

CONFERENCE OF EUROPEAN STATISTICIANS

Work Session on Statistical Data Editing
(Prague, Czech Republic, 14-17 October 1997)

Item 6 of the provisional agenda

A RECIPE FOR APPLYING CHERRYPI IN THE EDIT PROCESS

Submitted by Statistics Netherlands¹

¹ Prepared by Ton de Waal and Frank van de Pol.

Abstract

Data files collected by a statistical office generally contain errors. Since aggregate data such as totals and means are usually published, small errors in the data files are acceptable. However, relatively large errors have to be corrected in order to obtain acceptably accurate results. Correcting data can be done in many ways. Traditionally, it is done manually. Over the years several methods to make the traditional editing process more efficient have been developed. Examples of such methods are selective editing, macro editing, graphical editing, and automatic editing. A few years ago Statistics Netherlands initiated the development of a general system for automatically editing and imputing economic data called CHERRYPI. In this paper, the present functionality of CHERRYPI and its planned role in the editing process is discussed.

Keywords: editing, error localisation, Fellegi-Holt paradigm, imputation, economic data

I. INTRODUCTION

1. Data files collected by a statistical office generally contain errors. These errors may arise for a number of reasons: a respondent may have misunderstood some questions; he may have made mistakes while answering the questions; his database may contain errors; at the statistical office, mistakes may have been made while transferring the data from the questionnaires to the computer system, etc.

2. Since aggregate data such as totals and means are usually published, small errors in the data files are acceptable. Small errors will often cancel out when aggregated. However, relatively large errors have to be corrected in order to obtain acceptably accurate results.

3. Correcting data can be done in many ways. Traditionally, it is done manually, i.e. each record is checked either manually or by a computer and, if necessary, is corrected manually. However, this is a time and money consuming process. Over the years, several methods to make the traditional editing process more efficient have been developed. Examples of such methods are selective editing (*cf.* Cruddas, 1995; Engström, 1995; Van de Pol, 1997), macro editing (Barcaroli, Ceccarelli and Luzi, 1995; Granquist, 1995; Van de Pol and Diederer, 1996), graphical editing (*cf.* Engström and Ängsved, 1995; Buijs and Van de Pol, 1996; Esposito, Lin and Tidemann, 1997; Weir, Emery and Walker, 1997), and automatic editing (*cf.* Kovar and Withridge, 1990; Bankier et al., 1995; Winkler, 1996; Barcaroli and Venturi, 1997; Winkler and Petkunas, 1997).

4. A few years ago Statistics Netherlands initiated the development of a general system for automatically editing and imputing economic data called CHERRYPI (*cf.* De Waal, 1996). In this paper the present functionality of CHERRYPI and its planned role in the editing process is discussed. The functionality of CHERRYPI is examined in Section II. In Section III test results of CHERRYPI are briefly discussed. A recipe for using CHERRYPI in practice is given in Section IV. In the final section, Section V, some conclusions are drawn.

II. THE FUNCTIONALITY OF CHERRYPI

Technical remarks

5. CHERRYPI has been written in Borland DELPHI 2.0, and runs under Windows '95. It is suited for parallel computing on several PC's in a network, but it can also run on a single PC.

The edits

6. The edits that can be handled by CHERRYPI can be written as

$$Ax \geq b, \quad (2.1)$$

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. For each stratum a different set of edits can be defined. To enter the set of edits for each stratum quickly and easily, an interface has been developed.

Error localisation

7. The error localisation method of CHERRYPI is based on the Fellegi-Holt paradigm (*cf.* Fellegi and Holt, 1976). According to this paradigm a minimum number of fields should be imputed so that all edits are satisfied. Optionally, a weight may be assigned to each field indicating how trustworthy one considers the value of this field. The higher the weight of a field, the more trustworthy one considers the value of this field. In case weights are assigned to the fields CHERRYPI minimises the weighted number of fields such that all edits are satisfied.

8. To determine the fields that should be imputed according to the Fellegi-Holt paradigm, CHERRYPI applies the algorithm of Chernikova (*cf.* Chernikova, 1964 and 1965; Rubin, 1975). This algorithm generates a subset of the vertices of a set of linear inequalities. Chernikova's algorithm can be used to determine the fields that should be imputed (*cf.* Sande, 1978). However, the basic algorithm is too slow to be useful in practice. Several improvements by Statistics Canada to speed up the error localisation (*cf.* Schiopu-Kratina and Kovar, 1989) have been implemented in CHERRYPI. Through these improvements CHERRYPI has become fast enough to be applied in (most) practical situations.

Imputation

9. An imputation-variable, i.e. a variable that has been selected for imputation by CHERRYPI, that can assume only one value such that all edits can be satisfied is imputed deterministically, i.e. the only value allowed is imputed.

Besides deterministic imputation two kinds of imputation methods will be supported by CHERRYPI, namely imputation based on regression and hot deck imputation. Regression imputation allows historical imputation, ratio imputation and mean imputation. After the imputation-variables have been imputed by means of regression imputation, the resulting record may still violate the edits. Therefore, in a second step the imputed values are modified slightly in order to satisfy all edits. In this way a consistent record is always obtained. For

more information on regression imputation we refer to De Waal (1996).

10. The second imputation method that will be supported by CHERRYPI is hot deck imputation. At the time of writing (mid 1997) hot deck imputation has not yet been implemented, but it will be in near future. CHERRYPI will use a nearest neighbour hot deck approach. That is, CHERRYPI will use the donor record that is as close as possible to the recipient record and that leads to a resulting record that passes all edits. The appropriate values of the donor record are then imputed in the recipient record. The resulting record satisfies all edits, while multivariate distributions are preserved as much as possible. The matching variables that are used to find a suitable donor can either be specified by the user, or can be determined automatically by CHERRYPI.

11. Instead of imputing the values of a donor record in the recipient record, the possibility of imputing only the distribution of the imputation-variables is also being considered. For instance, suppose that variables that sum up to a certain total have to be imputed. In that case we can first divide the values of these variables in the donor-records by the corresponding value of the total. Next we impute the recipient-record by means of the nearest neighbour hot deck method. Finally, we multiply the values of the imputed variables by the value of the total in the recipient-record. In this way we can impute the distribution of the variables, while at the same time making sure that the edits are satisfied after imputation.

Output

12. After CHERRYPI has been executed, a consistent data set is returned. Each record in this data set satisfies all edits. Exceptions are those records for which the (weighted) number of variables that has to be imputed is higher than some user-specified maximum. These records are not imputed. They have to be edited in another way, e.g. manually, or the parameters of CHERRYPI should be adapted. Alternatively, the record could be discarded entirely.

13. CHERRYPI computes for each record the so-called OK-index. This number aims to reflect the trustworthiness of the imputations that have been applied by CHERRYPI. The value of the OK-index is influenced by several factors, such as the number of variables that have been imputed and the changes in estimated population totals due to the imputations. A high value indicates that we can trust the corresponding record, a low value indicates that we should distrust the corresponding record. In the latter case the record should be checked in another way, e.g. by editing it manually. At the moment we have only developed a provisional OK-index, which gives rather unsatisfactory results. That is, the provisional OK-index has difficulties in discriminating between correct imputations and incorrect imputations. We are trying to improve the provisional OK-index.

14. In addition to a consistent data set, statistical information on the applied imputations is returned. For instance, the number of times that a certain variable has been imputed and the total change in this variable due to these imputations are given.

III. TEST RESULTS

15. In an extensive study, the results of CHERRYPI have been tested (*cf.* John, 1997b). Three factors made this study hard to perform in a proper way, and the results somewhat difficult to interpret. Firstly, in the data set that was used, missing values were indicated by a zero, thereby making them indistinguishable from cases where the respondent reported a zero. As a result, CHERRYPI had difficulties in finding the correct variables to impute.

16. Secondly, not all variables seem to have been modified correctly when edited manually. Especially, variables that are considered unimportant seem to have been skipped during the manual edit process.

17. Thirdly, the set of edits that was specified was not correct, and had to be adapted. That the specified set of edits was not correct could be concluded from the fact that records that were edited and imputed manually, and that were considered as correct, did not satisfy all edits, and the fact that a (very) high percentage of the raw, unedited, records did not satisfy the edits.

18. The set of specified edits had to be adapted. In particular the edit bounds of the ratio edits had to be broadened in many cases. For this a simple explorative approach was used. For each ratio edit the raw records were sorted in increasing order of the corresponding ratio. For most ratio edits the following situation was found: for the first few records the ratio increases rather fast, for the majority of the subsequent records the ratio increases less fast, and for the last few records the ratio increases fast again. So, at two points there is a considerable change in the increase of the ratio. These two points were identified as our edit bounds, the first as a lower bound and the second as an upper bound for the corresponding ratio edit. For some other ratios edits there is only one considerable change in the increase of the ratio. In such a case only a lower bound or an upper bound for our ratio edit was identified, depending on whether the large change in increase occurred in the first few records or in the last few records, respectively. More information on the determination of the edit bounds can be found in John (1997a).

19. Using the edit bounds determined in the above way, CHERRYPI was used to obtain a consistent data set. A number of important estimates for population totals were calculated using this data set. These estimated population totals were compared to estimated population totals that were based on a data set that was corrected manually. Despite the problems with the data set and the edits, the results obtained seem quite satisfactory for most variables. For 41% of the variables the difference between the estimated population totals obtained from the data corrected by CHERRYPI and the estimated population totals obtained from the data corrected manually was at most 1%, for 56% of the variables the difference was at most 2%, for 71% of the variables the difference was at most 5%, and for 77% of the variables the difference was at most 10%. For the other variables, however, the difference between the results derived from the data corrected by CHERRYPI and the data corrected manually was larger than 10%. These substantial differences were partly due to the problems with the data set and the edits, but partly they were also due to the fact that some mistakes in records are difficult to correct automatically.

IV. THE ROLE OF CHERRYPI IN THE EDITING PROCESS

20. How should a system for automatic edit and imputation, such as CHERRYPI, be used in practice? Should we rely unconditionally on the results of CHERRYPI, or should we be more careful? The answer to the latter question is clear: CHERRYPI should not be considered as the complete edit process, but only a part of the complete edit process. A possible answer to the former question is given in remainder of this section.

21. Before running CHERRYPI, metadata should be constructed. For instance, when regression imputation is used, the corresponding parameter values should be specified. Once the meta-data have been specified, CHERRYPI can be used to produce a CHERRYPI corrected version of the data. These data are consistent and each record contains the value of the OK-index, which is a function of raw and corrected data (Section 2.5). Records with a low value on this index have undergone corrections with a big influence on publication figures or combine the presence of errors with a large share in publication figures.

22. Records with an OK-index above a certain threshold are considered as part of the 'non-critical stream', which means that the CHERRYPI corrections are accepted without further inquiries. Records with an OK-index below the chosen threshold, the 'critical stream', need manual editing. These records can be edited in the traditional manner, e.g. by comparing current data to data from previous periods or by re-contacting the respondent. Instead of editing the paper form, a tool like BLAISE may be used for computer-assisted micro editing. Contrary to a paper form, BLAISE will give instantaneous feedback on corrections. Corrections that do not match edit checks will be flagged. In addition to this standard BLAISE functionality it should be possible to use the results of CHERRYPI as suggestions in computer-assisted micro editing. To achieve this both the raw, unedited version and the CHERRYPI-corrected version of the record should be read into the BLAISE application.

23. When a sufficient proportion of the sample has been micro edited this way, records can also be edited by comparing them to similar records. For this purpose graphical editing tools, such as MACROVIEW (*cf.* Buijs and Van de Pol, 1996) may be used. The suspect record can be marked in several scattergrams at a time to find out which field is in error. This error tracing by looking for outliers in the multivariate distribution is known as the distribution method variant of macro editing.

24. Micro editing cannot reveal all errors in the data. Some of the unrevealed errors can be found by the aggregate method of macro editing (Granquist, 1995). In MACROVIEW preliminary publication results, that is certain estimates of population totals, are calculated to trace results that do not match expectations. In this process, the nonresponse part of the sample can be compensated for in the usual way, that is by imputation or by weighting. In suspicious publications cells multivariate outliers can be traced by the distribution method mentioned above. After localising a suspicious record in this way, computer-assisted micro editing can be applied for this record, including suggested improvements from CHERRYPI.

25. Thus CHERRYPI can also be used as a supporting tool for other editing tools, such as BLAISE or MACROVIEW. Furthermore, it should be possible to edit part of a record in BLAISE or MACROVIEW, and then run CHERRYPI to edit and impute the remaining fields of the record.

By applying CHERRYPI in the above way in the edit and imputation process (hopefully!) a

substantial part of the records can be edited automatically, while the quality of the automatically imputed values is ensured. The same, or almost the same, publication results will be obtained when the data are edited and imputed in the above way as when they were edited and imputed in a more traditional manner.

V. CONCLUSIONS

26. At the moment, not much work has to be done on CHERRYPI itself anymore. Only hot deck imputation should be implemented and the provisional OK-index should be improved. After that the system is capable of performing the necessary actions for automatic edit and imputation, i.e. localising errors and imputing the corresponding fields. It will be possible to carry out imputation in a variety of ways: deterministically (if possible), by means of regression imputation, or by means of hot deck imputation. After the records have been imputed, information is returned about the statistical effects of these imputations. As some records are too risky to 'correct' automatically, and therefore have to be corrected in an alternative way, an OK-index is computed. The most risky records can be identified by means of the value of the OK-index.

27. Nevertheless much work remains to be done. Firstly, CHERRYPI should be integrated with other editing tools, such as MACROVIEW and BLAISE. Integration of CHERRYPI with these editing tools would enhance the editing power of all systems involved.

28. Secondly, CHERRYPI should be put into practice. This seems to be the hardest task, because most statistical departments at Statistics Netherlands are very reluctant to change their editing process. But there is a sparkle of hope for CHERRYPI, because the management is finally beginning to realise that the editing process can be made much more efficient.

References

Bankier, M., J. Fillion, M. Luc and C. Nadeau, 1995. Imputing numeric and qualitative variables simultaneously. UN Work Session on Statistical Data Editing, 6-9 November 1995, Athens.

Barcaroli, G., C. Ceccarelli and O. Luzi, 1995. A edit and imputation system of quantitative variables based on macro editing techniques. UN Work Session on Statistical Data Editing, 6-9 November 1995, Athens.

Barcaroli, G. and M. Venturi, 1997. DAISY (Design, Analysis and Imputation System) structure, methodology and first applications. *Statistical Data Editing (Volume 2); Methods and Techniques*. United Nations.

Buijs, A. and F. Van de Pol, 1996. A short description of MacroView (in Dutch). Internal note, Statistics Netherlands, Voorburg.

Chernikova, N.V., 1964. Algorithm for finding a general formula for the non-negative solutions of a system of linear equations. *USSR Computational Mathematics and Mathematical Physics*, 4, 151-158.

Chernikova, N.V., 1965. Algorithm for finding a general formula for the non-negative solutions for a system of linear inequalities. *USSR Computational Mathematics and Mathematical Physics*, 5, 228-233.

Cruddas, M., 1995. Using selective editing CSO inquiries. UN Work Session on Statistical Data Editing, 6-9 November 1995, Athens.

De Waal, T., 1996. CHERRYPI: a computer program for automatic edit and imputation. UN Work Session on Statistical Data Editing, 4-7 November 1996, Voorburg.

Engström, P., 1995. A study on using selective editing in the Swedish survey on wages and employment in industry. UN Work Session on Statistical Data Editing, 6-9 November 1995, Athens.

Engström, P., and C. Ängsved. A description of a graphical macro-editing application. UN Work Session on Statistical Data Editing, 6-9 November 1995, Athens.

Esposito, R., D. Lin and K. Tidemann, 1995. The ARIES review system in the BLS current employment statistics program. *Statistical Data Editing (Volume 2); Methods and Techniques*. United Nations.

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

Granquist, L., 1995. Improving the Traditional Editing Process, In B.G. Cox, D.A. Binder, B.N. Chinappa, A. Christianson, M.J. Colledge, P.S. Kott, eds., *Business Survey Methods* (Wiley, New York), pp. 385-401.

John, P., 1997a. Explorative determination of edit bounds for economic data (in Dutch) Internal note, Statistics Netherlands, Voorburg.

John, P., 1997b. A first test of automatic editing with CHERRYPI (in Dutch). Internal note, Statistics Netherlands, Voorburg.

Kovar, J. and P. Whitridge, 1990. Generalized Edit and Imputation System: overview and applications. *Revista Brasileira de Estadística*, 51, 85-100.

Rubin, D.S, 1975. Vertex generation and cardinality constraint linear programs. *Operations Research*, 23, pp. 555-565.

Sande, G., 1978. An algorithm for the fields to impute problems of numerical and coded data. Technical report, Statistics Canada.

Schiopu-Kratina, I. and J.G. Kovar, 1989. Use of Chernikova's algorithm in the generalized edit and imputation system. Methodology Branch Working Paper BSMD 89-001E, Statistics Canada.

Van de Pol, F. and B. Diederens, 1996. A priority index for macro editing the Netherlands foreign trade survey. UN Work Session on Statistical Data Editing, 4-7 November 1996 Voorburg.

Van de Pol, F., 1997. Selective editing in the Netherlands annual construction survey

Statistical Data Editing (Volume 2); Methods and Techniques. United Nations.

Weir, P., R. Emery and J. Walker, 1997. The graphical editing analysis query system
Statistical Data Editing (Volume 2); Methods and Techniques. United Nations.

Winkler, W.E., 1996. The new SPEER edit system. UN Work Session on Statistical Data
Editing, 4-7 November 1996, Voorburg.

Winkler, W.E. and T.F. Petkunas, 1997. The DISCRETE edit system
Statistical Data Editing (Volume 2); Methods and Techniques. United Nations.