

**UNITED NATIONS STATISTICAL COMMISSION and  
ECONOMIC COMMISSION FOR EUROPE**

**CONFERENCE OF EUROPEAN STATISTICIANS**

**Work Session on Statistical Data Editing**  
(Vienna, Austria, 21-23 April 2008)

Topic (iv): New and emerging methods

**ALGORITHMS FOR DETECTING AND RESOLVING OBVIOUS INCONSISTENCIES IN  
BUSINESS SURVEY DATA**

Prepared by Sander Scholtus, Statistics Netherlands

**I. INTRODUCTION**

1. It is well known that data collected in a survey or register contain errors. In the case of a survey, these errors may be introduced when the respondent answers the survey questions or during the processing of survey forms at the statistical office. Publication figures based on erroneous data may be biased or logically inconsistent. For example, if for some respondents the reported operating results disagree with the difference between the reported operating returns and operating costs, then a similar disagreement may be present in the published totals, unless the inconsistencies are detected and dealt with somehow. Often the survey data are put through an editing process to resolve the errors. In the case of structural business statistics, all survey variables are quantitative and many (linear) relationships between them can typically be formulated. Thus a set of linear equality and inequality restrictions is established. These restrictions are called *edit rules*. If for a particular record any of these edit rules are violated, the record is found to be inconsistent and it is deduced that some variable(s) must have an erroneous value.

2. A distinction is often made between *systematic errors* and *random errors*. Errors of the former type are reported consistently over time by respondents (EDIMBUS, 2007). They usually occur because a respondent misunderstands a survey question, e.g. by reporting financial amounts in Euros rather than the requested multiples of 1,000 Euros. A fault in the data processing software might also introduce a systematic error. Since it is reported consistently between respondents, an undiscovered systematic error leads to biased publication figures. Once detected, a systematic error can be corrected (semi-)deductively because the underlying error mechanism is known. Random errors on the other hand occur by accident, usually due to inattention of the respondent or the data processing staff. An example of a random error occurs when a '1' on a survey form is read as a '7' during data processing. Because of their random nature, no deductive method can typically be applied to correct errors of this type.

3. At Statistics Netherlands *selective editing* is used for the data of structural business statistics (De Jong, 2002). This means that only records containing (potentially) influential errors are edited manually by subject-matter specialists, whereas the remaining records are edited automatically. For the latter step the software package SLICE was developed at Statistics Netherlands. SLICE uses an advanced error localisation algorithm based on the Fellegi-Holt paradigm (Fellegi and Holt, 1976), which states that the smallest possible number of variables should be labelled erroneous such that the record can be made consistent with every edit rule. This paradigm is based on the assumption that the data contain only random errors. A description of the error localisation algorithm implemented in SLICE can be found in De Waal and Quere (2003) and De Waal (2003).

4. For each record a plausibility indicator is calculated to determine whether it may contain influential errors and should be edited manually (Hoogland, 2006). The value of this indicator may be

any natural number between 0 and 10. A record receives a low score if it contains values that differ substantially from reference values. The plausibility indicator is calibrated such that all records that receive a score of 6 or higher are deemed suitable for automatic editing. Only the records that do not receive a sufficiently high plausibility indicator are edited manually.<sup>1</sup>

5. Selective editing leads to a more efficient editing process than traditional (completely manual) editing, since part of the data stream is not reviewed by subject-matter specialists anymore. The automatic error localisation algorithm implemented in SLICE has its limitations, however, as it is not suitable for records containing either influential or systematic errors. Moreover, in practice the error localisation problem becomes too complicated if a record violates too large a number of edit rules. To preserve the quality of the editing process, only records that contain a limited number of non-influential random errors should be edited automatically. Ideally, the plausibility indicator filters out all records containing influential errors or too many inconsistencies. Prior to this, several types of *obvious* errors can be detected and resolved automatically in a separate step. A systematic error is called obvious if it can be detected ‘easily’, i.e. by applying some basic, specific search algorithm. An example of an obvious error occurs when a respondent reports the values of some items in a sum, but leaves the corresponding total blank. This error can be resolved deductively by calculating and filling in the total value.

6. It is useful to detect and correct obvious inconsistencies as early as possible in the editing process, since it is a waste of resources if subject-matter specialists or advanced editing software such as SLICE have to deal with them. When obvious inconsistencies are corrected in a separate step, before the plausibility indicator is calculated, the efficiency of the selective editing process increases because more records will be eligible for automatic editing. Solving the error localisation problem becomes easier once obvious inconsistencies have been removed, since the number of violated edit rules becomes smaller. Moreover, since obvious inconsistencies are systematic errors, they can be corrected more accurately by a specific, deductive algorithm than by a general error localisation algorithm based on the Fellegi-Holt paradigm.

7. Currently, the editing process of the structural business statistics at Statistics Netherlands contains a step in which three obvious inconsistencies are detected and corrected (section II). An investigation of data collected in the structural business statistics of 2001 has revealed some additional obvious inconsistencies (Scholtus, 2008). One of these is discussed in some detail in section III, along with an algorithm for detection and correction. Subsequently rounding errors are discussed in section IV. Although rounding errors are not obvious inconsistencies in the true sense of the word<sup>2</sup>, the efficiency of the editing process will increase if these errors are also detected and resolved during the correction step for obvious inconsistencies. For this reason a method to resolve rounding errors will be described.

## II. CURRENT APPROACH AT STATISTICS NETHERLANDS

8. The currently implemented editing process for the structural business statistics at Statistics Netherlands contains a step in which three obvious systematic errors are detected. These errors are treated deductively before any other correction is made in the data of the processed survey forms.

9. The first of these obvious inconsistencies was already mentioned in section I: the amounts on the survey form are sometimes reported in Euros instead of multiples of 1,000 Euros. It is important to detect this error because otherwise publication figures of all financial items will be overestimated. This error can be detected in two ways, depending on which auxiliary information is available: for respondents that are present in the VAT-register, the amount of turnover in the register is compared to the reported turnover in the survey, for the other respondents the amount of reported turnover per reported number of employees (in FTE) is compared to its median in last year’s edited data. If a large discrepancy is found by either method, all reported financial amounts of the respondent are divided by 1,000.

---

<sup>1</sup> In addition to this the largest companies are always edited manually, since they are likely to have a substantial influence on publication figures.

<sup>2</sup> In fact a rounding error is not systematic, but random.

10. The second obvious inconsistency occurs when a respondent adds a minus sign to a value that is subtracted. The survey form contains a number of combinations of items where expenditures have to be subtracted from returns to obtain a balance (see also section III). If a respondent adds a minus sign to the reported costs, the value becomes wrongfully negative after data processing. The resulting inconsistency can be detected and corrected easily: the reported amount is simply replaced by its absolute value.

11. The third and final obvious inconsistency was also mentioned in section I: some respondents report component items of a sum but leave the corresponding total blank. When this is detected, the total value is calculated and filled in.

### III. SIGN ERRORS AND INTERCHANGED RETURNS AND COSTS

#### A. Context

12. The *results block* is a part of the structural business statistics where the respondent has to fill in a number of balance amounts. We refer to these balance variables by  $X_0, X_1, \dots, X_{n-1}$ . A final balance variable  $X_n$  called the *pre-tax results* is found by adding up the other balance variables. That is, the data should conform to the following edit rule:

$$X_0 + X_1 + \dots + X_{n-1} = X_n. \quad (1)$$

Rule (1) is sometimes referred to as the *external sum*. A balance variable is defined as the difference between a returns item and a costs item. If these items are also asked on the survey form, the following edit rule should hold:

$$X_{k,r} - X_{k,c} = X_k, \quad (2)$$

where  $X_{k,r}$  denotes the returns item and  $X_{k,c}$  the costs item. Rules of this form are referred to as *internal sums*. We will use a general description of the results block, in which the returns and costs are not necessarily asked for every balance variable in the survey.<sup>3</sup> To keep the notation simple we assume that the balance variables are arranged such that exactly the first  $m+1$  are split into returns and costs, for some  $m \in \{0, 1, \dots, n-1\}$ . Thus, the following set of edit rules is used:

$$\begin{cases} X_0 = X_{0,r} - X_{0,c} \\ \vdots \\ X_m = X_{m,r} - X_{m,c} \\ X_n = X_0 + X_1 + \dots + X_{n-1} \end{cases} \quad (3)$$

In this notation the 0<sup>th</sup> balance variable  $X_0$  stands for the operating results, and  $X_{0,r}$  and  $X_{0,c}$  represent operating returns and operating costs, respectively.

13. Table 1 displays the structure of the results block from the structural business statistics questionnaire that was used at Statistics Netherlands until 2005. The associated edit rules are given by (3), with  $n=4$  and  $m=n-1=3$ . (That is, all balance variables are split into returns and costs.) Table 1 also displays four example records, the last three of which are inconsistent. These example records have been constructed for this paper with nice ‘round’ amounts to improve readability, but the types of inconsistencies present were taken from actual records from the structural business statistics of 2001.

14. Example (a) from Table 1 does not violate any edit rule and therefore does not require any editing. In example (b) two edit rules are violated: the external sum and the internal sum with  $k=1$ . Interestingly, the results block can be made fully consistent with all edit rules by changing the value of  $X_1$  from ‘10’ to ‘-10’. This is the natural way to obtain a consistent results block here since any other

<sup>3</sup> In the new questionnaire of the structural business statistics at Statistics Netherlands (used from 2006 onward) the pre-tax results are split into five balance variables, of which only two break down into returns and costs.

explanation would require more variables to be changed. Moreover, it is quite conceivable that the minus sign in  $X_1$  was left out by the respondent or ‘lost’ during data processing.

15. Two internal sums are violated in example (c), but the external sum is valid. The natural way to obtain a consistent results block here is by interchanging the values of  $X_{1,r}$  and  $X_{1,c}$ , and also of  $X_{3,r}$  and  $X_{3,c}$ . By treating the inconsistencies this way, full use is made of the amounts actually filled in by the respondent and no imputation of synthetic values is necessary.

Table 1. Structure of a results block in the structural business statistics, with four example records.

<i>variable</i>	<i>full name</i>	<i>ex. (a)</i>	<i>ex. (b)</i>	<i>ex. (c)</i>	<i>ex. (d)</i>
$X_{0,r}$	operating returns	4,300	2,100	5,100	3,250
$X_{0,c}$	operating costs	3,900	1,950	4,650	3,550
$X_0$	operating results	400	150	450	300
$X_{1,r}$	financial revenues	30	0	0	110
$X_{1,c}$	financial expenditure	10	10	130	10
$X_1$	operating surplus	20	10	130	100
$X_{2,r}$	provisions rescinded	10	20	20	50
$X_{2,c}$	provisions added	50	5	0	90
$X_2$	balance of provisions	-40	15	20	40
$X_{3,r}$	exceptional income	25	50	15	30
$X_{3,c}$	exceptional expenses	0	10	25	10
$X_3$	exceptional result	25	40	10	20
$X_4$	pre-tax results	405	195	610	-140

## B. Detection and correction

16. The two types of errors found in examples (b) and (c) are quite common. We will refer to them as *sign errors* and *interchanged returns and costs*. They are closely related, and should therefore be searched for by one detection algorithm. We will now formulate such an algorithm, working from the assumption that if an inconsistent record can be made to satisfy all edit rules in (3) by only changing signs of balance variables and/or interchanging returns items and costs items, this is indeed the way the record should be corrected. It should be noted that the 0<sup>th</sup> returns and costs items differ from the other variables in the results block in the sense that they are also present in other edit rules, connecting them to items from other parts of the survey. E.g. the operating returns equal the sum of a list of operating returns items. If these variables were interchanged to suit the 0<sup>th</sup> internal sum, other edit rules might be violated. When detecting sign errors we therefore introduce the constraint that we are not allowed to interchange  $X_{0,r}$  and  $X_{0,c}$ .

17. The notions put forward in the previous paragraph can be given the following mathematical formulation. An inconsistent record contains a sign error and/or interchanged returns and costs if the following set of equations in  $(s_0, \dots, s_n; t_1, \dots, t_m) \in \{-1, 1\}^{n+1} \times \{-1, 1\}^m$  has a solution:

$$\begin{cases} X_0 s_0 = X_{0,r} - X_{0,c} \\ X_1 s_1 = (X_{1,r} - X_{1,c}) t_1 \\ \vdots \\ X_m s_m = (X_{m,r} - X_{m,c}) t_m \\ X_n s_n = X_0 s_0 + X_1 s_1 + \dots + X_{n-1} s_{n-1} \end{cases} \quad (4)$$

Note that in (4) the values of  $X_{0,r}$ ,  $X_{0,c}$ ,  $X_0$ , etc. are used as known constants rather than unknown variables. Thus a different set of equations in  $s_j$  and  $t_k$  is constructed for each record. Moreover, once a solution to (4) has been found, it immediately tells us how to obtain a consistent results block: if  $s_j = -1$  then the sign of  $X_j$  must be changed, and if  $t_k = -1$  then the values of  $X_{k,r}$  and  $X_{k,c}$  must be interchanged. It is easy to see that the resulting record satisfies all edit rules (3). It is also clear from (4) that we are not allowed to interchange  $X_{0,r}$  and  $X_{0,c}$ , since no variable  $t_0$  is present in the equations.

18. By way of illustration we set up (4) for example (d) from Table 1:

$$\begin{cases} 300s_0 = -300 \\ 100s_1 = 100t_1 \\ 40s_2 = -40t_2 \\ 20s_3 = 20t_3 \\ -140s_4 = 300s_0 + 100s_1 + 40s_2 + 20s_3 \end{cases}$$

This system has the following solution:  $s_0 = t_2 = -1$  and  $s_1 = s_2 = s_3 = s_4 = t_1 = t_3 = 1$ . This solution tells us that the value of  $X_0$  should be changed from '300' to '-300' and that the values of  $X_{2,r}$  and  $X_{2,c}$  should be interchanged. This correction indeed yields a fully consistent results block with respect to (3).

19. An important question is: does system (4) have a unique solution? It can be shown (Scholtus, 2008) that the following holds true: if  $X_0 \neq 0$ ,  $X_n \neq 0$  and if the equation

$$\lambda_0 X_0 + \lambda_1 X_1 + \dots + \lambda_{n-1} X_{n-1} = 0 \quad (5)$$

does not have any solution  $\lambda_0, \lambda_1, \dots, \lambda_{n-1} \in \{-1, 0, 1\}$  for which at least one term  $\lambda_j X_j \neq 0$ , then *if* the inconsistency in the record can be resolved by changing signs and/or interchanging returns and costs, it can be done so in a unique way. Translated roughly, this means that the inconsistency can be resolved uniquely *unless* there exists some simple linear relation between the balance amounts  $X_0, X_1, \dots, X_{n-1}$ , e.g. if two balance amounts happen to be equal. Since no such linear relation exists by design, for the great majority of inconsistent records containing sign errors the inconsistency can indeed be resolved uniquely. For instance, in the data of the structural business statistics of 2001 we found that over 95% of all records satisfied this condition for uniqueness.<sup>4</sup>

20. In the example above it is clear that the equation  $300\lambda_0 + 100\lambda_1 + 40\lambda_2 + 20\lambda_3 = 0$  has no non-trivial solution for  $\lambda_0, \lambda_1, \lambda_2, \lambda_3 \in \{-1, 0, 1\}$ , which shows that the solution obtained in paragraph 18 is in fact the only solution.

21. Finding a solution to (4) may be reformulated as a binary linear programming problem:

$$\min \left( \sum_{j=0}^n \sigma_j + \sum_{k=1}^m \tau_k \right), \text{ subject to:}$$

---

<sup>4</sup> Of course, the condition is irrelevant for most of these records, because they do not contain any sign error or interchanged returns and costs. It should also be noted that the condition is sufficient for uniqueness, but not necessary.

$$\begin{aligned}
& X_0(1-2\sigma_0) - (X_{0,r} - X_{0,c}) = 0 \\
& X_1(1-2\sigma_1) - (X_{1,r} - X_{1,c})(1-2\tau_1) = 0 \\
& \quad \quad \quad \vdots \\
& X_m(1-2\sigma_m) - (X_{m,r} - X_{m,c})(1-2\tau_m) = 0 \\
& X_0(1-2\sigma_0) + X_1(1-2\sigma_1) + \dots + X_{n-1}(1-2\sigma_{n-1}) - X_n(1-2\sigma_n) = 0 \\
& \quad \quad \quad \sigma_0, \sigma_1, \dots, \sigma_n; \tau_1, \dots, \tau_m \in \{0,1\}
\end{aligned} \tag{6}$$

where

$$\sigma_j = \frac{1-s_j}{2}, \quad \tau_k = \frac{1-t_k}{2}. \tag{7}$$

Note that in this formulation the number of variables  $s_j$  and  $t_k$  that are equal to  $-1$  is minimised, i.e. the solution is searched for that results in the smallest number of changes being made in the record. If a unique solution to (4) exists, then this is of course also the solution to (6). The binary linear programming problem may be solved by applying a standard branch and bound algorithm. Since  $m$  and  $n$  are small, very little computation time is needed to find the solution. By formulating the problem this way standard software may be used to solve it, making it easier to implement the search algorithm for sign errors.

22. Sign errors and interchanged returns and costs are quite common. An application of the preceding search algorithm to data collected in the structural business statistics of 2001 revealed that about one in five records contained such an error, or even one in three if only the records with an inconsistent results block are considered. It is useful to resolve these errors at the beginning of the editing process.

#### IV. ROUNDING ERRORS

##### A. Introduction

23. Many edit rules for the structural business statistics take the form of a linear equality, e.g. a number of component items that should add up to a total amount. These so-called *balance edit rules* are often violated by the smallest possible difference: the difference between the total amount and the sum of the items is equal to 1 or 2. We call such inconsistencies *rounding errors*, because they are likely to be caused by values being rounded off to multiples of 1,000 Euros. It is not straightforward to obtain a so-called *consistent rounding*, i.e. to make sure that the rounded off values satisfy the same restrictions as the original values.<sup>5</sup>

24. By their nature, rounding errors have virtually no influence on publication figures, and in this sense the choice of method to correct them is unimportant. However, the complexity of the automatic error localisation problem in SLICE increases rapidly as the number of violated edit rules becomes larger, irrespective of the magnitude of these violations. Thus a record containing many rounding errors and very few ‘real’ errors might not be suitable for automatic editing by means of SLICE and have to be edited manually. This is clearly a waste of resources. It is therefore advantageous to resolve all rounding errors at the beginning of the editing process, during the correction step for obvious inconsistencies.

25. In the remainder of this section, we describe a heuristic method to resolve rounding errors in business survey data. We call this method a heuristic method because it does not return a solution that is ‘optimal’ in some sense, e.g. that the number of changed variables or the total change in values is minimised. The rationale of using such a method is that the adaptations needed to resolve rounding errors are very small, and that it is therefore not necessary to use a sophisticated and potentially time-consuming search algorithm.

---

<sup>5</sup> The difficulty is illustrated by this example: suppose that the terms of the sum  $2.7 + 7.6 = 10.3$  are to be rounded off to natural numbers. If ordinary rounding is applied, the additivity is destroyed:  $3 + 8 \neq 10$ .

## B. The scapegoat algorithm

26. When the survey variables are denoted by the vector  $\mathbf{x} = [x_1 \ \dots \ x_v]'$ , the balance edit rules can be written as a linear system  $\mathbf{R}\mathbf{x} = \mathbf{0}$ , where each row of the  $r \times v$ -matrix  $\mathbf{R}$  defines an edit rule and each column corresponds to a survey variable. Denoting the  $i^{\text{th}}$  row of  $\mathbf{R}$  by  $\mathbf{r}'_i$ , an edit rule is violated when  $|\mathbf{r}'_i\mathbf{x}| > 0$ . The inconsistency is called a rounding error when  $0 < |\mathbf{r}'_i\mathbf{x}| \leq 2$ . Similarly, the edit rules that take the form of a linear inequality can be written as  $\mathbf{Q}\mathbf{x} \geq \mathbf{0}$ , where each edit rule is defined by a row of the  $q \times v$ -matrix  $\mathbf{Q}$ . We first assume that only balance edit rules are given.

27. The idea behind the heuristic method is as follows. For each record containing rounding errors, a number of variables is selected beforehand. Next, the rounding errors are resolved by only changing the values of the selected variables. For this reason the name ‘scapegoat algorithm’ has been suggested. In fact, the algorithm guarantees that exactly one choice of values exists for the selected variables such that the balance edit rules are satisfied. Different variables are selected for each record to minimise the effect of the adaptations on published aggregates.

28. It is assumed that the  $r \times v$ -matrix  $\mathbf{R}$  satisfies  $r \leq v$  and  $\text{rank}(\mathbf{R}) = r$ , that is: the number of variables should be at least as large as the number of restrictions and no redundant restrictions may be present. Clearly, these are very mild assumptions. Additionally, the scapegoat algorithm becomes simpler if  $\mathbf{R}$  is a *totally unimodular* matrix. A matrix is called totally unimodular if the determinant of every square submatrix is equal to  $-1$ ,  $0$  or  $1$  (Walukiewicz, 1990). So far we have found that matrices of balance edit rules used for structural business statistics at Statistics Netherlands are always of this type.

29. An inconsistent record  $\mathbf{x}$  is given, possibly containing both rounding errors and other errors. In the first step of the scapegoat algorithm, all rows of  $\mathbf{R}$  for which  $|\mathbf{r}'_i\mathbf{x}| > 2$  are removed from the matrix. We denote the resulting  $r_0 \times v$ -matrix by  $\mathbf{R}_0$ . It is easy to see that if  $\mathbf{R}$  satisfies the assumptions from paragraph 28, then so does  $\mathbf{R}_0$ . Hence  $\mathbf{R}_0$  has full row rank and has  $r_0$  linearly independent columns. The *first*  $r_0$  linearly independent columns may be found by applying Gaussian elimination. Since we want the choice of columns to vary between records, we perform a random permutation of columns beforehand. The variables of  $\mathbf{x}$  are given the same permutation. Next,  $\mathbf{R}_0$  is partitioned into two submatrices  $\mathbf{R}_1$  and  $\mathbf{R}_2$ . The first of these is a  $r_0 \times r_0$ -matrix that contains the linearly independent columns, the second is a  $r_0 \times (v - r_0)$ -matrix containing all other columns. The vector  $\mathbf{x}$  is also partitioned into subvectors  $\mathbf{x}_1$  and  $\mathbf{x}_2$  containing the variables associated with the columns of  $\mathbf{R}_1$  and  $\mathbf{R}_2$ , respectively. The linear system  $\mathbf{R}_0\mathbf{x} = \mathbf{0}$  is equivalent to the system  $\mathbf{R}_1\mathbf{x}_1 = -\mathbf{R}_2\mathbf{x}_2$ . At this point, the variables from  $\mathbf{x}_1$  are selected to be changed and the variables from  $\mathbf{x}_2$  remain fixed. Therefore the values of  $\mathbf{x}_2$  are filled in from the original record and we are left with the system  $\mathbf{R}_1\mathbf{x}_1 = \mathbf{c}$ , where  $\mathbf{c}$  is a vector of known constants. By construction the square matrix  $\mathbf{R}_1$  has full rank and is therefore non-singular. Thus a unique solution  $\tilde{\mathbf{x}}_1 = \mathbf{R}_1^{-1}\mathbf{c}$  exists.

30. In general this solution might contain fractional values, whereas most business survey variables are restricted to be integer-valued. If this is the case, a controlled rounding algorithm could be applied to the values of  $[\tilde{\mathbf{x}}_1' \ \mathbf{x}'_2]'$  to obtain an integer-valued solution to  $\mathbf{R}_1\mathbf{x}_1 = -\mathbf{R}_2\mathbf{x}_2$ . Note however that this is not possible without slightly changing the value of at least one variable from  $\mathbf{x}_2$ . If  $\mathbf{R}$  happens to be a totally unimodular matrix, this problem does not occur. In that case the determinant of the non-singular submatrix  $\mathbf{R}_1$  is equal to  $-1$  or  $1$  and it follows as an easy corollary of *Cramer's Rule*<sup>6</sup> that the solution to  $\mathbf{R}_1\mathbf{x}_1 = \mathbf{c}$  is always integer-valued.

<sup>6</sup> *Cramer's Rule* says that the linear system  $\mathbf{A}\mathbf{x} = \mathbf{b}$ , with  $\mathbf{A}$  a non-singular  $n \times n$ -matrix, has the unique solution

$$x_k = \frac{\det \mathbf{B}_k}{\det \mathbf{A}}, \quad k = 1, \dots, n,$$

### C. Illustration of the algorithm

31. To illustrate the scapegoat algorithm we give a small-scale example. Suppose a record of eleven variables  $x_1, \dots, x_{11}$  is given along with a set of five balance edit rules:

$$\begin{cases} x_3 = x_1 + x_2 \\ x_2 = x_4 \\ x_8 = x_5 + x_6 + x_7 \\ x_9 = x_3 + x_8 \\ x_{11} = x_9 - x_{10} \end{cases} \quad (8)$$

These balance edit rules can be written as a linear system  $\mathbf{R}\mathbf{x} = \mathbf{0}$ , where  $\mathbf{x} = [x_1 \ \dots \ x_{11}]'$  and

$$\mathbf{R} = \begin{bmatrix} 1 & 1 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 & 1 & -1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 1 & -1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & -1 & -1 \end{bmatrix}. \quad (9)$$

It is easily established that  $\mathbf{R}$  has full row rank. Moreover,  $\mathbf{R}$  is a totally unimodular matrix. An example of a record that is inconsistent with respect to (8) is given by:

$$\begin{array}{cccccccccccc} x_1 & x_2 & x_3 & x_4 & x_5 & x_6 & x_7 & x_8 & x_9 & x_{10} & x_{11} \\ 12 & 4 & 15 & 4 & 3 & 1 & 8 & 11 & 27 & 41 & -13 \end{array}$$

This record violates every edit rule, except for  $x_2 = x_4$ . In each instance the violation is small enough to qualify as a rounding error. Thus in this example the matrix  $\mathbf{R}_0$  is identical to  $\mathbf{R}$ .

32. A random permutation is applied to the elements of  $\mathbf{x}$  and the columns of  $\mathbf{R}$ . The following might be the result:

$$\mathbf{R} = \begin{bmatrix} 0 & -1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & -1 & 0 & 0 & 1 & 0 & 0 & 0 \\ -1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 1 & 1 & 0 \\ 1 & 1 & 0 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & -1 & 1 & 0 & -1 & 0 & 0 & 0 & 0 & 0 \end{bmatrix} \quad (10)$$

and

$$\mathbf{x} = [x_8 \ x_3 \ x_{10} \ x_9 \ x_4 \ x_{11} \ x_7 \ x_2 \ x_6 \ x_5 \ x_1]'. \quad (11)$$

It so happens that the first five columns of the permuted matrix  $\mathbf{R}$  are linearly independent. We obtain:

$$\mathbf{R}_1 = \begin{bmatrix} 0 & -1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & -1 \\ -1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & -1 & 0 \\ 0 & 0 & -1 & 1 & 0 \end{bmatrix} \text{ and } \mathbf{x}_1 = \begin{bmatrix} x_8 \\ x_3 \\ x_{10} \\ x_9 \\ x_4 \end{bmatrix} \quad (12)$$

and also

---

where the matrix  $\mathbf{B}_k$  is found by replacing the  $k^{\text{th}}$  column of  $\mathbf{A}$  by  $\mathbf{b}$ . The theorem is named after the Swiss mathematician Gabriel Cramer (1704-1752), although he never stated it in this form.

$$\mathbf{R}_2 = \begin{bmatrix} 0 & 0 & 1 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ -1 & 0 & 0 & 0 & 0 & 0 \end{bmatrix} \text{ and } \mathbf{x}_2 = \begin{bmatrix} x_{11} \\ x_7 \\ x_2 \\ x_6 \\ x_5 \\ x_1 \end{bmatrix}. \quad (13)$$

The values from the original record are put into  $\mathbf{x}_2$  and the constant vector  $\mathbf{c} = -\mathbf{R}_2 \mathbf{x}_2$  is calculated. We obtain the following system in  $\mathbf{x}_1$ :

$$\mathbf{R}_1 \mathbf{x}_1 = \begin{bmatrix} 0 & -1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & -1 \\ -1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & -1 & 0 \\ 0 & 0 & -1 & 1 & 0 \end{bmatrix} \cdot \begin{bmatrix} x_8 \\ x_3 \\ x_{10} \\ x_9 \\ x_4 \end{bmatrix} = \begin{bmatrix} -16 \\ -4 \\ -12 \\ 0 \\ -13 \end{bmatrix} = \mathbf{c}. \quad (14)$$

This system has the following solution:  $\tilde{\mathbf{x}}_1 = [\tilde{x}_8 \quad \tilde{x}_3 \quad \tilde{x}_{10} \quad \tilde{x}_9 \quad \tilde{x}_4]' = [12 \quad 16 \quad 41 \quad 28 \quad 4]'$ .

33. If the original values of the variables in  $\mathbf{x}_1$  are replaced by these new values, the record becomes consistent with respect to the balance edit rules (8):

$$\begin{array}{cccccccccccc} x_1 & x_2 & \tilde{x}_3 & \tilde{x}_4 & x_5 & x_6 & x_7 & \tilde{x}_8 & \tilde{x}_9 & \tilde{x}_{10} & x_{11} \\ 12 & 4 & 16 & 4 & 3 & 1 & 8 & 12 & 28 & 41 & -13 \end{array}$$

We remark that in this example it was not necessary to change the value of every variable in  $\mathbf{x}_1$ . In particular,  $x_4$  and  $x_{10}$  have retained their original value.

#### D. An extension to accommodate linear inequalities

34. In addition to balance edit rules, business survey variables usually have to satisfy a large number of edit rules that take the form of linear inequalities. For instance, it is very common that most variables are restricted to be non-negative. The scapegoat algorithm as described above does not take this into account. A non-negative variable might therefore be changed by the algorithm from 0 to  $-1$ , resulting in a new violation of an edit rule. We will now extend the algorithm to prevent this.

35. Suppose that in addition to the balance edit rules  $\mathbf{R}\mathbf{x} = \mathbf{0}$ , the survey data also have to satisfy  $\mathbf{Q}\mathbf{x} \geq \mathbf{0}$ , where each row of the  $q \times v$ -matrix  $\mathbf{Q}$  defines an edit rule. For a given record, we call a variable *critical* if it occurs in an inequality that becomes an exact equality when the current values of the survey variables are filled in. That is to say:

$$x_j \text{ is a critical variable} \Leftrightarrow \text{for some } i \text{ it holds that } q_{ij} \neq 0 \text{ and } \mathbf{q}'_i \mathbf{x} = 0,$$

where  $\mathbf{q}'_i = [q_{i1} \quad \dots \quad q_{iv}]$  denotes the  $i^{\text{th}}$  row of  $\mathbf{Q}$ . As a special case, any variable that is restricted to be non-negative and currently has the value 0 is critical. To prevent the violation of edit rules in  $\mathbf{Q}\mathbf{x} \geq \mathbf{0}$ , no critical variable should be selected for change during the execution of the scapegoat algorithm.

36. A way to achieve this works as follows: rather than randomly permuting all variables (and all columns of  $\mathbf{R}_0$ ), two separate permutations should be performed for the non-critical and the critical variables. The permuted columns associated with the non-critical variables are then placed to the left of the columns associated with the critical variables. This ensures that linearly independent columns are found among those that are associated with non-critical variables, provided there is a sufficient number of non-critical variables in each record. In practice this is almost always the case, because the number of survey variables is much larger than the number of balance edit rules.

## V. FINAL REMARK

37. A complication that we have not mentioned is the fact that rounding errors often occur in conjunction with obvious inconsistencies, thereby making them 'less obvious'. For instance, a sign error might be obscured by the presence of a rounding error. If the methods described in sections III and IV are applied directly, neither of these errors is detected. A simple way to circumvent this is to postpone the detection of rounding errors until all obvious inconsistencies have been resolved. The methods used for finding obvious inconsistencies should then be adapted to take the possibility of rounding errors into account. This is usually not difficult.

38. As an example, for the detection of sign errors and interchanged returns and costs, instead of (6) the following binary linear programming problem should be solved:

$$\min \left( \sum_{j=0}^n \sigma_j + \sum_{k=1}^m \tau_k \right), \text{ subject to:}$$

$$\begin{aligned} -2 &\leq X_0(1 - 2\sigma_0) - (X_{0,r} - X_{0,c}) \leq 2 \\ -2 &\leq X_1(1 - 2\sigma_1) - (X_{1,r} - X_{1,c})(1 - 2\tau_1) \leq 2 \\ &\vdots \\ -2 &\leq X_m(1 - 2\sigma_m) - (X_{m,r} - X_{m,c})(1 - 2\tau_m) \leq 2 \\ -2 &\leq X_0(1 - 2\sigma_0) + X_1(1 - 2\sigma_1) + \dots + X_{n-1}(1 - 2\sigma_{n-1}) - X_n(1 - 2\sigma_n) \leq 2 \\ &\sigma_0, \sigma_1, \dots, \sigma_n; \tau_1, \dots, \tau_m \in \{0, 1\} \end{aligned} \quad (15)$$

This ensures that a sign error is still detected and resolved if it is obscured by a rounding error. Once the sign error has been removed, the rounding error may be detected and resolved by applying the scapegoat algorithm of section IV.

## VI. REFERENCES

- De Jong, A. (2002), Uni-Edit: standardized processing of structural business statistics in The Netherlands. Paper presented at the UNECE Work Session on Statistical Data Editing, 27-29 May 2002, Helsinki, Finland.
- De Waal, T. (2003), A simple branching scheme for solving the error localisation problem. Research paper 03010, Statistics Netherlands.
- De Waal, T. and R. Quere (2003), A fast and simple algorithm for automatic editing in mixed data. *Journal of Official Statistics*, **19**, pp. 383-402.
- EDIMBUS (2007), *Recommended Practices for Editing and Imputation in Cross-Sectional Business Surveys*. Downloadable from the Eurostat website.
- Fellegi, I.P. and D. Holt (1976), A systematic approach to automatic edit and imputation. *Journal of the American Statistical Association*, **71**, pp. 17-35.
- Hoogland, J. (2006), Selective editing using Plausibility Indicators and SLICE. In: *Statistical Data Editing, Volume No. 3: Impact on Data Quality*. United Nations, New York and Geneva, pp. 106-130.
- Scholtus, S. (2008), Algorithms for correcting some obvious inconsistencies and rounding errors in business survey data. Research paper, Statistics Netherlands (forthcoming).
- Walukiewicz, S. (1990), *Integer Programming*. Kluwer Academic Publishers, Dordrecht.