WP. 43
ENGLISH ONLY

UNITED NATIONS STATISTICAL COMMISSION and
ECONOMIC COMMISSION FOR EUROPE

CONFERENCE OF EUROPEAN STATISTICIANS

Work Session on Statistical Data Editing
(Ottawa, Canada, 16-18 May 2005)

Topic (iv): New and emerging methods, including automation through machine learning, imputation, evaluation of methods

AUTOMATIC EDITNG SYSTEM FOR TWO SHORT-TERM BUSINESS SURVEYS

Supporting Paper

Submitted by Statistical Office of the Republic of Slovenia [1]

I. INTRODUCTION

1. Data editing is the phase of the statistical process that, in recent years, could be considered as one of the most "popular" areas for research and development of new methods. The main reason for such a wide range of activities is of course the possibility of cost reduction since it is estimated that, especially in the case of business surveys, the costs for data editing can represent 40-60% of the total survey budget. It is also known that usually in the case of business surveys only few errors contribute a significant impact on the final results, whereas there are many errors whose correction would change the final results negligibly. So if we were able to detect the errors with a large impact and check manually only these records and all the other erroneous data were corrected by some automatic method, we would be able to substantially reduce the data editing costs. Since in the business surveys we are usually aware of the fact that units are of great importance, there should be no problem to determine the units whose erroneous data could have a significant impact on the final results. A theoretically and methodologically much more demanding part is to set up an efficient data editing system for those units which are not manually rechecked. Throughout past few years we can find many papers and articles that develop theoretical concepts and methods, whereas on the other hand there is much less evidence about the application of these concepts into the reality of specific statistical processes.

2. The Statistical Office of the Republic of Slovenia (SORS) began developing methods for automatic data editing in 2004 and decided to start with the short-term business surveys. The reason for this decision was first of all the growing demand for quick results of these surveys and consequently the demand for quick and efficient flow of the statistical process. The efficient method for data editing would naturally contribute a lot to the goal of shortening the period between the time when the data are captured and the time of dissemination of the results. The main distinctive characteristics of the short-term business surveys, which should be taken into account, are:

- There is usually a short list of variables that we should control;
- The variables are usually continuous with non-negative values;
- The data could be examined cross-sectionally (data for all units in a fixed time period) or longitudinally (data for a fixed unit through time). In the paper we will refer to those two aspects of the data as a vertical (cross-sectional) aspect and horizontal (longitudinal) aspect of the data.

[1] Prepared by Rudi Seljak, Department for Sampling and Survey Methodology (rudi.seljak@gov.si) ; Tomaž Špeh, Department for Electronic Data Processing, Infrastructure and technology (tomaz.speh@gov.si)

3. The paper presents the application of two well-known methods – the Hidiroglou-Berthelot method for detection of outliers (H-B method) and the Fellegi-Holt method for errors localization and data imputation (F-H method) for the case of two short-term business surveys – the Monthly Survey on Turnover, New Orders and Value of Stocks in Industry and the Monthly Survey on Wages. First, we give some basic information about the surveys, then we present the general approach which should be developed into a generalized system for short-term business data editing, describe the application of this system for the two above-mentioned surveys and present some conclusions at the end.

II. ABOUT THE SURVEYS

A. Monthly Survey on Turnover, New Orders and Value of Stocks in Industry

4. The Monthly Survey on Turnover, New Orders and Value of Stocks in Industry has been conducted since January 2003 and provides the data for calculation and dissemination of indices of turnover and new orders. We started to calculate and disseminate these indices in 2004. For the future we are also planning to use these data for the calculation of the industrial production index (IPI), which is currently being calculated by using quantitative data gathered from another monthly survey where the units report the data on produced quantities. Since this survey presents quite a burden for the reporting units as well as for our office, we decided to use the deflated turnover and change of stocks as the basis for the IPI calculation in the future.

5. Some important characteristics of the survey are also:

- The observation unit of the survey, prescribed also by Eurostat's regulation, is a kind of activity unit (KAU).
- Approximately 2000 units are included in the survey. The method of selection is cut-off sample, meaning that all the units with more than 20 employees were selected and in those 2-digit NACE groups where the selected units didn't reach the threshold of 85% of the total number of employees, the largest of the remaining units were additionally selected in order to reach the desired threshold.
- The survey is carried out monthly by mail and the published results are indices only.
- The questionnaire consists of 5 questions, so we have five variables (Turnover – domestic market; Turnover – foreign market; New orders – domestic market; New orders – foreign market; Value of stocks).
- The survey is a typical example for the case where a small number of units is "responsible" for the great deal of the published results. At the present we still do the editing by the classical method, where the data that failed some of the edits are rechecked by contacting the reporting units or sometimes corrected by clerical staff themselves. For the future we are planning that approximately 30% of the units that we consider to be the most important for the final results would still be rechecked, whereas the erroneous data of the other 70% of the units would be corrected automatically.

B. Monthly Survey on Wages

6. The Monthly Survey on Wages is the largest short-term statistical survey conducted by SORS. By the current methodology approximately 24000 units are included in the survey, whereas by the new methodology, which should be implemented in the first half of 2005, even more than 40000 units should be included. In addition to the substantial enlargement of the observed population, the most important methodological change is the fact that our office will no longer be in charge of the data collection. This job will be taken over by one of our administrative agencies and the statistical office will only be responsible for statistical processing of the data and dissemination of the results. Most of the data will be collected through an Internet application containing the most important edits, which should ensure good quality of the data coming into our office. In spite of this system, we don't expect these data to be totally without errors while the system will not be able to detect for instance those errors connected to the distribution of the data. Therefore, we are planning to control the whole set of the data again in the office and all the data which will still fail some of the edits will be corrected automatically.

7. Other important characteristics of the survey are:

- The observation unit is local kind of activity unit (LKAU).
- By the old methodology all the local kind of activity units with at least three employees were included. By the new methodology all units with at least one employee will be included.
- The questionnaire consists of 9 questions, so we have 9 variables in the data. The most important among these variables are the total monthly net and gross wage paid out by the unit and the number of employees who have received the wage.
- The most important results of the survey are the average wages for different domains and indices of average wage.

III. GENERAL SYSTEM

8. Here we present the system for data editing which we are currently setting up for the case of short-term business surveys and which will in the future hopefully become a general system for the wide range of statistical surveys. The editing process could roughly be divided into two major parts:

- Detection of outliers, which is based on the Hideroglou-Berthelot method and uses the ratio of values of the variables in two consecutive time periods.
- Error localization and data imputation of the determined erroneous data. This part is based on the famous Fellegi-Holt method, using the properly defined, complete set of edits.

9. We will separately describe both of the parts of the system, but first we give some notation used. We assume that we have n variables observed in the time period $t: \{X_{it}\}_{i=1}^{n}$. All of these variables are assumed to be continuous with the non-negative values. For each of the variables X_{it} we also define the following auxiliary variables:

$$IX_{it} = \begin{cases} 1; & X_{it} > 0 \\ 0; & \text{otherwise} \end{cases}$$

We will call this variable sign-indicator for the variable X_{it}.

$$RX_{it} = \begin{cases} X_{it} / X_{it\text{-}1}; & \text{if } X_{it\text{-}1} \text{ exists and } X_{it\text{-}1} > 0 \\ 0; & \text{otherwise} \end{cases}$$

We will call this variable time-ratio for the variable X_{it}.

Many times we can find for the particular variable X_{it}, variable Y_{it} which is highly positively correlated to variable X_{it}. Since this correlation could be used especially in the phase of data imputation, we define:

$$CX_{it} = \begin{cases} \mathrm{Y_{it}}; & \text{if such variable exists} \\ 1; & \text{otherwise} \end{cases}$$

We will call this variable correlate for the variable X_{it}.

Detection of outliers

10. Following the procedures suggested in [4], we examine the distribution of time-ratios $r_i = RX_{it}$ inside a particular domain to detect erroneous or at least suspicious values of variable $X_{i,t}$. As described in [4], we need to calculate the median r_M of $\{r_i\}$ and then transform values of r_i twice:

$$s_i = \begin{cases} 1 - r_M / r_i; \text{if } r_i < r_M \\ r_i / r_M - 1; \text{if } r_M \le r_i \end{cases}$$

$$E_i = s_i \left(Max\{X_{it}, X_{it-1}\} \right)^U$$

Calculating $d_{Q1} = MaxE_M - E_{Q1}, |AE_M|)$, $d_{Q3} = \{MaxE_{Q3} - E_M, |AE_M|\}$, where E_{Q1}, E_M, E_{Q3} are first quartile, median and third quartile of $\{E_i\}$, we define outliers as those units for which the value of E_i is outside the interval $\left(E_M - C \cdot d_{Q1}, E_M + C \cdot d_{Q3}\right)$. The values of parameters U, A, C are subject of decision when we are applying the method.

11. The bounds of the interval are of course the same for all the units inside the chosen domain. Since for the need of the further process we would like to have these bounds defined for each individual record with regard to the value of variable X_{it}, we do the inverse transformations. If E_l and E_u are lower and upper bounds for E_i, we calculate:

$$s_l = E_l / \left(Max\{X_{it}, X_{it-1}\} \right)^U$$
$$s_u = E_u / \left(Max\{X_{it}, X_{it-1}\} \right)^U$$
$$r_l = r_M / (1 - s_l)$$
$$r_u = r_M \cdot (1 + s_u)$$
$$X_l = r_l \cdot X_{it-1}$$
$$X_u = r_u \cdot X_{it-1}$$

X_l and X_u are now lower and upper bounds for values of $X_{i,t}$ which should be indicated as outliers.

Error localisation and imputation

12. The basis for efficient data editing is the set of well-defined edits[2]. These edits could be given in many different forms, but it is preferable to find such a standardised form which would be as general as possible and would enable us to set up the procedure which would be easy to translate into computer language. From the case of two discussed short-term surveys and also for the most of the rest of our short-term surveys, all the edits could be divided into two groups:

- Edits which could be written in the form of inequalities. Mathematically this could be written as $a_0 + \sum_{j=1}^{n} a_j \cdot X_{jt} < 0$ where a_j could be a constant or a name of one of the auxiliary variables. We will call these edits type A edits. If X_l is the lower bound for outliers of variable X_{it}, calculated as described above, we have $a_0 = X_l$ and $a_j = \begin{cases} -1; \text{if} & j = i \\ 0; \text{otherwise} \end{cases}$. Similarly if X_u is the upper bound, we have $a_0 = -X_u$ and $a_j = \begin{cases} 1; & \text{if } j = i \\ 0 & ; \text{otherwise} \end{cases}$.

[2] Edit here means the condition for the erroneous record.

- Edits which are based on the rule that two variables must be either both positive or both equal 0. If X_{kt} and X_{lt} are the variables under consideration, the record fails the edit if $(X_{kt} > 0 \text{ and } X_{lt} = 0)$ or $(X_{kt} = 0 \text{ and } X_{lt} > 0)$. Using sign-indicators, these edits could be mathematically written in the form of equation $\sum_{j=1}^{n} b_j \cdot IX_{it} = 1,$ where exactly two of the coefficients b_j equal 1 and all others equal 0.

13. Now, let us assume that we have m_A edits of type A and m_B edits of type B, each of the edits of type A defined with $(n+1)$ coefficients and each of the edits of type B defined with n coefficients. The input for the computer procedure are then two matrices, the first one with dimension $m_A \times (n+1)$, defined with the coefficients of the edits of type A, and the second one with dimension $m_B \times n$, defined with the coefficients of the edits of type B. In the first phase of the procedure these two matrices are used to determine the failed records. The coefficients of the matrices for each edit also indicate weather a particular variable is explicitly involved in this edit or not. With this information it is quite easy to implement the procedure of finding the smallest set of variables to be corrected as it was suggested in [3].

14. After finding the smallest set of variables to be imputed, we have to determine the range of acceptance for each of these variables. Here we present the algorithm for the case when only one variable has to be imputed. By the system of sequential imputations it is not difficult to generalise the algorithm for the case when several variables have to be imputed. Let $X_{i,t}$ be the variable which has to be imputed. We first have to calculate the interval $[l_i, u_i]$ which will determine the range of acceptable values. This interval will be derived out of the set of intervals; in fact each of the edits, either of type A or of type B, determines one interval $[l_{ik}, u_{ki}]$. If the edit is of type A, determined with the coefficients $\{a_j\}_{j=1}^{n+1}$, the bounds of the interval are calculated as follows:

$$l_{ik} = \begin{cases} 0 & ; \quad \text{if } a_i \le 0 \\ -\dfrac{1}{a_i} \cdot s_{i0}; & \text{otherwise} \end{cases} ;$$

$$u_{ik} = \begin{cases} \infty & ; \text{if } a_i \ge 0 \\ -\dfrac{1}{a_i} \cdot s_{i0}; & \text{otherwise} \end{cases} ;$$

where $s_{i0} = a_1 \cdot X_{1k} + \cdots + 0 \cdot X_{ik} + \cdots + a_n \cdot X_{nk}$

15. A similar procedure could be employed in the case when the edit is of type B, determined with coefficients $\{b_j\}_{j=1}^{n}$:

$$l_{ik} = s_{i0} - 1$$

$$u_{ik} = \begin{cases} \infty ; & s_{i0} = 1 \\ 1 ; & \text{otherwise} \end{cases} ;$$

where $s_{i0} = b_1 \cdot X_{1k} + \cdots + 0 \cdot X_{ik} + \cdots + b_{nk} \cdot X_{nk}$.

If we have m_A edits of type A and m_B edits of type B, the interval $[l_i, u_i]$ is determined as the intersection of $(m_A + m_B)$ intervals, obtained from the above described procedure:

$$[l_i, u_i] = \bigcap_{k=1}^{m_A + m_B} ([l_{ik}, u_{ik}]).$$

16. After determining the acceptance interval, we can find the imputed value (out of this interval) in several different ways. Here we indicate only some of the options:

- Hot-deck method where the donor is found vertically, among "clean" records in the same time period.
- Hot-deck method where the donor is found horizontally, among the data of the "imputed" unit from previous time periods.
- Nearest neighbour method where the donor is chosen (vertically or horizontally) in the way that the absolute difference $\left|CX_{it} - CX^{d}_{iit}\right|$ between the correlated variable of the imputed variable and the correlated variable of the donor is minimized. If we want to employ this method, we need to add the third matrix as the input. In this matrix for each variable is given the information, which is the correlated variable. In our case this is the matrix $\{c_{ij}\}_{i,j=1}^{n}$, where the elements are determined by a rule:

  $$c_{ij} = \begin{cases} 1; \text{if } j\text{-th variable is a correlate for i-th varible} \\ 0; \qquad \text{otherwise} \end{cases}.$$

  It is also reasonable that in this case we use the value of the correlate directly for the calculation of the imputed value. If we define $RX_{it} = \begin{cases} \dfrac{X_{it}}{CX_{it}}; CX_{it} > 0 \\ X_{it}; \text{otherwise} \end{cases}$, we can calculate the imputed value: $X_{it\_imp} = RX^{d}_{iit} \cdot CX_{it}$. If we want that these imputed value will fall in the acceptance interval, we have to adjust the bounds of the interval which have been determined for the direct values of the X^{d}_{it} into the bounds of the interval for the RX_{it}. These bounds could be calculated as follows: $l^{r}_{i} = \begin{cases} \dfrac{l_i}{CX_{it}}; \quad Y_i > 0 \\ l_i; \text{otherwise} \end{cases}$; $u^{r}_{i} = \begin{cases} \dfrac{u_i}{CX_{it}} \; ; Y_i > 0 \\ l_i \; ; \text{otherwise} \end{cases}$. Now, the particular value could be used as a donor if $RX^{d}_{it} \in \left[l^{r}_{i}; u^{r}_{i}\right]$.

IV. APPLICATIONS OF THE SYSTEM

A. Monthly Survey on Turnover, New Orders and Value of Stocks in Industry

17. For the case of this survey we are planning to apply the system described in the previous chapter in the second half of this year. At the moment all the units that fail some of the edits are still checked manually, while at the same time we also do the automatic editing for a part of the units. Such a system enables efficient testing of the system before putting it into a regular process. The main purpose of the testing is to find the proper values of parameters for the H-B method and to choose the best method for data imputation.

18. In order to determine the appropriate values for the parameters of H-B procedure, we used the data from years 2003 and 2004 which have been controlled and rechecked manually, so we can consider these data to be clean. Through the simulation of the procedure we were changing the values of parameters in order to reach the situation when most of the data would pass the test for outliers. So far the results of testing led us to the following values of the parameters: U=0.5; A=0.05; C=30-50 (depending on the activity domain). Regarding the imputation method, we will probably set the method which will be the combination of the Hot-deck method and the Historical-trend method.

19. For the future we are planning to divide the units at the beginning of the process. Approximately 30% of the most important units will still be subject of the old system, based on the manual checking, while the other 70% of the units will go through the system of automatic editing. Also all the values that

have been imputed for unit non-response will be checked and automatically corrected. The planned statistical process is presented in the next picture:

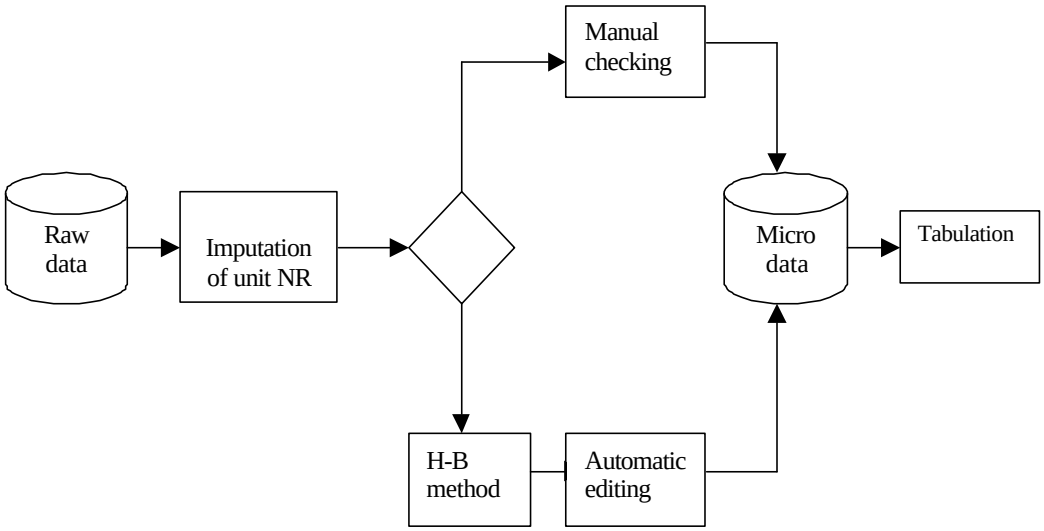

Figure 1: Statistical process in the case of Monthly Survey on Turnover, New Orders and Value of Stocks in Industry

B. Monthly Survey on Wages

20. The Monthly Survey on Wages is, as has already been described, the largest monthly survey conducted by our office, which is just now subject of revision of methodology. Since data collection is now in the hand of an external organization, it is of great importance to have a reliable and efficient system for automatic data editing. As in the case of the above described application, we first determined the values of the parameters of the H-B method. For this purpose data from the previous years were used. All of these data were checked and corrected manually, so we can consider these data to be without errors. The values of the parameters which will be used in the application of the H-B method are: U=0.7; A=0.05; C=30-40 (depending on the activity domain).

21. The first step of the process will be imputation for unit non-response. For the units for which we have data from the previous month, we will use the Historical-trend imputation, whereas for those units for which the data from the previous month are not available, the (adjusted) mean imputation method will be used. Regarding item non-response, we will fix the missing values to 0 and only if these values fail some of the edits, 0 will be replaced with an imputed value. After imputations we will use the H-B method to determine the limits for outliers. Calculating RX_{it} (which distribution is the basis for determination of the limits) we will take into account only those units which have reported (and not imputed) data from the previous month. The limits for outliers will be used to define some of the edits of type A, which will be used in the next step of the process. Together with the rest of the edits of type A and all the edits of type B, these limits will form the basis for the automatic editing, which will be accomplished by the procedure that has been described before. After this procedure all the data should be clean in the sense that they should pass all the defined edits. These clean data will be stored in the final micro-database prepared for tabulation.

22. In order to enable the survey manager good overview and control over the process, we also set up the system of calculating the quality indicators. Indicators for the current month are calculated automatically during the processing and inserted into an Excel spreadsheet, where they can be viewed together with the indicators from all the previous months of the current year. As far as data editing is concerned, the number and rate of corrected reported and corrected imputed values are calculated.

23. The process is graphically presented in the next picture:

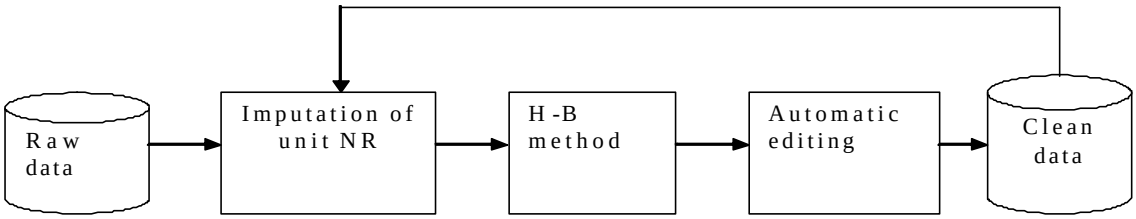

Figure 2: Statistical process for the Monthly Survey on Wages

V. CONCLUSIONS

24. In the paper we shortly described the system for automatic editing which has been firstly applied for the case of two short-term business statistics. While the system is still in the phase of formation and testing, a lot of practical work is still to be expected if we want the system to become more generally applicable. Some of the most important tasks that we anticipate for the future should be:

- To set up the computer program for derivation of implied edits out of the basic set of edits. For the case of the two discussed surveys this work has still been done partly manually.
- To set up the most appropriate imputation method for the case of short-term business statistics. Through the testing and comparing of results we would like to find the best way of combining vertical and horizontal data for the purposes of imputation.
- To set up user friendly computer environment for presentation of input matrices of the edit coefficients and the matrix on the correlates. The plan is to set up such a system that these three matrices, together with the data, would be the only input for the whole system.
- To set up the system for efficient selective editing. In the case of Monthly Survey on Turnover, New Orders and Value of Stocks in Industry we already have some kind of selective editing when only 30% of the units are checked manually. At the moment these units, usually called key respondents are determined once a year by some straightforward rules. For the future we are planning to introduce the system that would use the score function, suggested by several authors.

References

[1] Allen, R.G.D. (1975);" Index Numbers in Theory and Practice": Chicago: Aldine.
[2] Biemer Paul., Lieberg Lars E.," Introduction to Survey Quality": John Willey & sons, 2003.
[3] Fellegi, I.P and Holt D. (1976): A systematic approach to automatic edit and imputation. Journal of the American Statistical Association, 71, 17-35.
[4] Hidiroglou, M.A. and J.M. Berthelot (1986), "Statistical Editing and Imputation for Periodic Business Surveys", Yurvey Methodology, 12, pp. 73-83.
[5] Granquist, L.(1991), "Macro Editing – A Review of Some Methods for Rationalizing the Editing of Survey Data": Statistical Journal, 8, pp. 137-145.
[6] L. Lyberg et al. – Survey measurement and Survey Quality, Wiley, 1997.
[7] Mick Silver (1997),"Business Statistics": The McGraw-Hull Companies.
[8] Methodology of short-term business statistics – Interpretation and guidelines; Eurostat, 2002.
