that are involved in these operations, as well as the number of times each operation was found in the manually edited data (column 'Freq.').

In manual data editing, the most frequent occurrence of a 'transfer amount' (TA) correction was for 10% of all records. For the CS and CV operations these percentages are 2% and 4%, respectively. This demonstrates that the special operations are actually important in manual data editing, most especially the TA operation. Note that there is no overlap between the interchange and transfer operations in Table 1. That is, the two variables involved with a CV operation do not occur together in a TA operation.

The weights for the special operations were chosen in an ad hoc manner. For a CS operation the weight is half of the Fellegi-Holt weight for the same variable ($w_i^{CS} = 0.5w_i^{FH}$). For the other two special operations (CV and TA) the weight is 0.5 higher than the largest Fellegi-Holt weight of the two variables involved ($w_{ij}^{TA} = w_{ij}^{CV} = \max\{w_i^{FH}, w_j^{FH}\} + 0.5$). Because integer-valued confidence weights are used in statistical production, the weight of a CV or TA action is smaller than the sum of the

FH-weights of the underlying variables ($w_{ij}^{CS} < w_i^{FH} + w_j^{FH}$); a necessary condition for the occurrence of a CV or TA operation in the optimal solution of a data editing problem (see Section 3). To limit the effects of computational performance problems, maximum computation time for each record was set to 60 seconds.

Table 1. Occurrence of special editing operations in manually edited data

	Involved variable(s)	Variable description	Freq.
Change Sign (CS)			
CS1	FINREST100000	Financial result	3 (1%)
CS2	RESULTS130000	Pre-tax result	1 (0%)
Interchange Values (CV)			
CV1	INKWRDE132000	Costs of outsourced work	15
	INKWRDE110000	Costs of purchased goods for trade	(4%)
CV2	INKWRDE110000	Costs of purchased goods for trade	8
	INKWRDE133000	Costs of other purchases	(2%)
CV3	OMZETPS213009	Turnover from services (former main activity)	7
	OMZETPS213900	Turnover from other services	(2%)
CV4	INKWRDE120000	Costs of purchased raw and auxiliary materials	7
	INKWRDE110000	Costs of purchased goods for trade	(2%)
Transfer Amount (TA)			
TA1	LOONSOM121100	Employers share of social benefit costs	39
	LOONSOM110002	Gross salary costs	(10%)
TA2	OMZETPS213000	Turnover from services	15
	OMZETPS212200	Turnover from retail trade	(4%)
TA3	OMZETPS213000	Turnover from services	12
	OMZETPS212100	Turnover from wholesale trade	(3%)
TA4	INKWRDE120000	Costs of purchased raw and auxiliary materials	12
	INKWRDE112000	Costs of purchased goods for retail trade	(3%)
TA5	PERSONS110000	Employed persons (total FTE)	11
	PERSONS111000	Employed persons (total number)	(3%)
TA6	OMZETPS213000	Turnover from services	11
	OMZETPS212000	Turnover from trade	(3%)
TA7	BEDRLST345400	Payments to temp agencies / other hired staff	9
	LOONSOM110002	Gross salary costs	(2%)
TA8	BEDRLST345900	Other staff costs	9
	LOONSOM110002	Gross salary costs	(2%)
TA9	INKWRDE120000	Costs of purchased raw and auxiliary materials	9
	INKWRDE111000	Costs of purchased goods for wholesale trade	(2%)

4.2 Main results

Table 2 below summarises the main results. Most importantly, it shows the feasibility of the extended paradigm for a real-life problem. Computational performance is somewhat worse than for the original FH-approach: total processing time is more than twice as large and a larger share of records cannot be processed within 60 seconds. However, the computational performance of the MIP formulation is much better than would be expected of the original error localisation algorithm proposed by Scholtus (2014), for a data set with this number of variables and edit rules.

Table 2 also shows that the special data editing operations do appear in the automatic data editing solutions. The appearance is however less often than in the manually edited data. Special operations occurred for 14.9% of all records for the automatically edited data and for 36.3% of the manually edited data (the latter is not shown in the table).

Table 2. Performance measures for automatic editing

Measure	Fellegi-Holt (original)	Extended
Total computation time (in s.)	444.42	1043.71
Number of records processed within 60 s.	373 (98.9%)	365 (96.8%)
Number of records with special operation(s)	0	56 (14.9%)

4.3 Detailed results

In this subsection we compare the results of the manually and automatically edited data for each of the special operations. The comparison is based on the set of 365 records that were successfully processed in automatic data editing (i.e., within 60 seconds).

Indicators have been computed that count the number of special corrections that are

- correctly detected, i.e., special operations in the manually edited data that have also been detected by automatic methods;
- missed, i.e., special operations in the manually edited data that were not identified by automatic methods;
- wrongly detected, i.e., automatically determined special operations that did not occur in manual editing.

Note that manually edited data are used as a benchmark; these data are considered correct, which is a quite usual assumption in a comparison between automated and manual data editing.

A comparison is made between the manually edited data and the automatically edited data according to the FH approach and the extended paradigm. The FH edited data are examined after imputation, since special operations affect both error localisation and imputation. Imputation was carried out by a ratio method that is

used for production. Subsequently, the imputed data were further adjusted to ensure that these are consistent with the edit rules (see also Section 6).

4.3.1 Change of sign

As shown above in Table 1, we consider "Change of sign" operations for two variables. These operations were observed for 4 records of the manually edited data. However, one record could not be automatically processed within 60 seconds. Therefore, we only consider three records below.

Table 3. Detailed results for edit operations that change the sign of a variable

Count of operations	FH edited	Extended paradigm
Sign change in manually edited data		
- Correctly detected	1	3
- Missed	2	0
Total	**3**	**3**
Wrongly detected	0	0

The results show the three records with CS operations in the manually edited data have been detected by following the extended paradigm. On the contrary, in a traditional FH-process two cases were not detected. In one of these cases the variable subject to a change of sign was identified in error localisation, but a different imputation was conducted than a change of sign. The aforementioned results clearly show the added value of the extended paradigm.

4.3.2 Change values and transfer amounts

In this subsection we present the results for the Change Values and Transfer Amounts operations together. In the presentation of the results a distinction is made between:

- instances that are not observed in FH-based results;
- instances of operations that appear at least once in the FH-based results.

The former group is considered first. Table 4 below summarises results for two CV operations and seven TA operations. These operations were made in 116 records of the manually data, but not in any record of a FH process. It follows that the operations were not performed by the extended method either, with the exception of just one case.

Table 4. Detailed results for edit operations that interchange or transfer an amount between two variables (not observed in FH-based results)

Count of operations	FH edited	Extended paradigm
Special operation in manually edited data		
- Correctly detected	0	1
- Missed	116	115
Total	**116**	**116**
Wrongly detected	0	1

The main reason for the absence of certain special operations in the automatically edited data is the lack of an edit rule, which causes an automatic correction to occur. An example is operation CV4: an interchange of values between INKWRDE120000 (Costs of purchased raw and auxiliary materials) and INKWRDE110000 (Costs of purchased goods for trade). An editor might know from the other data provided by a respondent whether the unit is involved with trade. Hence, an editor can assess the odds that one value is higher than the other. In automatic data editing there is not a constraint on the relation between INKWRDE120000 and INKWRDE110000. Therefore, there is no direct reason for an interchange of values.

We now look at instances of special operations that can be observed in FH-edited data. These special operation include two cases of CV and two cases of TA.

Table 5. Detailed results for edit operations that interchange or transfer an amount between two variables (observed at least once in FH-based results)

Count of operations	FH edited	Extended paradigm
Special operation in manually edited data		
- Correctly detected	17	28
- Missed	24	13
Total	**41**	**41**
Wrongly detected	10	23

The table above shows that the extended paradigm reproduces more special data editing corrections than the original FH-approach. This is however offset by a larger number of “wrongly detected operations”, operations that are conducted in the automatically edited data but not in the manually edited data.

A generic reason for differences in results between manual and automatic data editing is that different corrections are chosen to achieve compliance with the rules. Sometimes consistency is achieved by a different special editing operation than the one applied in manual editing.

A further investigation of particular records might lead to particular solutions to obtain results that are more similar to manually edited data. One solution might be a more advanced weighting scheme.

One could however not expect that differences in results between manual and automatic data editing can disappear altogether. The outcomes of both approaches partly rely on arbitrary choices. For instance, in automatic error localisation many optimal solutions may exist for one record, i.e., solutions with the same value for the objective function. In that case, one of these optimal solutions is randomly selected. It might be expected that different choices exist in manual editing as well.

5. Simulation study

5.1 Aim and setup

The results in the previous section were obtained for a single data set with real errors. To gain more insight into the possibility of using the extended Fellegi-Holt paradigm to improve the quality of automatic editing, we obtained additional results by generating various types of simulated errors in a real data set.

As a starting point for the simulations, we used the production-edited data of the 2011 SBS survey for four industries in the sector Transport. This dataset consists of 1080 records that satisfy all edit rules. For the purpose of this study, these data were considered error-free. In terms of number of variables and edit rules, they are very similar to the data used in Section 4. We added random errors to the original dataset by drawing from a set of potential error generating mechanisms $\mathcal{E}$ through the following algorithm.

For each record $i = 1, \ldots, 1080$, do the following:

1. Draw a random number k_i of errors to add (where $1 \leq k_i \leq k_{\max}$).
2. Define $\mathcal{E}_i = \mathcal{E}$.
3. For each $j = 1, \ldots, k_i$, do the following:
    a. Draw a random type of error $e \in \mathcal{E}_i$.
    b. Remove e from $\mathcal{E}_i$, i.e., replace $\mathcal{E}_i$ by $\mathcal{E}_i \backslash \{e\}$.
    c. Check whether an error of the form e is meaningful for the current record. If so, apply it to this record. If not, return to Step 3.a.

We conducted several simulations with different settings for these steps (see below). Some details were the same in all simulations. Firstly, the probability distribution for drawing the number of errors in Step 1 was always based on the empirical distribution of the number of erroneous values found in a sample of manually edited records. For this, we used a subset of the sample of 377 records from Section 4, restricted to the four Transport industries and to records that contain at least one error. This left 154 records. The empirical distribution of the number of erroneous values for these units is shown in Table 6. The probability distribution used in Step 1 was obtained by truncating the empirical distribution in Table 6 to $[1, k_{\max}]$, where $k_{\max}$ depended on the setting.

Table 6. Empirical distribution of number of erroneous variables in edited sample

Number	1	2	3	4	5	6	7	8	9	10	>10
Freq.	26	30	8	13	12	16	9	11	7	3	19

Secondly, the potential error generating mechanisms $\mathcal{E}$ included errors based on the standard Fellegi-Holt edit operations (i.e., a random error in a single variable) and the fifteen special edit operations listed in Table 1. Moreover, we restricted the set of potential errors to only errors that affect at least one variable in $V_{\geq 2}$, where $V_{\geq 2}$

denotes the subset of all variables that are involved in at least two edit rules besides non-negativity edits of the form $x \geq 0$. The reason for this restriction is that in this study we want to focus on the effect of the error localisation criterion on the quality of automatic editing. In general, the outcome of automatic editing also depends on other factors, including the strength of the edit rules that have been defined. By focussing on a subset of the variables for which relatively strong edits exist, we hope to obtain results that are mainly informative about the effect of the minimisation criterion on the quality of error localisation, rather than about, e.g., the absence of meaningful edit rules. (It should be noted that different industries and size classes within Transport have slightly different questionnaires with different sets of edit rules, so that $V_{\geq 2}$ - and therefore also $\mathcal{E}$ - is not identical for all units.)

Some more details on the different types of errors now follow.

- *Errors corresponding to CS operations (sign errors)*: These errors were generated by multiplying the original value of a variable by -1. If the original value in a record happened to be zero, this error was rejected in Step 3.c.
- *Errors corresponding to CV operations (interchanged values)*: These errors were generated by interchanging the original values of two variables. If the two original values happened to be the same in a record, the error was rejected in Step 3.c.
- *Errors corresponding to TA operations (transferred amounts)*: For two variables with original values v_1 and v_2, an error of this form was generated by drawing a random non-zero value a from the uniform distribution on the set of integers
$$\{-v_2, -v_2 + 1, \dots, -1, 1, \dots, v_1 - 1, v_1\}$$
and defining the new values $v_1' = v_1 - a$ and $v_2' = v_2 + a$, so that $v_1' \geq 0$ and $v_2' \geq 0$ (as required by the edit rules for these variables). The error was rejected in Step 3.c for records with either $v_1 = v_2$ or $v_1 \leq 1$ or $v_2 \leq 1$.
- *Errors corresponding to FH operations*: To create a random error in a single variable, the original value v was replaced by the following procedure:
    a. If $v > 0$, the new value $v' = v + a$, where a is either (with 50% probability) a value from a uniform distribution on the set of positive integers
    $$\{\max\{1, 0.5v\}, \dots, 9v\}$$
    or (with 50% probability) a value from a uniform distribution on the set of negative integers
    $$\{-v, \dots, \min\{-1, -0.5v\}\}.$$
    In this way, we aimed to draw new values that are moderately different from the original values and in particular do not cause a positive value to become negative. An exception was made if the variable does not have to satisfy a non-negativity edit; here, the set of possible negative values of a was chosen wider:
    $$\{-11v, \dots, \min\{-1, -0.5v\}\}.$$
    b. If $v < 0$, the new value $v' = v + a$, where a is either (with 50% probability) a positive value from a uniform distribution on the set
    $$\{\max\{1, -0.5v\}, \dots, -11v\}$$
    or (with 50% probability) a negative value from the uniform distribution on the set
    $$\{9v, \dots, \min\{-1, 0.5v\}\}.$$

c. If $v = 0$, the new value $v' = a$, where a is the mean of all non-zero values in the current record.

An error of this form was never rejected in Step 3.c. However, if it happened that $v' = -v$, the error was re-classified as a sign error corresponding to a CS operation.

We conducted a simulation study with four different settings in the above procedure:

- Setting 1: In Step 1, the maximum number of errors $k_{\max} = 5$. In each iteration of Step 3, all error types in $\mathcal{E}_i$ have the same drawing probability.
- Setting 2: As Setting 1, but in each iteration of Step 3, error types related to special edit operations have a drawing probability that is three times as large as that of the error types related to FH operations.
- Setting 3: As Setting 2, but for each drawn FH operation there is a 10% probability of introducing a missing value.
- Setting 4: As Setting 2, but with $k_{\max} = 10$.

For Settings 1, 2 and 3, five datasets with random errors were generated according to the above procedure. For Setting 4, we ran error localisation on only one dataset, because it required a much longer computation time (see Subsection 5.3).

We tested error localisation according to the original Fellegi-Holt paradigm and according to the generalised paradigm, with all 15 special edit operations in Table 1 included. In both cases, all confidence weights were chosen equal to 1. For Setting 1, this is the natural choice, as all error types were equally likely to occur. For Settings 2, 3 and 4, we also tested a version of the generalised paradigm with 'optimal' weights. According to Scholtus (2016), the generalised Fellegi-Holt paradigm can be seen as an approximate maximum-likelihood procedure if the confidence weight of the edit operation that corresponds to error $e_g \in \mathcal{E}$ is chosen as

$$w_g = -\log\left(\frac{\pi_g}{1-\pi_g}\right), \tag{17}$$

where π_g denotes the probability that error e_g occurs. For Settings 2, 3 and 4, we used weights based on approximate probabilities that were computed as follows. Let $\mathcal{E} = \mathcal{E}_{FH} \cup \mathcal{E}_{EO}$, where $\mathcal{E}_{FH}$ denotes the set of Fellegi-Holt operations and $\mathcal{E}_{EO}$ the set of special edit operations. In addition, let $\bar{k}$ denote the mean of the truncated distribution of the number of errors, given $k_{\max}$, based on Table 6. The approximate probability of an error that corresponds to a Fellegi-Holt operation is given by

$$\pi_{FH} = \frac{\bar{k}}{|\mathcal{E}_{FH}| + 3|\mathcal{E}_{EO}|}$$

and the approximate probability of an error that corresponds to a special edit operation is given by

$$\pi_{EO} = \frac{3\bar{k}}{|\mathcal{E}_{FH}| + 3|\mathcal{E}_{EO}|} = 3\pi_{FH}.$$

Note that

$$\sum_{e_g \in \mathcal{E}} \pi_g = |\mathcal{E}_{FH}|\pi_{FH} + |\mathcal{E}_{EO}|\pi_{EO} = \bar{k},$$

the expected number of errors. We obtained the 'optimal' weights by substituting these probabilities into (17). These weights automatically satisfy all relevant weight conditions from Section 3, since FH operations were assigned larger weights than special operations and also there was no overlap between CV and TA operations.

As in Section 4, error localisation was done using a prototype implementation in R, based on the editrules package. The maximum computation time per record was again set to 60 seconds.

5.2 Evaluation measures

To evaluate the quality of error localisation, we compared the supposed errors that were found by the algorithms with the true errors in each dataset. Using the fact that these true errors were known in this simulation study, we compiled the following contingency table of errors that occurred and associated edit operations that were selected, across all combinations of records and edit operations. (So each record i contributes $n + |\mathcal{E}_{EO}|$ counts to this table.)

	edit operation was selected	**edit operation was not selected**
error occurred	TP	FN
error did not occur	FP	TN

The following three indicators were computed from this table: the proportion of false negatives,

$$\alpha = \frac{FN}{TP + FN};$$

the proportion of false positives,

$$\beta = \frac{FP}{FP + TN};$$

and the overall proportion of wrong decisions,

$$\delta = \frac{FN + FP}{TP + FN + FP + TN}.$$

A good error localisation method should have low values on all three indicators (cf. De Waal *et al.*, 2011, pp. 410-411). The quantities $1 - \alpha$ and $1 - \beta$ are sometimes referred to as *sensitivity* and *specificity*, respectively.

These indicators measure the quality of error localisation at the level of individual errors. As a fourth indicator on the level of whole records, we computed $\rho^c = 1 - \rho$, where ρ denotes the proportion of records in the dataset for which exactly the right solution was found (i.e., the solution that chooses all the right edit operations and only these). Again, lower values of ρ^c indicate that the quality of error localisation is better.

The above indicators refer to the correct use of edit operations. This places the original Fellegi-Holt approach at a disadvantage, since this method does not use all relevant edit operations. It might sometimes happen that the Fellegi-Holt approach does not choose the right edit operation but still correctly identifies (some of) the involved variables as erroneous. To account for this, we looked at a second contingency table of true erroneous variables and variables that were identified as erroneous. (Here, a special edit operation is supposed to identify all affected

variables as erroneous.) From this table, indicators α, β and δ were computed in the same way as before. A second indicator ρ^c was also defined in terms of erroneous values. The same set of evaluation measures was used by Scholtus (2016).

5.3 Results

Table 7 shows the average (mean and median) computing time per record in seconds for each setting and each error localisation method. It also shows the percentage of records for which an optimal solution was found (column 'optimal'), for which a possibly suboptimal solution was found (column 'suboptim.') and for which no solution was found within the time limit of 60 seconds (column 'not solved'). For Settings 1, 2 and 3, each cell in the tables in this subsection shows the mean value across five simulated datasets, as well as the minimal and maximal value in brackets. As noted above, only one dataset was used for Setting 4 to save time.

Table 7. Results of error localisation on simulated data: computing times per record and percentages of solutions found within 60 seconds

	computing time per rec.		**solutions found within 60 s**		
	mean (s)	**median (s)**	**optimal**	**suboptim.**	**not solved**
Setting 1: maximum 5 errors					
FH original	1.94 (1.52; 2.22)	0.52 (0.48; 0.55)	99.6% (99.3; 99.9)	0.0% (0.0; 0.0)	0.4% (0.1; 0.7)
FH extended	2.94 (2.40; 3.49)	0.56 (0.50; 0.61)	98.8% (98.4; 99.3)	0.1% (0.0; 0.1)	1.2% (0.6; 1.6)
Setting 2: maximum 5 errors, unequal drawing probabilities					
FH original	1.71 (1.53; 2.01)	0.49 (0.48; 0.50)	99.8% (99.6; 99.9)	0.0% (0.0; 0.0)	0.2% (0.1; 0.4)
FH extended	2.30 (2.20; 2.46)	0.51 (0.49; 0.53)	99.4% (99.1; 99.9)	0.0% (0.0; 0.1)	0.6% (0.1; 0.9)
FH extended (weighted)	3.02 (2.68; 3.30)	0.67 (0.64; 0.70)	98.9% (98.6; 99.4)	0.1% (0.1; 0.2)	0.9% (0.6; 1.3)
Setting 3: maximum 5 errors, unequal drawing probabilities, missing values					
FH original	1.46 (1.40; 1.51)	0.47 (0.44; 0.50)	99.8% (99.6; 100.0)	0.0% (0.0; 0.0)	0.2% (0.0; 0.4)
FH extended	2.03 (1.90; 2.28)	0.52 (0.49; 0.56)	99.4% (99.1; 99.4)	0.0% (0.0; 0.1)	0.6% (0.6; 0.9)
FH extended (weighted)	2.61 (2.52; 2.70)	0.62 (0.58; 0.66)	99.2% (99.0; 99.4)	0.0% (0.0; 0.2)	0.8% (0.6; 1.0)
Setting 4: maximum 10 errors, unequal drawing probabilities					
FH original	10.29	1.02	91.4%	0.0%	8.6%
FH extended	12.92	1.26	87.1%	0.0%	12.9%
FH extended (weighted)	15.55	1.78	82.2%	0.0%	17.8%

It is seen that the extended paradigm led to an increase in computation time and, consequently, an increase in the number of records for which the error localisation problem was not solved to (guaranteed) optimality within 60 seconds. For Settings 1, 2 and 3, this increase was not large. For Setting 4, the average computation time per record was much longer, due to the larger number of errors per record. In practice, manual editing may be required for records that contain many errors. However, this is true for both the original Fellegi-Holt paradigm and its extension.

The next two tables contain the evaluation measures α, β, δ and ρ^c as defined in Subsection 5.2. The indicators that are based on counting edit operations are shown in Table 8; the indicators that are based on erroneous values are shown in Table 9.

Table 8. Results of error localisation on simulated data: evaluation measures with respect to edit operations

	evaluation measures			
	α	β	δ	ρ^c
Setting 1: maximum 5 errors				
FH original	23.6% (22.3; 25.2)	0.5% (0.5; 0.6)	1.2% (1.1; 1.3)	44.2% (42.3; 46.7)
FH extended	22.1% (20.7; 23.4)	0.4% (0.4; 0.4)	1.0% (0.9; 1.1)	40.2% (38.0; 42.5)
Setting 2: maximum 5 errors, unequal drawing probabilities				
FH original	38.7% (38.1; 39.4)	0.9% (0.5; 1.0)	2.0% (1.9; 2.0)	63.8% (62.7; 65.1)
FH extended	32.3% (31.1; 33.5)	0.6% (0.5; 0.6)	1.4% (1.4; 1.5)	55.7% (54.8; 57.5)
FH extended (weighted)	24.2% (23.4; 25.6)	0.3% (0.3; 0.3)	1.0% (0.9; 1.0)	43.6% (42.6; 44.8)
Setting 3: maximum 5 errors, unequal drawing probabilities, missing values				
FH original	37.3% (35.5; 39.4)	0.9% (0.8; 0.9)	1.9% (1.8; 1.9)	62.3% (60.1; 64.7)
FH extended	31.3% (29.9; 32.2)	0.5% (0.5; 0.5)	1.4% (1.3; 1.4)	54.5% (53.3; 56.6)
FH extended (weighted)	22.8% (22.4; 23.7)	0.3% (0.3; 0.3)	0.9% (0.9; 0.9)	41.7% (40.6; 43.1)
Setting 4: maximum 10 errors, unequal drawing probabilities				
FH original	51.2%	1.3%	3.6%	75.7%
FH extended	50.0%	0.7%	3.0%	69.5%
FH extended (weighted)	49.5%	0.4%	2.6%	60.4%

Table 9. Results of error localisation on simulated data: evaluation measures with respect to erroneous values

	evaluation measures			
	α	β	δ	ρ^c
Setting 1: maximum 5 errors				
FH original	27.0% (25.9; 28.2)	0.4% (0.4; 0.5)	1.4% (1.3; 1.5)	41.2% (39.4; 42.9)
FH extended	27.2% (26.0; 28.4)	0.4% (0.4; 0.4)	1.3% (1.2; 1.4)	40.2% (38.0; 42.4)
Setting 2: maximum 5 errors, unequal drawing probabilities				
FH original	41.3% (40.4; 42.2)	0.6% (0.6; 0.7)	2.2% (2.1; 2.3)	58.3% (56.3; 59.8)
FH extended	39.2% (37.8; 40.3)	0.5% (0.5; 0.6)	2.0% (2.0; 2.1)	55.6% (54.5; 57.5)
FH extended (weighted)	26.4% (25.6; 27.5)	0.3% (0.3; 0.3)	1.3% (1.3; 1.4)	43.2% (42.1; 44.7)
Setting 3: maximum 5 errors, unequal drawing probabilities, missing values				
FH original	40.5% (38.1; 41.9)	0.6% (0.6; 0.6)	2.1% (2.0; 2.2)	57.0% (54.5; 60.0)
FH extended	38.7% (36.6; 39.6)	0.5% (0.5; 0.5)	2.0% (1.8; 2.0)	54.4% (53.1; 56.6)
FH extended (weighted)	25.6% (25.1; 26.3)	0.3% (0.2; 0.3)	1.2% (1.2; 1.3)	41.5% (40.4; 42.8)
Setting 4: maximum 10 errors, unequal drawing probabilities				
FH original	52.7%	0.9%	4.2%	71.4%
FH extended	53.9%	0.6%	4.0%	69.5%
FH extended (weighted)	49.3%	0.3%	3.4%	59.9%

It is seen that for the settings considered here, the extended paradigm always leads to an improvement for the evaluation measures with respect to edit operations, and nearly always for the evaluation measures with respect to erroneous values. For Setting 1, where the special edit operations are relatively rarely needed, the improvement is not substantial. For the other settings, however, large improvements are seen when the optimal confidence weights based on (17) are used. These improvements are also observed in terms of the evaluation measures in Table 9 which look at erroneous values instead of edit operations.

Three further remarks should be made about Table 8 and 9. Firstly, the evaluation measures were computed on all 1080 records, including any records for which no solution was found. These records will impact negatively on the evaluation measures, in particular on indicators α and ρ^c. It is clear that this puts the extended paradigm at a disadvantage, in particular for Setting 4, where it led to a significant increase in the number of records that could not be solved (see Table 7). Nevertheless, the extended

paradigm performed well in terms of the indicators in Table 8 and 9 even for Setting 4, which suggests that the improvement compared to the original Fellegi-Holt paradigm was even larger on the subset of records for which an optimal solution could be found within 60 seconds.

Secondly, the datasets generated for Setting 3 contained missing values. These missing values were identified correctly as erroneous by all three error localisation methods. This should have a positive effect on the quality of error localisation compared to Setting 2. However, there could also be a negative effect, because a missing value can 'hide' other errors: if a missing value occurs in a balance edit rule, for instance, an error localisation algorithm based on the original Fellegi-Holt paradigm will be inclined to not identify other variables that occur in the same balance edit as erroneous, even when these variables have smaller confidence weights. The results in Table 8 and 9 for Settings 2 and 3 suggest that in this case the positive effect of having missing values slightly outweighed the negative effect.

Thirdly, the majority of variables in the datasets used in this study does not occur in any special edit operation. For these variables, any differences between the original and extended Fellegi-Holt paradigm can arise only as an indirect result of including special edit operations. Therefore, it seems plausible that the different error localisation methods will yield more similar results for these variables than for variables for which special edit operations have been defined. To check this, we have also computed the evaluation measures from Table 8 and 9 on the subset of variables that are involved in special edit operations. The results are shown in Table 10 (edit operations) and Table 11 (erroneous values) in Appendix A. As expected, the differences between the methods were more pronounced for this subset.

6. Consistent imputation

So far in this paper, we have focussed on the problem of error localisation. After the errors have been identified, new values have to be imputed for the variables that were deemed to be erroneous. As was noted in the introduction, it is important that the imputed values are consistent with the edit rules. In this section, we briefly review how consistent imputations can be obtained after error localisation under the original Fellegi-Holt paradigm, and we discuss how this approach can be extended to our generalised error localisation problem of Section 3.

In practice at Statistics Netherlands, consistent imputation is usually achieved in two steps. In the first step we impute values using a relatively simple method, such as regression imputation, ratio imputation or random hot deck imputation. In the second step these initial imputations are minimally adjusted, according to some criterion, to become consistent with the edit rules. In both steps, only the values of variables that were identified as erroneous (or missing) are allowed to change. See, e.g., De Waal *et al.* (2011, Chapter 10) for a general discussion of this topic.

As before, let $\boldsymbol{p} = (p_1, \dots, p_n)'$ denote an original observed record and assume that the edit rules are given by (1). Suppose that error localisation has been done according to the original Fellegi-Holt paradigm by solving problem (2), and suppose that the optimal solution is given by δ_i^{FH} $(i = 1, \dots, n)$. Let $\tilde{\boldsymbol{x}} = (\tilde{x}_1, \dots, \tilde{x}_n)'$ denote the initial imputed record after step 1, with $\tilde{x}_i = p_i$ for all variables with $\delta_i^{FH} = 0$. In step 2, the final imputed record is obtained by solving a problem of the form

$$\begin{aligned}
&\text{Minimise}_{\boldsymbol{x}} \quad D(\boldsymbol{x}, \tilde{\boldsymbol{x}}), \\
&\text{subject to} \\
&\mathbf{A}\boldsymbol{x} + \boldsymbol{b} \odot \mathbf{0}, \\
&p_i - C_i^{FH} \le x_i \le p_i + C_i^{FH} \quad (i = 1, \dots, n), \\
&C_i^{FH} = M\delta_i^{FH} \quad (i = 1, \dots, n), \\
&\boldsymbol{x} \in \mathbb{R}^n.
\end{aligned} \tag{18}$$

Here, $D(\boldsymbol{x}, \tilde{\boldsymbol{x}})$ is a distance function. Two common choices are: a weighted sum of absolute adjustments,

$$D(\boldsymbol{x}, \tilde{\boldsymbol{x}}) = \sum_{i=1}^{n} w_i |x_i - \tilde{x}_i|, \tag{19}$$

or a weighted sum of quadratic adjustments (squared weighted Euclidean distance),

$$D(\boldsymbol{x}, \tilde{\boldsymbol{x}}) = \sum_{i=1}^{n} w_i (x_i - \tilde{x}_i)^2. \tag{20}$$

Note that in problem (18), the binary values δ_i^{FH} are fixed and not part of the minimisation criterion. If these values have been obtained by solving error localisation problem (2), then problem (18) is guaranteed to have a feasible solution. Moreover, the fact that all decision variables are now real-valued means that (18) can be solved much more efficiently than (2). In particular, if $D(\boldsymbol{x}, \tilde{\boldsymbol{x}})$ is given by (19), the problem can be written as a standard linear programming problem (after some transformation to get rid of the absolute signs), and if $D(\boldsymbol{x}, \tilde{\boldsymbol{x}})$ is given by (20), it is a

standard quadratic programming problem. In R, the former can be solved using the lpSolveAPI package. For solving quadratic programming problems, the R package rspa is available (Van der Loo, 2015).

The sum of absolute adjustments (19) is used to obtain consistent imputations in the software SLICE which is part of the production process of SBS at Statistics Netherlands (De Waal, 2005). Here, the confidence weights of the original error localisation problem are re-used as weights in (19). The squared weighted Euclidean distance (20) is used in the production process of the Dutch statistics on health care, which is implemented in R using editrules and rspa (Van der Loo and Pannekoek, 2013). Here, the weights are chosen equal to $1/|\tilde{x}_i|$.

The above approach can be generalised easily to the extended error localisation problem of Section 3. For a given original record $\boldsymbol{p}$ we now have the error localisation indicators δ_i^{FH} and δ_k^{EO}. We can start by obtaining an initial imputed record $\tilde{\boldsymbol{x}}$, where now it must hold that $\tilde{x}_i = p_i$ for all variables with $\delta_i^{FH} + \sum_{k=1}^{K} I_{ki}^{EO} \delta_k^{EO} = 0$. The final imputed record can be obtained by solving the following minimisation problem, based on (7):

$$
\begin{aligned}
&\text{Minimise}_{\boldsymbol{x},\boldsymbol{\alpha}_k} \quad D(\boldsymbol{x}, \tilde{\boldsymbol{x}}),\\
&\text{subject to}\\
&\mathbf{A}\boldsymbol{x} + \boldsymbol{b} \odot \mathbf{0},\\
&p_i - C_i^{FH} + C_i^{EO} \le x_i \le p_i + C_i^{FH} + C_i^{EO} \quad (i = 1, \dots, n),\\
&C_i^{FH} = M\delta_i^{FH} \quad (i = 1, \dots, n),\\
&C_i^{EO} = \textstyle\sum_{k=1}^{K} I_{ki}^{EO}(C_{ki}^{D} + C_{ki}^{V}) \quad (i = 1, \dots, n),\\
&C_{ki}^{D} = \delta_k^{EO}\left(\textstyle\sum_{j=1}^{n} t_{kij} p_j + c_{ki} - p_i\right) \quad (k = 1, \dots, K;\ i = 1, \dots, n),\\
&C_{ki}^{V} = \textstyle\sum_{r=1}^{m_k} s_{kir}\alpha_{kr} \quad (k = 1, \dots, K;\ i = 1, \dots, n),\\
&-M\delta_k^{EO} \le \alpha_{kr} \le M\delta_k^{EO} \quad (k = 1, \dots, K;\ r = 1, \dots, m_k),\\
&\boldsymbol{x} \in \mathbb{R}^n, \boldsymbol{\alpha}_k \in \mathbb{R}^{m_k} \quad (k = 1, \dots, K).
\end{aligned}
\tag{21}
$$

Again, the values of the binary indicators δ_i^{FH} and δ_k^{EO} are now fixed, which makes the problem much less complex than (7): by choosing (19) or (20) as the distance function, we again obtain a standard linear or quadratic programming problem, respectively. Provided that the values of δ_i^{FH} and δ_k^{EO} have been obtained by solving problem (7), it is guaranteed that (21) has a feasible solution.

The above problem formulations in (18) and (21) are generic. It should be noted that, in practice, these problems can often be simplified greatly before solving. In (18), we can substitute $x_i = p_i$ for all variables with $\delta_i^{FH} = 0$ into the edit rules $\mathbf{A}\boldsymbol{x} + \boldsymbol{b} \odot \mathbf{0}$. The problem can then be reduced to a problem that only involves the variables x_i with $\delta_i^{FH} = 1$. Similarly, in problem (21), we can substitute $x_i = p_i$ for all variables with $\delta_i^{FH} + \sum_{k=1}^{K} I_{ki}^{EO} \delta_k^{EO} = 0$ into the edit rules. We can also remove any restrictions related to edit operations with $\delta_k^{EO} = 0$. Simplification is also possible for any variable that is affected only by a special edit operation without a variable part; i.e., all edit operations with $m_k = 0$ so that $C_{ki}^{V} = 0$. (For instance, the change-sign and interchange-values operations.) For these cases, we can substitute $x_i = p_i + C_i^{EO}$ into the edits, since C_i^{EO} is a fixed value.

A further simplification can be useful, based on the fact that some of the variable parts C_{ki}^{V} in problem (21) may be implicitly fixed by the other restrictions. In fact, in the original problem (18) it is also true that some of the x_i values with $\delta_i^{FH} = 1$ may be implicitly fixed by the other restrictions. This can be investigated by solving two sets of different minimisation problems, with the same restrictions as in (18) or (21) but with the criterion functions given by x_i and $-x_i$, respectively. Any variables for which both minimisation problems yield the same optimum are implicitly fixed by the restrictions to this common value. This is known as *deductive imputation*. For particular classes of edit rules, even simpler approaches are available to derive deductive imputations (De Waal *et al.*, 2011, Chapter 9). Note that deductive imputation can be done directly after the error localisation step, even before the initial imputations $\tilde{x}_i$ are created. This has the advantage that more auxiliary information is available to create the initial imputations for the remaining variables, which may improve the quality of these imputations and (indirectly) the quality of the final, adjusted imputations.

7. Conclusion

In this paper, we have introduced a MIP formulation for the generalised Fellegi-Holt paradigm for automatic error localisation that was proposed by Scholtus (2014, 2016). In comparison to the original proposal, we have introduced two additional assumptions (in Section 3):

1. In the optimal solution all special edit operations occur as if they are applied to the original observed record.
2. In the optimal solution Fellegi-Holt operations are always applied *after* special edit operations.

These assumptions remove the need to check for order-dependency between edit operations. This reduces the complexity of the generalised error localisation problem. In fact, for an application to a dataset of realistic dimensions, this reduction is necessary to obtain an error localisation problem that is solvable in practice. As was noted in Section 3, these assumptions are not very restrictive in practice and in fact may help to clarify the role of special edit operations in the generalised error localisation problem.

To give an example, in the application from Section 4 it was not possible to apply the special edit operations CV4 and TA4 from Table 1 to the same record, since these edit operations would both change the value of the variable INKWRDE120000. However, if this would be considered a meaningful combination of edit operations (e.g., because manual editors often apply both operations to the same record), then we could define a new edit operation accordingly. Depending on the context, we could define one edit operation which performs both operations in a particular order, or even two different edit operations for both orders.

We have made a prototype implementation in R of an error localisation algorithm based on the MIP formulation. The results of applying this algorithm to data with real (Section 4) and simulated errors (Section 5) show that solving the generalised error localisation problem is technically feasible for datasets that occur in practice, with on the order of 100 variables and 100 edit rules. The new MIP problem has a higher computational cost than the MIP formulation of the original Fellegi-Holt paradigm, but this increase is not dramatic. The results with simulated errors also show that a significant improvement is possible in the quality of error localisation with the new paradigm, compared to the original Fellegi-Holt paradigm. This does require that meaningful edit operations have been defined, i.e., edit operations that represent types of errors that occur relatively often in the data at hand. It also appears that specifying ‘optimal’ weights for these edit operations according to (17) leads to significantly better results. In practice, this would require prior information about the relative frequencies with which different types of errors occur.

The new paradigm introduces additional parameters that have to be specified: the choice of admissible edit operations and the associated confidence weights. The results in this paper suggest that some care must be taken to make appropriate

choices for these parameters. If the selected edit operations do not correspond to errors that are made often by respondents, or if the weights do not adequately reflect the relative frequencies with which these errors occur, then the new approach may not be worthwhile, as it then comes with an increase in complexity and computation time without necessarily leading to a better quality of error localisation. In practice it is therefore important, if the new approach is used, to find appropriate edit operations and associated weights. The required information could be obtained by analysing historical manually-edited data and documentation of the manual editing process, and from interviews with (supervisors of) editors. In the application to real data in Section 4, we have made a start with this analysis, but it is clear that setting up an actual application for statistical production would require more work. The main aim here was to test whether the approach is technically feasible.

We can conclude that the generalised Fellegi-Holt paradigm may be useful to improve the quality of automatic editing in practice. It therefore has the potential to make automatic editing applicable on a wider scale than it is currently used, thereby improving the efficiency of data editing processes. From a technical point of view, the new approach is feasible. The next step would be to test the approach on a real application, which would require experiments to find an appropriate choice of edit operations and weights.

References

Daalmans, J.A. (2018). Constraint Simplification for Data Editing of Numerical Variables. *Journal of Official Statistics* **34**, 27–39.

de Jonge, E. and M. van der Loo (2014). *Error Localization as a Mixed Integer Problem with the Editrules Package*. Discussion Paper 2014-07, Statistics Netherlands, The Hague. Available at: http://www.cbs.nl.

de Waal, T. (2005). *SLICE 1.5: A Software Framework for Automatic Edit and Imputation*. Working Paper No. 39, UN/ECE Work Session on Statistical Data Editing, Ottawa.

de Waal, T., J. Pannekoek, and S. Scholtus (2011). *Handbook of Statistical Data Editing and Imputation*. New York: John Wiley & Sons. DOI: 10.1002/9780470904848.

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

Granquist, L. and J. Kovar (1997). Editing of Survey Data: How Much is Enough? In: *Survey Measurement and Process Quality* (eds. L.E. Lyberg, P. Biemer, M. Collins, E.D. de Leeuw, C. Dippo, N. Schwartz, and D. Trewin), New York: John Wiley & Sons, pp. 415–435.

Konis, K. (2014). *lpSolveAPI: R interface for lp_solve version 5.5.2.0*. http://lpsolve.r-forge.r-project.org.

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**, 511–537. DOI: 10.2478/jos-2013-0038.

Scholtus, S. (2014). *Error Localisation using General Edit Operations*. Discussion Paper 2014-14, Statistics Netherlands, The Hague. Available at: http://www.cbs.nl.

Scholtus, S. (2016). A Generalized Fellegi-Holt Paradigm for Automatic Error Localization. *Survey Methodology* **42**, 1–18.

van der Loo, M. (2015). *The rspa package for minimal record adjustment*.

van der Loo, M. and E. de Jonge (2012). *Automatic Data Editing with Open Source R*. Working Paper No.33, UN/ECE Work Session on Statistical Data Editing, Oslo.

van der Loo, M. and J. Pannekoek (2013). *An Automatic Data Editing System for Healthcare Statistics*. Internal report (in Dutch), Statistics Netherlands, The Hague.

Appendix: Error localisation on simulated data - additional results

Table 10. Results of error localisation on simulated data: evaluation measures with respect to edit operations (subset: variables involved in special edit operations)

	evaluation measures			
	α	β	δ	ρ^c
Setting 1: maximum 5 errors				
FH original	53.8% (52.3; 56.7)	0.9% (0.8; 1.0)	3.1% (2.9; 3.4)	42.0% (40.3; 44.8)
FH extended	44.8% (42.3; 46.6)	0.4% (0.4; 0.5)	2.2% (2.1; 2.4)	37.3% (35.3; 39.0)
Setting 2: maximum 5 errors, unequal drawing probabilities				
FH original	72.9% (70.0; 74.3)	1.8% (1.7; 1.8)	5.7% (5.6; 5.8)	62.5% (61.2; 63.3)
FH extended	58.0% (56.0; 60.1)	0.7% (0.6; 0.8)	3.8% (3.7; 4.1)	54.1% (53.3; 55.6)
FH extended (weighted)	39.9% (38.7; 41.6)	0.6% (0.5; 0.6)	2.7% (2.6; 2.9)	41.3% (40.0; 42.9)
Setting 3: maximum 5 errors, unequal drawing probabilities, missing values				
FH original	72.0% (71.2; 73.1)	1.7% (1.7; 1.8)	5.5% (5.2; 5.6)	61.3% (59.2; 63.9)
FH extended	57.5% (56.8; 57.9)	0.6% (0.5; 0.7)	3.7% (3.5; 3.8)	53.1% (51.8; 55.0)
FH extended (weighted)	39.0% (38.0; 39.7)	0.5% (0.4; 0.5)	2.5% (2.5; 2.6)	39.9% (38.9; 41.2)
Setting 4: maximum 10 errors, unequal drawing probabilities				
FH original	78.1%	2.4%	8.9%	75.2%
FH extended	69.3%	1.0%	6.9%	68.7%
FH extended (weighted)	59.6%	0.8%	5.9%	58.9%

Table 11. Results of error localisation on simulated data: evaluation measures with respect to erroneous values (subset: variables involved in special edit operations)

	evaluation measures			
	α	β	δ	ρ^c
Setting 1: maximum 5 errors				
FH original	54.0% (52.8; 55.5)	0.5% (0.4; 0.6)	5.2% (5.0; 5.6)	38.8% (37.1; 40.7)
FH extended	51.0% (49.1; 53.1)	0.5% (0.5; 0.6)	5.0% (4.7; 5.3)	37.1% (35.0; 38.7)
Setting 2: maximum 5 errors, unequal drawing probabilities				
FH original	65.4% (63.0; 66.6)	0.5% (0.4; 0.6)	9.4% (9.1; 9.7)	56.7% (54.8; 57.6)
FH extended	60.9% (59.2; 62.4)	0.6% (0.5; 0.7)	8.9% (8.6; 9.2)	53.8% (53.1; 55.4)
FH extended (weighted)	38.3% (37.0; 39.6)	0.7% (0.6; 0.7)	5.8% (5.6; 6.0)	40.8% (39.5; 42.1)
Setting 3: maximum 5 errors, unequal drawing probabilities, missing values				
FH original	65.3% (63.6; 67.6)	0.5% (0.4; 0.6)	9.2% (8.4; 9.8)	55.8% (53.5; 59.0)
FH extended	61.0% (59.5; 61.9)	0.5% (0.4; 0.6)	8.7% (8.0; 9.0)	52.9% (51.4; 55.0)
FH extended (weighted)	38.1% (37.2; 38.8)	0.6% (0.5; 0.6)	5.6% (5.4; 5.8)	39.6% (38.4; 40.8)
Setting 4: maximum 10 errors, unequal drawing probabilities				
FH original	71.8%	0.8%	15.9%	70.8%
FH extended	69.8%	0.6%	15.3%	68.4%
FH extended (weighted)	55.9%	0.8%	12.5%	58.1%

Explanation of symbols

Empty cell	Figure not applicable
.	Figure is unknown, insufficiently reliable or confidential
*	Provisional figure
**	Revised provisional figure
2017–2018	2017 to 2018 inclusive
2017/2018	Average for 2017 to 2018 inclusive
2017/'18	Crop year, financial year, school year, etc., beginning in 2017 and ending in 2018
2015/'16–2017/'18	Crop year, financial year, etc., 2015/'16 to 2017/'18 inclusive

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

Colophon

Publisher
Statistics Netherlands
Henri Faasdreef 312, 2492 JP The Hague
www.cbs.nl

Prepress
Statistics Netherlands, CCN Creation and visualisation

Design
Edenspiekermann

Information
Telephone +31 88 570 70 70, fax +31 70 337 59 94
Via contactform: www.cbsl.nl/information