

**CONFERENCE OF EUROPEAN STATISTICIANS**

**UN/ECE Work Session on Statistical Data Editing**  
(Rome, Italy, 2-4 June 1999)

Topic (i): Measuring the impact of editing in various phases of statistical survey processing

**AUTOMATIC EDITING IN THE DUTCH LABOUR COST SURVEY  
USING CHERRYPI**

Submitted by Statistics Netherlands<sup>1</sup>

**Contributed paper**

**I. Introduction**

1. Statistical offices collect large amounts of data for statistical purposes. An important part of the statistical process consists of checking the collected data and correcting the errors found. This part of the process is called statistical data editing. Traditionally, it is a very costly process as, until recently, it was mostly done by hand. Since this is not very efficient, a lot of research has been devoted to try to automate the editing process. Moreover, very often it is not important to have a data set that is 100 % correct. It is often only important to be able to produce reliable tables. To this end it is usually sufficient to correct the most influential errors. Therefore, it is interesting to test at several stages of the editing process what the resulting key tables would look like if the editing process would be stopped at those stages. By comparing these key tables at several stages it is usually found that the effect of editing diminishes quickly. As soon as no major changes in the key tables are found anymore the editing process may be stopped. This monitoring always has to be an important part of statistical data editing.

2. Changing views on the editing process have important implications for the statistical process of national statistical institutes like Statistics Netherlands. More information about current and future changes in the statistical production process of Statistics Netherlands can be found in Van Bochove (1996). Numerous evaluations show that the substantial costs of editing cannot be justified by the quality improvements obtained. Modern views on necessary changes in the editing strategy can be found in Granquist (1997) and Nordbotten (1998). Different techniques for improving the traditional editing process like selective editing and macro editing could be combined as described in De Jong (1996). An example of including modern imputation techniques into the statistical production process is Schulte Nordholt (1998a).

3. This report deals with the changes in the edit process of the Dutch Labour Cost Survey (LCS). The LCS is a business survey that is conducted every four years by Statistics Netherlands. It contains a few thousand kinds of activity unit records with a lot of detailed information on the cost structure of the production factor labour. The most recent reference year of the LCS is 1996. Up till the last LCS (with reference year 1992) editing was mainly done manually. This implies that the data quality was improved in a rather inefficient way. Therefore, the editing process was changed for the 1996 LCS by introducing selective and automatic edits using a special software tool called CherryPi. This software tool is developed by Statistics Netherlands in order to replace a large part of the manual editing process of

---

<sup>1</sup> Prepared by Eric Schulte Nordholt and Ton de Waal. The views expressed in this paper are those of the authors and do not necessarily reflect the policies of Statistics Netherlands.

economic statistics by automatic editing. Thus, it is possible to reduce costs while keeping the same output available for the different LCS users.

4. In Section II the data of the LCS are described. The software package CherryPi is applied to edit the data of the LCS with reference year 1996 automatically. More information about CherryPi and how this package was applied on the LCS data can be found in Section III. In Section IV the results are presented.

## **II. The LCS data**

5. As a lot of information that is asked in the LCS is not available in an electronic form from the different business units, a paper form is used that has to be filled in and sent back to Statistics Netherlands before a given deadline. To diminish the response burden, the sample size went back from 10,000 kinds of activity units in 1992 to 7,353 in 1996. However, business units with 500 or more employees (firm size 9) remained in the sample. The sample was taken from the Dutch General Business Register that contains data on business units in the Netherlands. Business units with less than 500 employees in the General Business Register were stratified by economic activity and firm size for the sampling process. Unfortunately, not all business units managed to send their form back to Statistics Netherlands on time. Moreover, not all answers were of a reasonable quality. Therefore, a data set of only 5,189 records (70.6 % of the sample size) for the 1996 LCS resulted. To further diminish the response burden, business units that sent their information for the Dutch Annual Survey on Employment and Earnings (ASEE) quarterly to Statistics Netherlands through Electronic Data Capture (EDC) received a shorter form for the LCS. This shorter form contains only one third of the questions of the complete form that is sent to the other business units.

6. Most data for the ASEE are no longer collected in paper form but by EDC from wage administrations. At the moment, the percentage of firms that responds electronically is still modest, although increasing rapidly. Moreover, mainly the large firms switch to EDC quickly, so the number of employee records that are sent electronically to Statistics Netherlands is substantial. From the firms that switched to EDC in principle samples of employees are no longer received, but we do receive tapes with all the employee records. More information about the changes in the data collection process of the ASEE can be found in Arnoldus (1997). We thus have much more earnings information than before. In 1995, the first reference year of the survey, the ASEE data set contains approximately 1,500,000 records with detailed earnings information. The ASEE data set of 1996 contains approximately 2,200,000 job records. The challenge is to enlarge this number of records in a few years to all 6,000,000 employees in the Netherlands. This would imply that in the long run, in principle all business units in the LCS could get the shorter questionnaire and the missing information could be added from the ASEE for all business units.

7. Before the edit process can start the data from the LCS and ASEE are matched. As it is important that different surveys by Statistics Netherlands with overlapping data do not give rise to different conclusions, the data are also compared to other sources. Also some specific information for the LCS is obtained from other sources. The sources concerned are social security files, collective labour agreements, statistics on vocational training, statistics on sick leave and statistics on short-time working, strikes and working hours lost due to frost. That way, the quality of the resulting LCS micro-data is enhanced and the consistency among different data sources is assured. Finally, the LCS data set is ready to be edited.

8. As a small number of large firms with many employees has to be absolutely correct, the data for these firms were corrected manually. It is clear that if we would not have made these manual corrections, the key tables would be of an unacceptably low quality. However, the work involved was done by only one employee of Statistics Netherlands, whereas more than ten people were involved in editing the data manually four years earlier.

9. Some edits that are easy to correct were performed by a quality control computer programme. The input data for CherryPi were thus already partly cleaned. It was considered more efficient to do this partial cleaning before using CherryPi. A lot of information from other sources was used that thus did not have to be read by CherryPi separately. The imputations performed by the quality control programme are sometimes called deductive imputations and can be executed without any risk of damaging the quality of the data set. On the contrary, one can be sure that the quality is greatly enhanced by these deductive imputations. CherryPi could now concentrate on the remaining errors. A given record could still fail a specific edit due to these remaining errors. In this case it is not always clear what the correct or best solution is. The user of CherryPi has to specify which edits and imputations have to be executed on the partly cleaned data. Finally, the user must specify which criterion has to be used to choose from several possible sets of corrections to satisfy the edits.

### III. CherryPi

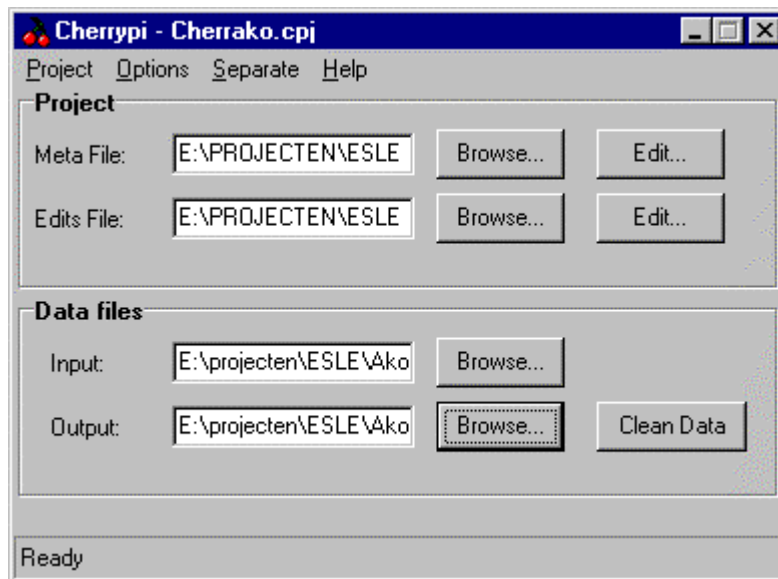
10. The statistical software package CherryPi has been written in Borland Delphi 3.0, and runs under Windows 95. It is a general system for automatic editing and imputing economic data. CherryPi can handle both linear and ratio edit checks. Both kinds of edits can be written as  $Ax \geq b$ , where  $A$  is a constant matrix,  $b$  is a constant vector and  $x$  is a vector corresponding to the values in a given record. The matrix  $A$  and the vector  $b$  together define the set of edits one wants to apply on the micro data. The set of edits does not need to be the same for each stratum. It is possible to enter the different strata and edits both manually in ASCII and by using the interface of CherryPi.

11. The error detection method of CherryPi is based on the Fellegi-Holt methodology (Fellegi and Holt, 1976). The idea is to find the minimum number of fields to be imputed in each record so that all edits can be satisfied. CherryPi, like the Generalized Edit and Imputation System (GEIS) of Statistics Canada, uses a modified version of the Chernikova algorithm to determine these minimal sets of fields per record.

12. The imputation method supported in the current version of CherryPi is regression imputation. At the moment only one auxiliary variable can be used in the regression, but the functionality of the regression imputation module in CherryPi can be extended easily. An improved version of CherryPi and other edit and imputation methods, such as graphical macro editing and the hot deck imputation method, will be implemented in a new software package called SLICE (Statistical Localisation, Imputation and Correction of Errors). SLICE is planned to be totally integrated in the Blaise suite. This will allow the user of the Blaise suite to design the questionnaire, interview respondents, enter data, edit data, impute missing values and weight records with one software tool only. For more information about different imputation methods, their quality and some examples how they can be used in practice we refer to Schulte Nordholt (1998b), and for more information on SLICE we refer to De Waal and Wings (1999).

13. The CherryPi programme consists of four parts. In the first part, the localisation part, (optimal) sets of fields are determined so that all edits can be satisfied by imputing these fields. Of course, one set has to be selected and that happens in the second part, the selection part. Two methods to select a set of imputation variables from several (optimal) sets have been implemented in CherryPi. The first one is simply to select one randomly. In the second method the set of imputation variables with the maximum number of edit checks in which no variables from that set occur is selected. If several sets fulfil this criterion, CherryPi selects one of these sets randomly. After a set of imputation variables has been selected, values have to be imputed for these variables. This occurs in the third part, the imputation part. After the imputation step has been carried out, the resulting records may still fail one or more edit checks. If this is the case the imputed values are adapted to satisfy all edit checks. This is handled in the fourth part, the modification part. CherryPi uses a linear target function that is minimised subject to the edit constraints.

14. To give an idea of what the software package CherryPi looks like, the main window is shown below. After having identified in this window which data to use, the four parts of CherryPi can be executed one by one via the module Separate, or all at once using the Clean Data button.



15. After CherryPi has finished, an imputed micro data set is ready. Of course, this output data set has to be checked on a macro basis. This means that key tables before and after editing have to be compared. This comparison could lead to changes in the way CherryPi is used. The edits in CherryPi could be changed and it is also possible to change the imputation models used before running CherryPi again. Finally the fully edited and imputed data set results. More information about CherryPi can be found in De Waal (1996 and 1998), and De Waal and Van de Pol (1997).

#### IV. The results

16. CherryPi was used on the partly cleaned data set of the 1996 Labour Cost Survey described in Section II. Altogether 45 different edits were used and five different strata were distinguished. Not every edit was applied to all strata, although in one stratum all edits applied. The different strata were created using the variables economic sector (NACE) and category code of the kind of activity unit. In particular, it was necessary to treat employment agencies differently from other business units. In all the edits together 59 variables are involved. Of these variables 56 were allowed to change, but 3 wage variables were forced to stay constant. This means that if an edit with such a wage variable leads to an error, at least one of the other variables involved in this edit has to be changed.

17. In total 1,960 missing values were imputed and 5,040 other imputations were necessary to satisfy all edits. When we study the effect of automatic editing using CherryPi, we can look at the effect on the sum of the values of a variable that is changed in the editing process. The sum of the values of the 56 variables involved that were allowed to change was sometimes increased. However, more often the sum of the values of such a variable was diminished because too high values according to the edit rules were replaced by lower scores.

#### V. Discussion

18. This was the first time CherryPi has been applied to the Labour Cost Survey. Inevitably, there were some problems to apply CherryPi effectively. In particular, it took some time to decide which edits should be executed by the quality control computer programme, and which edits should be executed by CherryPi. Other problems were designing proper edit checks and proper imputation models. Despite these initial problems the application of CherryPi has made the editing process considerably more efficient.

## References

- Arnoldus, F., 1997. Electronic supply of data for labour statistics. In: Netherlands Official Statistics, Volume 12, autumn 1997, pp. 60-68.
- Bochove, C. van, 1996. From assembly line to electronic highway junction: a twin-track transformation of the statistical process. In: Netherlands Official Statistics, Volume 11, summer 1996, pp. 5-36.
- Fellegi, I. and D. Holt, 1976. A systematic approach to automatic edit and imputation. In: Journal of the American Statistical Association, March 1976, Volume 71, Nr. 353, pp. 17-35.
- Granquist, L., 1997. The new view on editing. In: International Statistical Review, Volume 65, Nr. 3, pp. 381-387.
- Jong, W. de, 1996. Designing a complete edit strategy; combining techniques. Report presented at the Work Session on Statistical Data Editing of the United Nations Statistical Commission and Economic Commission for Europe, Voorburg, 4-7 November 1996 (Research Report, Statistics Netherlands, Voorburg).
- Nordbotten, S., 1998. Improving editing strategies. Report presented at the SCB Conference on statistical methods, Stockholm, 12-13 October 1998 (Research Report, University of Bergen, Norway).
- Schulte Nordholt, E., 1998a. Imputation, the alternative for surveying earning patterns. In: Netherlands Official Statistics, Volume 13, spring 1998, pp. 14-15.
- Schulte Nordholt, E., 1998b. Imputation: methods, simulation experiments and practical examples. In: International Statistical Review, Volume 66, Nr. 2, pp. 157-180.
- Waal, T. de, 1996. CherryPi: a computer program for automatic edit and imputation. Report presented at the Work Session on Statistical Data Editing of the United Nations Statistical Commission and Economic Commission for Europe, Voorburg, 4-7 November 1996 (Research Report, Statistics Netherlands, Voorburg).
- Waal, T. de and F. van de Pol, 1997. A recipe for applying CherryPi in the edit process. Report presented at the Work Session on Statistical Data Editing of the United Nations Statistical Commission and Economic Commission for Europe, Prague, 14-17 October 1997 (Research Report, Statistics Netherlands, Voorburg).
- Waal, T. de, 1998. An Introduction to CherryPi and MacroView. Report, Statistics Netherlands, Voorburg.
- Waal, T. de and H. Wings, 1999. From CherryPi to SLICE. Report, Statistics Netherlands.