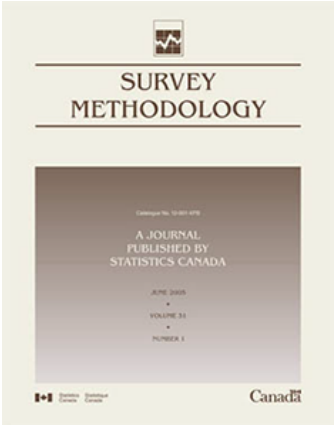
Catalogue no. 12-001-X
ISSN 1492-0921

Survey Methodology

A generalized Fellegi-Holt paradigm for automatic error localization



by Sander Scholtus

Release date: June 22, 2016

Statistics Canada Statistique Canada

Canada

How to obtain more information

For information about this product or the wide range of services and data available from Statistics Canada, visit our website, www.statcan.gc.ca.

You can also contact us by

email at STATCAN.infostats-infostats.STATCAN@canada.ca

telephone, from Monday to Friday, 8:30 a.m. to 4:30 p.m., at the following toll-free numbers:

- Statistical Information Service 1-800-263-1136
- National telecommunications device for the hearing impaired 1-800-363-7629
- Fax line 1-877-287-4369

Depository Services Program

- Inquiries line 1-800-635-7943
- Fax line 1-800-565-7757

Standards of service to the public

Statistics Canada is committed to serving its clients in a prompt, reliable and courteous manner. To this end, Statistics Canada has developed standards of service that its employees observe. To obtain a copy of these service standards, please contact Statistics Canada toll-free at 1-800-263-1136. The service standards are also published on www.statcan.gc.ca under "Contact us" > "Standards of service to the public."

Note of appreciation

Canada owes the success of its statistical system to a long-standing partnership between Statistics Canada, the citizens of Canada, its businesses, governments and other institutions. Accurate and timely statistical information could not be produced without their continued co-operation and goodwill.

Standard table symbols

The following symbols are used in Statistics Canada publications:

. not available for any reference period
.. not available for a specific reference period
... not applicable
0 true zero or a value rounded to zero
0^s value rounded to 0 (zero) where there is a meaningful distinction between true zero and the value that was rounded
^{p} preliminary
^{r} revised
x suppressed to meet the confidentiality requirements of the *Statistics Act*
^{E} use with caution
F too unreliable to be published
* significantly different from reference category ($p < 0.05$)

Published by authority of the Minister responsible for Statistics Canada

An HTML version is also available.

Cette publication est aussi disponible en français.

Survey Methodology, June 2016
Vol. 42, No. 1, pp. 1-18
Statistics Canada, Catalogue No. 12-001-X

A generalized Fellegi-Holt paradigm for automatic error localization

Sander Scholtus[1]

Abstract

The aim of automatic editing is to use a computer to detect and amend erroneous values in a data set, without human intervention. Most automatic editing methods that are currently used in official statistics are based on the seminal work of Fellegi and Holt (1976). Applications of this methodology in practice have shown systematic differences between data that are edited manually and automatically, because human editors may perform complex edit operations. In this paper, a generalization of the Fellegi-Holt paradigm is proposed that can incorporate a large class of edit operations in a natural way. In addition, an algorithm is outlined that solves the resulting generalized error localization problem. It is hoped that this generalization may be used to increase the suitability of automatic editing in practice, and hence to improve the efficiency of data editing processes. Some first results on synthetic data are promising in this respect.

Key Words: Automatic editing; Edit operations; Maximum likelihood; Numerical data; Linear edits.

1 Introduction

Data that have been collected for the production of statistics inevitably contain errors. A data editing process is needed to detect and amend these errors, at least in so far as they have an appreciable impact on the quality of statistical output (Granquist and Kovar 1997). Traditionally, data editing has been a manual task, ideally performed by professional editors with extensive subject-matter knowledge. To improve the efficiency, timeliness, and reproducibility of editing, many statistical institutes have attempted to automate parts of this process (Pannekoek, Scholtus and van der Loo 2013). This has resulted in deductive correction methods for *systematic errors* and error localization algorithms for *random errors* (de Waal, Pannekoek and Scholtus 2011, Chapter 1). In this article, I will focus on automatic editing for random errors.

Methods for this task usually proceed by minimally adjusting each record of data, according to some optimization criterion, so that it becomes consistent with a given set of constraints known as *edit rules*, or *edits* for short. Depending on the effectiveness of the optimization criterion and the strength of the edit rules, automatic editing may be used as a partial alternative to traditional manual editing. In practice, automatic editing is applied nearly always in combination with some form of *selective editing*, which means that the most influential errors are treated manually (Hidiroglou and Berthelot 1986; Granquist 1995, 1997; Granquist and Kovar 1997; Lawrence and McKenzie 2000; Hedlin 2003; de Waal et al. 2011).

Most automatic editing methods that are currently used in official statistics are based on the paradigm of Fellegi and Holt (1976): for each record, the smallest subset of variables is identified as erroneous that can be imputed so that the record becomes consistent with the edits. A slight generalization is obtained by assigning so-called *confidence weights* to the variables and minimizing the total weight of the imputed variables. Once this *error localization problem* is solved, suitable new values have to be found in a separate step for the variables that were identified as erroneous. This is the so-called *consistent imputation problem*; see de Waal et al. (2011) and their references. In this article, I will focus on the error localization problem.

1. Sander Scholtus, Statistics Netherlands, Department of Process Development and Methodology, P.O. Box 24500, 2490 HA, The Hague, The Netherlands. E-mail: sshs@cbs.nl.

At *Statistics Netherlands*, error localization based on the Fellegi-Holt paradigm has been a part of the data editing process for Structural Business Statistics (SBS) for over a decade now. In evaluation studies, where the same SBS data were edited both automatically and manually, a number of systematic differences were found between the two editing efforts. Many of these differences could be explained by the fact that human editors performed certain types of adjustments that were suboptimal under the Fellegi-Holt paradigm. For instance, editors sometimes interchanged the values of associated costs and revenues items, or transferred parts of reported amounts between variables.

In practice, 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; however, see EDIMBUS (2007, pages 34-35). 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).

To some extent, systematic differences between automatic and manual editing could be prevented by a clever choice of confidence weights. In general, however, the effects of a modification of the confidence weights on the results of automatic editing are difficult to predict. 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 correction of systematic errors through automatic correction rules (de Waal et al. 2011; Scholtus 2011). This approach has practical limitations, however, because it may require a large collection of if-then rules, which would be difficult to design and maintain over time (Chen, Thibaudeau and Winkler 2003). 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.

In this article, a different approach is suggested. A new definition of the error localization problem is proposed that allows for the possibility that errors affect more than one variable at a time. It is shown that this problem contains error localization under the original Fellegi-Holt paradigm as a special case. Throughout this article, I restrict attention to numerical data and linear edits; a possible extension to categorical and mixed data will be discussed briefly in Section 8.

The remainder of this article is organized as follows. Section 2 briefly reviews relevant previous work done in this area. In Section 3, the concept of an edit operation is introduced and illustrated. The new error localization problem is formulated in terms of these edit operations in Section 4. Section 5 generalizes an existing method for identifying solutions to the Fellegi-Holt-based error localization problem, and this result is used in Section 6 to outline a possible algorithm for solving the new problem. A small simulation study is discussed in Section 7. Finally, some conclusions and questions for further research follow in Section 8.

2 Background and related work

Let $\mathbf{x} = (x_1, \ldots, x_p)' \in \mathbb{R}^p$ 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:

$$\mathbf{A}\mathbf{x} + \mathbf{b} \odot \mathbf{0}, \tag{2.1}$$

where $\mathbf{A} = (a_{rj})$ is a $k \times p$ – matrix of coefficients and $\mathbf{b} = (b_1, \ldots, b_k)'$ is a vector of constants. Here and elsewhere, $\mathbf{0}$ represents a vector of zeros of appropriate length; similarly, $\odot$ represents a symbolic vector of operators from the set $\{\geq, \leq, =\}$.

For a given record $\mathbf{x}$ that does not satisfy all edits in (2.1), the Fellegi-Holt-based error localization problem amounts to finding the minimum of

$$\sum_{j=1}^{p} w_j \delta_j, \tag{2.2}$$

with $w_j > 0$ the confidence weight of variable x_j and $\delta_j \in \{0,1\}$, under the condition that the original record can be made consistent with the edits by imputing only those x_j with $\delta_j = 1$ (de Waal et al. 2011, page 66).

Fellegi and Holt (1976) also proposed a method for solving the above error localization problem, based on the generation of a sufficient set of so-called *implied edits* (see below). Unfortunately, the number of implied edits needed by this method is often extremely large in practice. Over the past decades, various dedicated algorithms for the error localization problem have been developed by, among others, Schaffer (1987), Garfinkel, Kunnathur and Liepins (1988), Kovar and Whitridge (1990), Ragsdale and McKeown (1996), de Waal (2003), de Waal and Quere (2003), Riera-Ledesma and Salazar-González (2003, 2007), Bruni (2004), and de Jonge and van der Loo (2014). Early algorithms mostly focused on strengthening the original method of Fellegi and Holt (1976) by reducing the number of required implied edits. More recent algorithms rely on the fact that the error localization problem can be written as a mixed-integer programming problem, which makes it possible to apply standard optimization techniques. See also de Waal and Coutinho (2005) or de Waal et al. (2011) for an overview and comparison of various error localization algorithms.

Implied edits are constraints that follow logically from the original edits (2.1). In the present context (numerical data, linear edits), all relevant implied edits may be generated by a technique called *Fourier-Motzkin elimination* (FM elimination; cf. Williams 1986). 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. For mathematical details, see the appendix.

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. In error localization under the Fellegi-Holt paradigm, by repeatedly applying this fundamental property, one may verify whether any particular combination of variables can be imputed to obtain a consistent record, given the original values of the other variables. A clear illustration of this use of FM elimination is provided by the error localization algorithm of de Waal and Quere (2003).

To conclude this section, it is interesting to look briefly at the statistical interpretation of the error localization problem. In fact, in motivating their paradigm for automatic error localization, 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, page 18).

A statistical argument for minimizing the weighted number of imputed variables was provided by Liepins (1980) and Liepins, Garfinkel and Kunnathur (1982), elaborating on earlier results of Naus, Johnson and Montalvo (1972). Suppose that errors occur according to a stochastic process, with each variable x_j being observed in error with a probability p_j that does not depend on its true value and with errors being independent across variables. Suppose furthermore that the confidence weights are defined as follows:

$$w_j = -\log\left(\frac{p_j}{1-p_j}\right). \tag{2.3}$$

Then it can be shown that minimizing expression (2.2) is approximately equivalent to maximizing the likelihood of the unobserved error-free record. Note that these authors tacitly assume that an error always affects one variable at a time.

Alternative error localization procedures that are based more directly on statistical models have been proposed by, e.g., Little and Smith (1987) and Ghosh-Dastidar and Schafer (2006). These procedures use outlier detection techniques and require an explicit model for the true data. Unfortunately, they cannot handle edit rules such as (2.1) in a straightforward manner.

3 Edit operations

Continuing with the notation from Section 2, I define an *edit operation* g to be an affine function of the general form

$$g(\mathbf{x}) = \mathbf{T}\mathbf{x} + \mathbf{S}\boldsymbol{\alpha} + \mathbf{c}, \tag{3.1}$$

where $\mathbf{T}$ and $\mathbf{S}$ are known coefficient matrices of dimensions $p \times p$ and $p \times m$, respectively, $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_m)'$ is a vector of free parameters that may occur in g, and $\mathbf{c}$ is a $p-$vector of known constants. In the special case that g does not involve any free parameters $(m = 0)$, the second term in (3.1) vanishes. Sometimes, it may be useful to impose one or several linear constraints on the free parameters in g:

$$\mathbf{R}\boldsymbol{\alpha} + \mathbf{d} \odot \mathbf{0}, \tag{3.2}$$

with $\mathbf{R}$ a known matrix, and $\mathbf{d}$ a known vector of constants. (Note: Matrix-vector notation will be used throughout this article because it leads to a concise description of results; however, using matrices to represent edits and edit operations is probably not the most efficient way to implement these results on a computer.)

As a first example, consider the operation that replaces one of the original values in $\mathbf{x}$ by an arbitrary new value (imputation). I will call this an *FH operation*, in view of its central role in automatic editing based on the Fellegi-Holt paradigm. Let $\mathbf{I}$ denote the $p \times p$ identity matrix and $\mathbf{e}_i$ the i^{th} standard basis vector in $\mathbb{R}^p$. The FH operation that imputes the variable x_j is given by (3.1) with $\mathbf{T} = \mathbf{I} - \mathbf{e}_j\mathbf{e}_j'$, $\mathbf{S} = \mathbf{e}_j$, and $\mathbf{c} = \mathbf{0}$. This yields: $g(\mathbf{x}) = \mathbf{x} + \mathbf{e}_j(\alpha - x_j) = (x_1, \ldots, x_{j-1}, \alpha, x_{j+1}, \ldots, x_p)'$, with $\alpha \in \mathbb{R}$ a free parameter that

represents the imputed value. It should be noted that for a record of p variables, p distinct FH operations can be defined.

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

- An edit operation that changes the sign of one of the variables:

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

- An edit operation that interchanges the values of two adjacent items:

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

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

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

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

- An edit operation that imputes two variables simultaneously using a fixed ratio:

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

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

Intuitively, an edit operation is supposed to "reverse the effects" of a particular type of error 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 $\mathbf{x}$, then $g(\mathbf{x})$ is the record that would have been observed if that error had not occurred. Somewhat more formally, it is assumed here that errors occurring in the data can be modeled by a stochastic "error generating process" $\mathcal{E}$, and that each edit operation acts as a "corrector" for one particular error that can occur under $\mathcal{E}$ (see Remark 4 in the next section).

If the edit operation g contains free parameters, the record $g(\mathbf{x})$ might not be determined uniquely even when the restrictions (2.1) and (3.2) are taken into account. In that case, one has to "impute" values for the free parameters that occur in an edit operation, which in turn means that some of the variables in $\mathbf{x}$

are imputed via the affine transformation given by (3.1). As in traditional Fellegi-Holt-based editing, finding appropriate "imputations" for the free parameters will not be considered part of the error localization problem here. On the other hand, if g does not contain any free parameters, the imputed values in $g(\mathbf{x})$ follow directly from the edit operation itself and the distinction between error localization and imputation is blurred.

In any particular application, only a small subset of potential edit operations of the form (3.1) would have a substantively meaningful interpretation, in the sense that the associated types of errors are known to occur. In what follows, I assume that a finite set of specific edit operations of the form (3.1) 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 8.

4 A generalized error localization problem

Let $\mathcal{G}$ be a finite set of allowed edit operations for a given application of automatic editing. Informally, I propose to generalize the error localization 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 generalized error localization problem, some new notation and concepts need to be introduced.

Consider a sequence of points $\mathbf{x} = \mathbf{x}_0, \mathbf{x}_1, \ldots, \mathbf{x}_t = \mathbf{y}$ in $\mathbb{R}^p$. A *path* from $\mathbf{x}$ to $\mathbf{y}$ is defined as a sequence of *distinct* edit operations $g_1, \ldots, g_t \in \mathcal{G}$ such that $\mathbf{x}_n = g_n(\mathbf{x}_{n-1})$ for all $n \in \{1, \ldots, t\}$. (Note: In the case that g_n contains free parameters, one should interpret this equality as "there exist feasible parameter values such that g_n maps $\mathbf{x}_{n-1}$ to $\mathbf{x}_n$".) A path is denoted by $P = [g_1, \ldots, g_t]$. The set of all possible paths from $\mathbf{x}$ to $\mathbf{y}$ is denoted by $\mathcal{P}(\mathbf{x}, \mathbf{y})$. This set may be empty. Later, I will use $\mathcal{P}(\mathbf{x}; G)$ to denote, for a given subset $G \subseteq \mathcal{G}$, the set of all paths starting in $\mathbf{x}$ that consist of the edit operations in G in some order (without specifying the free parameters); if G contains t elements, $\mathcal{P}(\mathbf{x}; G)$ contains $t!$ paths.

To each edit operation $g \in \mathcal{G}$, one can associate a weight $w_g > 0$ that expresses the costs of applying edit operation g. In particular, the weight of an FH operation is to be chosen equal to the confidence weight of the variable that it imputes. Now the *length* of a path $P = [g_1, \ldots, g_t]$ can be defined as the sum of the weights of its constituent edit operations: $\ell(P) = \sum_{n=1}^{t} w_{g_n}$, where, by convention, the empty path has length zero. The *distance* from $\mathbf{x}$ to $\mathbf{y}$ is defined as the length of the shortest path that connects $\mathbf{x}$ to $\mathbf{y}$:

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

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

The distance from $\mathbf{x}$ to any closed, non-empty subset $D \subseteq \mathbb{R}^p$ is defined as the distance to the nearest $\mathbf{y} \in D$: $d(\mathbf{x}, D) = \min\{d(\mathbf{x}, \mathbf{y}) | \mathbf{y} \in D\}$. For the purpose of error localization, the closed, non-empty subset of $\mathbb{R}^p$ that is of particular interest is the set D_0 of all points that satisfy (2.1).

I can now formulate the generalized error localization problem.

Problem. Consider a given set of consistent records D_0, a given set of allowed edit operations $\mathcal{G}$, and a given record $\mathbf{x}$. If $d(\mathbf{x}, D_0) = \infty$, then the error localization problem for $\mathbf{x}$ is infeasible. Otherwise, any shortest path leading to a record $\mathbf{y} \in D_0$ such that $d(\mathbf{x}, \mathbf{y}) < \infty$ is called a *feasible solution* to the error localization problem for $\mathbf{x}$. A feasible solution is called *optimal* if it leads to a record $\mathbf{x}^* \in D_0$ such that

$$d(\mathbf{x}, \mathbf{x}^*) = d(\mathbf{x}, D_0). \tag{4.1}$$

Formally, then, the generalized error localization problem consists of finding an optimal path of edit operations.

Remark 1. In general, there may be infinitely many records $\mathbf{x}^*$ in D_0 that satisfy (4.1) and can be reached by the same path of edit operations. To solve the error localization problem, it is sufficient to find an optimal path. Constructing an associated record $\mathbf{x}^* \in D_0$ may then be regarded as a generalization of the consistent imputation problem; cf. the discussion on imputation at the end of Section 3.

Remark 2. The above error localization problem is infeasible for records that cannot be mapped onto D_0 by any combination of distinct edit operations in $\mathcal{G}$. To avoid this situation, $\mathcal{G}$ should be chosen sufficiently large so that $d(\mathbf{x}, D_0) < \infty$ for all $\mathbf{x} \in \mathbb{R}^p$. In what follows, I tacitly assume that $\mathcal{G}$ has this property. An easy way – not necessarily the only way – to achieve this is by letting $\mathcal{G}$ contain at least all FH operations. That this is sufficient follows from the fact that any two points in $\mathbb{R}^p$ are connected by a path that concatenates the FH operations associated with the coordinates on which they differ.

Remark 3. It is not difficult to see that the above error localization problem reduces to the original problem of Fellegi and Holt (1976) in the special case that $\mathcal{G}$ contains only the FH operations.

Remark 4. As with the original Fellegi-Holt-based error localization problem, it can be shown that, under certain assumptions, minimizing $d(\mathbf{x}, \mathbf{y})$ over all $\mathbf{y} \in D_0$ for a given observed record $\mathbf{x}$ is approximately equivalent to maximizing the likelihood of the associated unobserved error-free record. The argument closely follows that of Kruskal (1983, pages 38-39) for the so-called Levenshtein distance in the context of approximate string matching. This requires first of all that the edits (2.1) be hard edits, i.e., failed only by erroneous values. In addition, it must be assumed that the stochastic "error generating process" $\mathcal{E}$ introduced in Section 3 has the following properties:

- There exists a one-to-one correspondence between the set of errors that can occur under $\mathcal{E}$ and the set of allowed edit operations $\mathcal{G}$ that correct them.
- The errors in $\mathcal{E}$ occur independently of each other.
- The error corresponding to operation g occurs with known probability p_g.

Finally, analogous to (2.3), the weights w_g should be chosen according to

$$w_g = -\log\left(\frac{p_g}{1 - p_g}\right). \tag{4.2}$$

Under these assumptions, Scholtus (2014) adapted the argument of Kruskal (1983) to show that the optimal solution to error localization problem (4.1) can be justified as an approximate maximum likelihood

estimator. [Note: The derivation in Scholtus (2014) assumed in addition that all $p_g \ll 1$, in which case $w_g \approx -\log p_g$. This assumption is unnecessary; cf. Liepins (1980).]

5 Implied edits for general edit operations

In this section, a result will be derived that establishes whether a given path of edit operations of the form (3.1) can be used to make a given record consistent with a given system of edit rules (i.e., is a feasible solution to the error localization problem). This result uses the FM elimination technique discussed in Section 2.

Let $\mathbf{x}$ be a given record and let $\mathbf{y}_t$ be any record that can be obtained by applying, in sequence, the edit operations $g_1, \ldots, g_t$ to $\mathbf{x}$:

$$\mathbf{y}_t = g_t \circ g_{t-1} \circ \cdots \circ g_1(\mathbf{x}). \tag{5.1}$$

Write $g_n(\mathbf{x}) = \mathbf{T}_n\mathbf{x} + \mathbf{S}_n\boldsymbol{\alpha}_n + \mathbf{c}_n$, for $n \in \{1, \ldots, t\}$. From (5.1) it follows by induction that

$$\begin{aligned} \mathbf{y}_1 &= \mathbf{T}_1\mathbf{x} + \mathbf{S}_1\boldsymbol{\alpha}_1 + \mathbf{c}_1, \\ \mathbf{y}_2 &= \mathbf{T}_2\mathbf{T}_1\mathbf{x} + \mathbf{S}_2\boldsymbol{\alpha}_2 + \mathbf{c}_2 + \mathbf{T}_2(\mathbf{S}_1\boldsymbol{\alpha}_1 + \mathbf{c}_1), \end{aligned}$$

and, in general,

$$\mathbf{y}_t = \mathbf{T}_t \cdots \mathbf{T}_1\mathbf{x} + \mathbf{S}_t\boldsymbol{\alpha}_t + \mathbf{c}_t + \sum_{n=2}^{t} \mathbf{T}_t \cdots \mathbf{T}_n(\mathbf{S}_{n-1}\boldsymbol{\alpha}_{n-1} + \mathbf{c}_{n-1}), \tag{5.2}$$

where the sum over n is defined to be zero when $t = 1$. Moreover, all terms involving $\mathbf{S}_n\boldsymbol{\alpha}_n$ vanish in these expressions when g_n does not contain any free parameters.

The path of edit operations $P = [g_1, \ldots, g_t]$ can be applied to $\mathbf{x}$ to obtain a record that is consistent with the edits (2.1) if, and only if, there exists a $\mathbf{y}_t$ of the form (5.2) that satisfies $\mathbf{A}\mathbf{y}_t + \mathbf{b} \odot \mathbf{0}$ and all relevant additional restrictions of the form (3.2) on $\boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_t$. Using (5.2), $\mathbf{A}\mathbf{y}_t + \mathbf{b} \odot \mathbf{0}$ can be written as:

$$(\mathbf{A}\mathbf{T}_t \cdots \mathbf{T}_1)\mathbf{x} + (\mathbf{A}\mathbf{S}_t)\boldsymbol{\alpha}_t + \sum_{n=2}^{t} (\mathbf{A}\mathbf{T}_t \cdots \mathbf{T}_n\mathbf{S}_{n-1})\boldsymbol{\alpha}_{n-1} + \mathbf{b}_t \odot \mathbf{0}, \tag{5.3}$$

with $\mathbf{b}_t = \mathbf{b} + \mathbf{A}\mathbf{c}_t + \sum_{n=2}^{t} \mathbf{A}\mathbf{T}_t \cdots \mathbf{T}_n\mathbf{c}_{n-1}$ a vector of constants.

Interestingly, (5.3) and the possible additional restrictions of the form (3.2) constitute a linear system of the form (2.1) on the extended record $(\mathbf{x}', \boldsymbol{\alpha}_1', \ldots, \boldsymbol{\alpha}_t')'$. Therefore, FM elimination may be used to remove all free parameters from this system. This yields a system of implied restrictions for $\mathbf{x}$. Moreover, a repeated application of the fundamental property of FM elimination establishes that $\mathbf{x}$ satisfies this system of implied edits if, and only if, there exist parameter values for $\boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_t$ that, together with $\mathbf{x}$, satisfy (5.3) and (3.2). Hence, it follows that the path of edit operations $P = [g_1, \ldots, g_t]$ can lead to a consistent record for $\mathbf{x}$ if, and only if, $\mathbf{x}$ satisfies the system of implied edits obtained by eliminating $\boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_t$ from (5.3) and (if relevant) additional restrictions of the form (3.2).

Example. Consider the following edits in x_1 and x_2:

$$x_1 \geq 0, \tag{5.4}$$

$$x_2 \geq 0, \tag{5.5}$$

$$x_1 + x_2 \leq 5. \tag{5.6}$$

Let g be the edit operation that transfers an amount of at most four units between x_1 and x_2, in either direction: $g((x_1, x_2)') = (x_1 + \alpha, x_2 - \alpha)'$ with $-4 \leq \alpha \leq 4$. For this single edit operation, the system of transformed edits (5.3) is:

$$x_1 + \alpha \geq 0, \tag{5.7}$$

$$x_2 - \alpha \geq 0, \tag{5.8}$$

$$x_1 + x_2 \leq 5. \tag{5.9}$$

I also add the following restrictions of the form (3.2) on α:

$$\alpha \geq -4, \tag{5.10}$$

$$\alpha \leq 4. \tag{5.11}$$

This yields five linear constraints (5.7)-(5.11) on x_1, x_2, and α, from which α may be removed by FM elimination to obtain:

$$x_1 \geq -4, \tag{5.12}$$

$$x_2 \geq -4, \tag{5.13}$$

$$x_1 + x_2 \geq 0, \tag{5.14}$$

$$x_1 + x_2 \leq 5. \tag{5.15}$$

According to the theory, any record $(x_1, x_2)'$ that satisfies (5.12)-(5.15) can be made consistent with the original edits (5.4)-(5.6) by transferring a certain amount $-4 \leq \alpha \leq 4$ between x_1 and x_2. The example record $(x_1, x_2)' = (-2, 3)'$ is inconsistent with the original edit rules (5.4)-(5.6) but satisfies (5.12)-(5.15). This implies that the record can be made consistent with the original edits by applying g. It is easy to see that this is true; any choice $2 \leq \alpha \leq 3$ will do.

It is interesting to note that, for the special case that P consists of the single FH operation that imputes x_j, the transformed system of edits (5.3) is obtained by replacing every occurrence of x_j in the original edits by an unrestricted parameter α. Eliminating α from (5.3) is equivalent in this case to eliminating x_j directly from the original edits. In this sense, the above result generalizes the fundamental property of FM elimination for FH operations to all edit operations of the form (3.1).

In general, the set of records defined by expression (5.2) depends on the way the edit operations are ordered. Thus, two paths consisting of the same set of edit operations in a different order need not yield the same solution to the error localization problem. In this respect, general edit operations differ from FH operations (Scholtus 2014).

6 An error localization algorithm

In this section, I propose a relatively simple algorithm to solve the error localization problem of Section 4, using the theoretical result from the previous section.

Step 0. Let $\mathbf{x}$ be a given record and $\mathcal{G}$ a given set of allowed edit operations. Initialize: $\mathcal{L} := \emptyset$; $\mathcal{B}_0 := \{\emptyset\}$; $W := \infty$; and $t := 1$.

Step 1. Determine all subsets $G \subseteq \mathcal{G}$ of cardinality t that satisfy these conditions:

1. Every subset of $t-1$ elements in G is part of $\mathcal{B}_{t-1}$.
2. It holds that $\sum_{g \in G} w_g \leq W$.

Step 2. For each G found in step 1, construct $\mathcal{P}(\mathbf{x};G)$ and, for each path $P \in \mathcal{P}(\mathbf{x};G)$, evaluate whether it can lead to a consistent record. If so, then:

- if $\ell(P) < W$, define $\mathcal{L} := \{P\}$ and $W := \ell(P)$;
- if $\ell(P) = W$, define $\mathcal{L} := \mathcal{L} \cup \{P\}$.

If *none* of the paths $P \in \mathcal{P}(\mathbf{x};G)$ lead to a consistent record, add G to $\mathcal{B}_t$.

Step 3. If $t < R$ and $\mathcal{B}_t \neq \emptyset$, define $t := t + 1$ and return to step 1.

Figure 6.1 An algorithm that finds all optimal paths of edit operations for problem (4.1).

In practical applications of error localization in official statistics, it is not unusual to have records of over 100 variables. To obtain a problem that is computationally feasible, existing applications of automatic editing based on the Fellegi-Holt paradigm usually specify an upper bound M on the number of variables that may be imputed in a single record (e.g., $M = 12$ or $M = 15$). de Waal and Coutinho (2005) argued that the introduction of such an upper bound is reasonable because a record that requires more than, say, fifteen imputations should be considered unfit for automatic editing anyway. Following this tradition, one can also introduce an upper bound R on the number of distinct edit operations that may be applied to a single record. Even with this additional restriction, the search space of potential solutions to (4.1) will usually be too large in practice to find the optimal solution by an exhaustive search.

Figure 6.1 summarizes the proposed error localization algorithm. Its basic set-up was inspired by the *apriori algorithm* of Agrawal and Srikant (1994) for data mining. Upon completion, the algorithm returns a set $\mathcal{L}$ containing all paths of allowed edit operations that correspond to an optimal solution to (4.1), as well as the optimal path length W. [Note: An error localization problem may have multiple optimal solutions, and it may be beneficial to find all of them (Giles 1988; de Waal et al. 2011, pages 66-67).]

After initialization in step 0, the algorithm cycles through steps 1, 2, and 3 at most R times. In step 1 of the algorithm, the search space is limited by using the following fact: if G has a proper subset $H \subset G$ for which $\mathcal{P}(\mathbf{x};H)$ contains a path that leads to a consistent record, then $\mathcal{P}(\mathbf{x};G)$ can contain only suboptimal solutions. Thus, any set G that has such a subset may be ignored by the algorithm. Similarly, G may also be ignored whenever the total weight of the edit operations in G exceeds the path length of the best feasible solution found so far.

During the t^{th} iteration, the number of subsets G encountered in step 1 of the algorithm equals $\binom{N}{t}$. For each of these subsets, the conditions in step 1 have to be checked. If a subset G passes these checks, in step 2 all $t!$ paths in $\mathcal{P}(\mathbf{x};G)$ are evaluated using the theory of Section 5. The idea behind the apriori

algorithm is that, as t becomes larger, the majority of subsets will not pass the checks in the first step, so that the total amount of computational work remains limited. In the context of data mining, this desirable behavior has indeed been observed in practice. Whether it also occurs in the context of error localization remains to be seen.

One possible improvement to the algorithm can be made by observing that the order in which edit operations are applied does not matter in all cases. Sometimes two paths in $\mathcal{P}(\mathbf{x};G)$ are *equivalent* in the sense that any record that can be reached from $\mathbf{x}$ by the first path can also be reached by the second path, and vice versa. This property defines an equivalence relation on $\mathcal{P}(\mathbf{x};G)$. Let $\tilde{\mathcal{P}}(\mathbf{x};G)$ be a set that contains one representative from each equivalence class of $\mathcal{P}(\mathbf{x};G)$ under this relation. Clearly, the algorithm in Figure 6.1 remains correct if in step 2 the search is limited to $\tilde{\mathcal{P}}(\mathbf{x};G)$ instead of $\mathcal{P}(\mathbf{x};G)$. Scholtus (2014) provides a simple method for constructing $\tilde{\mathcal{P}}(\mathbf{x};G)$ from $\mathcal{P}(\mathbf{x};G)$.

A detailed example illustrating the above algorithm can be found in Scholtus (2014).

7 Simulation study

To test the potential usefulness of the new error localization approach, I conducted a small simulation study, using the R environment for statistical computing (R Development Core Team 2015). A prototype implementation was created in R of the algorithm in Figure 6.1. This prototype made liberal use of the existing functionality for Fellegi-Holt-based automatic editing available in the `editrules` package (van der Loo and de Jonge 2012; de Jonge and van der Loo 2014). The program was not optimized for computational efficiency, but it turned out to work sufficiently fast for the relatively small error localization problems encountered in this simulation study. (Note: The R code used in this study is available from the author upon request.)

The simulation study involved records of five numerical variables that should satisfy the following nine linear edit rules:

$$\begin{aligned} x_1 + x_2 &= x_3, \\ x_3 - x_4 &= x_5, \\ x_j &\geq 0, \qquad j \in \{1,2,3,4\}, \\ x_1 &\geq x_2, \\ x_5 &\geq -0.1x_3, \\ x_5 &\leq 0.5x_3. \end{aligned}$$

Edits of this form might typically be encountered for SBS, as part of a much larger set of edit rules (Scholtus 2014).

I created a random error-free data set of 2,000 records by drawing from a multivariate normal distribution (using the `mvtnorm` package) with the following parameters:

$$\boldsymbol{\mu} = \begin{pmatrix} 500 \\ 250 \\ 750 \\ 600 \\ 150 \end{pmatrix} \quad \text{and} \quad \boldsymbol{\Sigma} = \begin{pmatrix} 10{,}000 & -1{,}250 & 8{,}750 & 7{,}500 & 1{,}250 \\ -1{,}250 & 5{,}000 & 3{,}750 & 4{,}000 & -250 \\ 8{,}750 & 3{,}750 & 12{,}500 & 11{,}500 & 1{,}000 \\ 7{,}500 & 4{,}000 & 11{,}500 & 11{,}750 & -250 \\ 1{,}250 & -250 & 1{,}000 & -250 & 1{,}250 \end{pmatrix}.$$

Only records that satisfied all of the above edits were added to the data set. Note that $\boldsymbol{\Sigma}$ is a singular covariance matrix that incorporates the two equality edits. Technically, the resulting data follow a so-called truncated multivariate singular normal distribution; see de Waal et al. (2011, pages 318ff) or Tempelman (2007).

Table 7.1 lists the nine allowed edit operations that were considered in this study. Note that the first five lines contain the FH operations for this data set. As indicated in the table, each edit operation has an associated type of error. A synthetic data set to be edited was created by randomly adding errors of these types to the above-mentioned error-free data set. The probability of each type of error is listed in the fourth column of Table 7.1. The associated "ideal" weight according to (4.2) is shown in the last column.

To limit the amount of computational work, I only considered records that required three edit operations or less. Records without errors were also removed. This left 1,025 records to be edited, each containing one, two, or three of the errors listed in Table 7.1.

Table 7.1
Allowed edit operations for the simulation study

name	operation	associated type of error	p_g	w_g
FH1	impute x_1	erroneous value of x_1	0.10	2.20
FH2	impute x_2	erroneous value of x_2	0.08	2.44
FH3	impute x_3	erroneous value of x_3	0.06	2.75
FH4	impute x_4	erroneous value of x_4	0.04	3.18
FH5	impute x_5	erroneous value of x_5	0.02	3.89
IC34	interchange x_3 and x_4	true values of x_3 and x_4 interchanged	0.07	2.59
TF21	transfer an amount from x_2 to x_1	part of the true value of x_1 reported as part of x_2	0.09	2.31
CS4	change the sign of x_4	sign error in x_4	0.11	2.09
CS5	change the sign of x_5	sign error in x_5	0.13	1.90

Several error localization approaches were applied to this data set. First of all, I tested error localization according to the Fellegi-Holt paradigm (i.e., using only the edit operations FH1–FH5) and according to the new paradigm (i.e., using all edit operations in Table 7.1). Both approaches were tested once using the "ideal" weights listed in Table 7.1 and once with all weights equal to 1 ("no weights"). The latter case simulates a situation where the relevant edit operations would be known, but not their respective frequencies. Finally, to test the robustness of the new error localization approach to a lack of information about relevant edit operations, I also applied this approach with one of the non-FH operations in Table 7.1 missing from the set of allowed edit operations.

The quality of error localization was evaluated in two ways. Firstly, I evaluated how well the optimal paths of edit operations found by the algorithm matched the true distribution of errors, using the following contingency table for all $1{,}025 \times 9 = 9{,}225$ combinations of records and edit operations:

Table 7.2
Contingency table of errors and edit operations suggested by the algorithm

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

From this table, I computed indicators that measure the proportion of false negatives, false positives, and overall wrong decisions, respectively:

$$\alpha = \frac{FN}{TP + FN}; \quad \beta = \frac{FP}{FP + TN}; \quad \delta = \frac{FN + FP}{TP + FN + FP + TN}.$$

Similar indicators are discussed by de Waal et al. (2011, pages 410-411). I also computed $\bar{\rho} = 1 - \rho$, with ρ the fraction of records in the data set for which the error localization algorithm found exactly the right solution. A good error localization algorithm should have low scores on all four indicators.

It should be noted that the above quality indicators put the original Fellegi-Holt approach at a disadvantage, as this approach does not use all the edit operations listed in Table 7.1. Therefore, I also calculated a second set of quality indicators $\alpha, \beta, \delta,$ and $\bar{\rho}$ that look at erroneous values rather than edit operations. In this case, α measures the proportion of values in the data set that were affected by errors but left unchanged by the optimal solution of the error localization problem, and similarly for the other measures.

Table 7.3 displays the results of the simulation study for both sets of quality indicators. In both cases, a considerable improvement in the quality of the error localization results is seen for the approach that used all edit operations, compared to the approach that used only FH operations. In addition, leaving one relevant edit operation out of the set of allowed edit operations had a negative effect on the quality of error localization. In some cases this effect was quite large – particularly in terms of edit operations used –, but the results of the new error localization approach still remained substantially better than those of the Fellegi-Holt approach. Contrary to expectation, not using different confidence weights actually improved the quality of the error localization results somewhat for this data set under the Fellegi-Holt approach (both sets of indicators) and to some extent also under the new approach (only the second set of indicators). Finally, it is seen that using all edit operations led to an increase in computing time compared to using only FH operations, but this increase was not dramatic.

Table 7.3
Quality of error localization in terms of edit operations used and identified erroneous values; computing time required

	quality indicators (edit operations)				quality indicators (erroneous values)				
approach	α	β	δ	$\bar{\rho}$	α	β	δ	$\bar{\rho}$	**time***
Fellegi-Holt (weights)	74%	12%	23%	80%	19%	10%	13%	32%	46
Fellegi-Holt (no weights)	70%	12%	21%	74%	13%	8%	9%	24%	33
all operations (weights)	14%	3%	5%	24%	10%	5%	7%	17%	98
except IC34	29%	5%	9%	35%	15%	9%	11%	29%	113
except TF21	34%	5%	10%	37%	10%	5%	7%	18%	80
except CS4	28%	6%	9%	39%	10%	5%	7%	17%	80
except CS5	35%	7%	10%	47%	11%	6%	7%	18%	82
all operations (no weights)	27%	5%	8%	36%	6%	4%	5%	13%	99

* Total computing time (in seconds) on a laptop PC with a 2.5 GHz CPU under Windows 7.

8 Conclusion

In this article, a new formulation was proposed of the error localization 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 localization problem can be seen as a generalization of the problem proposed in a seminal paper 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 main focus here has been on developing the mathematical theory behind the new error localization problem. It turns out that FM elimination, a technique that has been used in the past to solve the Fellegi-Holt-based error localization problem, can be applied also in the context of the new problem (Section 5). Nevertheless, the task of solving the new error localization 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 at statistical institutes. A possible error localization algorithm was outlined in Section 6. More efficient algorithms probably could and should be developed. Similarly to FM elimination, it may be possible to adapt other ideas that have been used to solve the Fellegi-Holt-based problem to the generalized problem considered here.

The discussion in this article was restricted to numerical data and linear edits. The original Fellegi-Holt paradigm has been applied also to categorical and mixed data. Several authors, including Bruni (2004) and de Jonge and van der Loo (2014), have shown that a large class of edits for mixed data can be re-formulated in terms of numerical data and linear edits, with the additional restriction that some of the variables have to be integer-valued. In principle, this means that the results in this article could be applied also to mixed data. To accommodate the fact that some variables are integer-valued, Pugh's (1992) extension of FM elimination to integers could be used; see also de Waal et al. (2011) for a discussion of this extended elimination technique in the context of Fellegi-Holt-based error localization. It remains to be seen whether this approach is computationally feasible.

Remark 4 in Section 4 hinted at an analogy between error localization in statistical microdata and the field of approximate string matching. In approximate string matching, text strings are compared under the assumption that they may have been partially corrupted (Navarro 2001). Various distance functions have been proposed for this task. The Hamming distance, which counts the number of positions on which two strings differ, may be seen as an analogue of the Fellegi-Holt-based target function (2.2). The generalized error localization problem defined in this paper has its counterpart in the use of the Levenshtein distance or "edit distance" for approximate string matching. It may be interesting to explore this analogy further. In particular, efficient algorithms have been developed for computing edit distances between strings; it might be possible to apply some of the underlying ideas also to the generalized error localization problem.

The new error localization algorithm was applied successfully to a small synthetic data set (Section 7). Overall, the results of this simulation study suggest that the new error localization approach has the potential to achieve a substantial improvement of the quality of automatic editing compared to the approach that is currently used in practice. However, this does require that sufficient information be available to identify all – or at least most – of the relevant edit operations in a particular application. Possible gains in the quality of error localization also have to be weighed in practice against the higher computational demands of the generalized error localization problem.

An obvious candidate for applying the new methodology in practice would be the SBS. However, more research is needed before this method could be applied during regular production. 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, editing instructions and other documentation used by the editors, and interviews with editors and/or supervisors of editing.

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

Ultimately, the aim of the new methodology proposed in this article is to improve the usefulness of automatic editing in practice. So far, the results are promising.

Acknowledgements

The views expressed in this article are those of the author and do not necessarily reflect the policies of *Statistics Netherlands*. The author would like to thank Jeroen Pannekoek, Ton de Waal, and Mark van der Loo for their comments on earlier versions of this article, as well as the Associate Editor and two anonymous referees.

Appendix

Fourier-Motzkin elimination

Consider a system of linear constraints (2.1) and let x_f be the variable to be eliminated. First, suppose that x_f is involved only in inequalities. For ease of exposition, suppose that the edits are normalized so that all inequalities use the $\geq$ operator. The FM elimination method considers all pairs (r,s) of inequalities in which the coefficients of x_f have opposite signs; that is, $a_{rf}a_{sf} < 0$. Suppose without loss of generality that $a_{rf} < 0$ and $a_{sf} > 0$. From the original pair of edits, the following implied constraint is derived:

$$\sum_{j=1}^{p} a_j^* x_j + b^* \geq 0, \tag{A.1}$$

with $a_j^* = a_{sf}a_{rj} - a_{rf}a_{sj}$ and $b^* = a_{sf}b_r - a_{rf}b_s$. Note that $a_f^* = 0$, so x_f is not involved in (A.1). An inequality of the form (A.1) is derived from each of the above-mentioned pairs (r,s). The full implied system of constraints obtained by FM elimination now consists of these derived constraints, together with all original constraints that do not involve x_f.

If there are linear equalities that involve x_f, the above technique could be applied after replacing each linear equality with two equivalent linear inequalities. de Waal and Quere (2003) suggested a more efficient

alternative for this case. Suppose that the r^{th} constraint in (2.1) is an equality that involves x_f. This constraint can be rewritten as

$$x_f = \frac{-1}{a_{rf}}\left(b_r + \sum_{j \neq f} a_{rj} x_j\right). \tag{A.2}$$

By substituting the expression on the right-hand-side of (A.2) for x_f in all other constraints, one again obtains an implied system of constraints that does not involve x_f and that can be rewritten in the form (2.1).

For a proof that FM elimination has the fundamental property mentioned in Section 2, see, e.g., de Waal et al. (2011, pages 69-70).

References

Agrawal, R., and Srikant, R. (1994). *Fast Algorithms for Mining Association Rules*. Technical report, IBM Almaden Research Center, San Jose, California.

Bruni, R. (2004). Discrete models for data imputation. *Discrete Applied Mathematics*, 144, 59-69.

Chen, B., Thibaudeau, Y. and Winkler, W.E. (2003). *A Comparison Study of ACS If-Then-Else, NIM, DISCRETE Edit and Imputation Systems Using ACS Data*. Working Paper No. 7, UN/ECE Work Session on Statistical Data Editing, Madrid.

de Jonge, E., and van der Loo, M. (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. (2003). Solving the error localization problem by means of vertex generation. *Survey Methodology*, 29, 1, 71-79.

de Waal, T., and Coutinho, W. (2005). Automatic editing for business surveys: An assessment for selected algorithms. *International Statistical Review*, 73, 73-102.

de Waal, T., and Quere, R. (2003). A fast and simple algorithm for automatic editing of mixed data. *Journal of Official Statistics*, 19, 383-402.

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

EDIMBUS (2007). *Recommended Practices for Editing and Imputation in Cross-Sectional Business Surveys*. Eurostat manual prepared by ISTAT, Statistics Netherlands, and SFSO.

Fellegi, I.P., and Holt, D. (1976). A systematic approach to automatic edit and imputation. *Journal of the American Statistical Association*, 71, 17-35.

Garfinkel, R.S., Kunnathur, A.S. and Liepins, G.E. (1988). Error localization for erroneous data: Continuous data, linear constraints. *SIAM Journal on Scientific and Statistical Computing*, 9, 922-931.

Ghosh-Dastidar, B., and Schafer, J.L. (2006). Outlier detection and editing procedures for continuous multivariate data. *Journal of Official Statistics*, 22, 487-506.

Giles, P. (1988). A model for generalized edit and imputation of survey data. *The Canadian Journal of Statistics*, 16, 57-73.

Granquist, L. (1995). Improving the traditional editing process. In *Business Survey Methods*, (Eds., B.G. Cox, D.A. Binder, B.N. Chinnappa, A. Christianson, M.J. Colledge and P.S. Kott), John Wiley & Sons, Inc., 385-401.

Granquist, L. (1997). The new view on editing. *International Statistical Review*, 65, 381-387.

Granquist, L., and Kovar, J. (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), John Wiley & Sons, Inc., 415-435.

Hedlin, D. (2003). Score functions to reduce business survey editing at the U.K. Office for National Statistics. *Journal of Official Statistics*, 19, 177-199.

Hidiroglou, M.A., and Berthelot, J.-M. (1986). Statistical editing and imputation for periodic business surveys. *Survey Methodology*, 12, 1, 73-83.

Kovar, J., and Whitridge, P. (1990). Generalized edit and imputation system; Overview and applications. *Revista Brasileira de Estadistica*, 51, 85-100.

Kruskal, J.B. (1983). An overview of sequence comparison. In *Time Warps, String Edits, and Macromolecules: The Theory and Practice of Sequence Comparison*, (Eds., D. Sankoff and J.B. Kruskal), Addison-Wesley, 1-44.

Lawrence, D., and McKenzie, R. (2000). The general application of significance editing. *Journal of Official Statistics*, 16, 243-253.

Liepins, G.E. (1980). *A Rigorous, Systematic Approach to Automatic Data Editing and its Statistical Basis*. Report ORNL/TM-7126, Oak Ridge National Laboratory.

Liepins, G.E., Garfinkel, R.S. and Kunnathur, A.S. (1982). Error localization for erroneous data: A survey. *TIMS/Studies in the Management Sciences*, 19, 205-219.

Little, R.J.A., and Smith, P.J. (1987). Editing and imputation of quantitative survey data. *Journal of the American Statistical Association*, 82, 58-68.

Naus, J.I., Johnson, T.G. and Montalvo, R. (1972). A probabilistic model for identifying errors in data editing. *Journal of the American Statistical Association*, 67, 943-950.

Navarro, G. (2001). A guided tour to approximate string matching. *ACM Computing Surveys*, 33, 31-88.

Pannekoek, J., Scholtus, S. and van der Loo, M. (2013). Automated and manual data editing: A view on process design and methodology. *Journal of Official Statistics*, 29, 511-537.

Pugh, W. (1992). The omega test: A fast and practical integer programming algorithm for data dependence analysis. *Communications of the ACM*, 35, 102-114.

R Development Core Team (2015). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. URL: http://www.R-project.org/.

Ragsdale, C.T., and McKeown, P.G. (1996). On solving the continuous data editing problem. *Computers & Operations Research*, 23, 263-273.

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

Riera-Ledesma, J., and Salazar-González, J.J. (2007). A branch-and-cut algorithm for the continuous error localization problem in data cleaning. *Computers & Operations Research*, 34, 2790-2804.

Schaffer, J. (1987). Procedure for solving the data-editing problem with both continuous and discrete data types. *Naval Research Logistics*, 34, 879-890.

Scholtus, S. (2011). Algorithms for correcting sign errors and rounding errors in business survey data. *Journal of Official Statistics*, 27, 467-490.

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

Tempelman, D.C.G. (2007). *Imputation of Restricted Data*. Ph. D. Thesis, University of Groningen. Available at: http://www.cbs.nl.

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

Williams, H.P. (1986). Fourier's method of linear programming and its dual. *The American Mathematical Monthly*, 93, 681-695.