STATISTICAL COMMISSION and
ECONOMIC COMMISSION FOR EUROPE

Working paper No. 32
English only

CONFERENCE OF EUROPEAN STATISTICIANS

UNECE Work Session on Statistical Data Editing
(27 – 29 May 2002, Helsinki, Finland)

Topic (iv): Impact of new technologies on statistical data editing

DEVELOPMENT OF MODERN EDIT AND IMPUTATION METHODS AT STATISTICS NETHERLANDS

Contributed paper

Submitted by Statistics Netherlands[1]

Abstract: The development of modern edit and imputation (E&I) methods and software is one of the spearheads of the Methods and Informatics Department of Statistics Netherlands. Many aspects of E&I are covered by the work that is currently being carried out. Software development focuses on the further development of SLICE, a general software framework for automatic E&I. At the moment, SLICE is being extended with the new Cherry Pie module for automatic error localisation in a mixture of categorical and continuous data, with a module for regression imputation of missing continuous data, and with the EC System module for modification of imputed values so that all user-specified edits become satisfied. Besides SLICE, the WAID program for donor imputation of missing values is being developed further. Finally, a software tool for graphical macro-editing based on the functionality offered by SPSS enhanced with Visual Basic modules that are integrated into the SPSS environment is being developed. Methodological research focuses on selective editing and automatic E&I. The merits of classification/regression trees and (logistic) regression for selective editing purposes are currently being investigated. For automatic E&I an ambitious research project is planned to start early 2002. This project will last four years. Aim of this research project is to develop an approach for automatic E&I that integrates the approach based on the Fellegi-Holt paradigm and the NIM approach. Like NIM our approach will use 'predicted' values to locate and correct errors. In contrast to the NIM, the approach will not be based on hot-deck donor imputation exclusively, but will also allow other types of imputation. If no imputation method is specified, the new approach will be the same as the approach based on the Fellegi-Holt paradigm. In this paper the above-mentioned development of E&I methods and software at Statistics Netherlands is described in some detail.

I. INTRODUCTION

1. The Methods and Informatics Department (MID) of Statistics Netherlands (CBS) fulfils both a facilitating and an innovating role. In its facilitating role the department gives advise and provides support on methodological and IT problems to other, statistics producing departments. In its innovating role, the department is responsible for developing new methodology and software that it considers to be promising. In this role the department explores possible directions for the other departments of the bureau. Those other departments may later decide themselves whether or not they use the new methodology and software developed by MID. Research currently carried out by MID is therefore not necessarily the methodology that will be used in future practice. However, generally there is a strong link between research carried out by MID and methodology that is later adopted by statistics producing departments. In this sense, current research at MID is an indicator for the paths CBS may adopt in future.

2. The aim of this paper is to describe research and software development planned to be carried out by MID in the near future, i.e. generally within the next year. Section II briefly sketches the edit and imputation (E&I) process for economic data at CBS. Section III describes the work recently completed

[1] Prepared by Ton de Waal (twal@cbs.nl).

and work planned to be done by MID with respect to selective editing. Section IV is devoted to automatic E&I. It describes the further development of SLICE (a software framework for E&I developed by MID; see e.g. De Waal and Wings, 1999), the further development of WAID, plans for experiments with new error localisation algorithms, and plans for a so-called strategic project. Section V focuses on software for graphical macro-editing. In Section VI we discuss our involvement in the EUREDIT project. Finally, Section VII concludes the paper with a brief discussion.

II. THE E&I PROCESS FOR BUSINESS SURVEYS AT CBS

3. For CBS, E&I for business surveys is a much more serious problem than E&I for social surveys. Social surveys are often collected by means of Blaise, the integrated survey-processing system developed by CBS (see e.g. *Blaise Reference Manual*, 1998; *Blaise Developer's Guide*, 1998). This allows us to correct erroneous and missing data during the data collection phase. Business surveys are usually still collected by means of paper questionnaires. The erroneous and missing answers in these questionnaires hence have to be edited at CBS during the edit and imputation phase.

4. The E&I process for business surveys at CBS has changed rapidly in the last few years. This is due to the IMPECT project. This project has resulted in a new uniform system for E&I for all our structural business surveys. For MID, the IMPECT project is by far the most important client with respect to E&I methodology and software.

5. MID have offered methodological and IT support during the development of the plausibility indexes that are used for selective editing in the IMPECT project. In collaboration with statisticians and subject-matter specialists from the Division of Business Statistics a complicated score function, based on seven constituent score functions, has been developed. For more information regarding this score function and its constituent parts, we refer to Hoogland (2002). The IMPECT project is also by far the most important user of SLICE. Components of SLICE have been integrated into the software system that has been developed for the IMPECT project.

6. Within the IMPECT project, and hence for basically all our structural business surveys, the new editing process is split up into four successive stages: detecting and correcting simple errors, splitting the data into a critical and a non-critical stream by means of selective editing techniques, automatic E&I, and (graphical) macro-editing. The development of modern E&I methods by MID focuses on the latter three stages, and is described in subsequent sections of this paper. For more information on the IMPECT project and the E&I process for business surveys at CBS, we refer to De Jong (2002).

III. SELECTIVE EDITING

7. The application of selective editing techniques to split records into a critical stream, i.e. the records that need to be corrected interactively, and a non-critical stream, i.e. records that can be corrected automatically, is fundamental in order to obtain data of sufficiently high quality. Without solid selective editing techniques, application of automatic editing is bound to lead to data of poor quality.

8. At CBS, selective editing is traditionally based on calculating a score function for each record. The records with scores below a certain threshold value form the critical stream, the remaining records form the non-critical stream. Setting the threshold value is not a serious problem in practice. The threshold value is dictated on the one hand by a minimum for the data quality and on the other hand by the available capacity for cleaning the data. Determining a suitable score function, however, is a serious problem. Statisticians and subject-matter specialists need to devote quite a substantial amount of time to develop a score function that succeeds in identifying the records with influential errors.

9. To reduce the time and resources needed to develop a score function, we have explored two different paths. Firstly, we have examined to the possibility to construct a "recipe book" for score functions (see Van Langen, 2002a). Our aim is to develop a computer program that allows a user to quickly generate a suitable score function in an interactive manner. The user should specify certain characteristics of the data set to be edited, and certain preferences based on subject-matter knowledge.

The computer program should then return a proposal for a suitable score function. Parameters required for the score function should be calculated automatically by the computer program. So far, we have not yet succeeded in writing such a recipe book. In near future, we plan to continue working on construction of such a recipe book and a related computer program.

10. Secondly, we have explored the possibility to replace score functions by techniques of a more statistical nature, such as (logistic) regression, and classification and regression trees (see Sanders, 2002, and Van Langen, 2002b). Our experiments so far show mixed results. When we want to order errors in a single variable in order of influence, the results of logistic regression and classification/regression trees are generally rather disappointing: the results for a very simple score function are at least as good as the results for complicated logistic regression models and complicated classification/regression trees. When we want to order the total error in entire records in order of influence, the situation is somewhat more encouraging: simple score functions are still at least as good as complicated logistic regression models or classification trees, but regression trees outperform the simple score functions. At the moment of writing the present paper, it is not yet clear whether regression trees also outperform more complex score functions. We plan to continue examining the possibility of using regression trees for selective editing purposes.

IV. AUTOMATIC EDITING AND IMPUTATION

11. Software for automatic E&I forms MID's most important product with respect to data editing and imputation. Research and development of automatic E&I methods and software are hence very important issues for MID on which much work is being done. In Subsection A we discuss the development of SLICE. The strategic project on automatic E&I is the subject of Subsection B. The further development of WAID is discussed in Subsection C. Finally, Subsection D describes planned work on alternative algorithms for automatic error localisation.

A. Development of SLICE

12. At the moment, the most important development with respect to automatic E&I is the design and implementation of the second version of SLICE. Whereas SLICE 1.0 contains a module for error localisation in continuous data, SLICE 2.0 will contain a module for error location in a mix of categorical and continuous data. Both the error localisation module in SLICE 1.0 as well as the module in SLICE 2.0 are based on the Fellegi-Holt paradigm that says that data should be made to satisfy all edits by changing the fewest possible (weighted) number of variables (see Fellegi and Holt, 1976). The new error localisation module will be called Cherry Pie – a small variation on the name of the old module (CherryPi). The mathematical algorithm underlying this module is reported on in Quere and De Waal (2000), and De Waal (2002a).

13. SLICE 2.0 will also contain an improved module for imputation of missing data. This module will use regression imputation to impute for missing data. It will be able to take one-dimensional balance restrictions into account, i.e. after imputation one-dimensional balance edit rules will be satisfied. More complicated edit rules are, however, not taken into account by this imputation module.

14. Finally, SLICE 2.0 will also contain a module to solve the so-called consistent imputation problem, i.e. the problem of imputing for missing data subject to the constraint that all edits become satisfied. At CBS we follow a two-step approach to solve this problem. In the first step we simply impute for missing data without necessarily taking all edits into account, e.g. by means of the above-mentioned imputation module. In a second step we then slightly modify these imputed values such that all edits become satisfied. Whenever we refer to consistent imputation in this paper, we therefore actually mean the modification of imputed values such that all edits become satisfied.

15. The consistent imputation module of SLICE 2.0 is referred to as EC System (an abbreviation of Edit Check System). EC System modifies only imputed values; the original, non-imputed values are not altered by the module. It aims to modify the imputed values as little as possible. EC System can handle a

mix of categorical and continuous data. The mathematical algorithm on which EC System is based is reported on in Kartika (2001) and De Waal (2002b).

16. The combination of an error localisation module based on the Fellegi-Holt paradigm, Cherry Pie, and a consistent imputation module, such as e.g. EC System, allows us to use any imputation method we prefer. Provided the modules are applied in the appropriate order the final data are guaranteed to pass all edits by changing the fewest possible (weighted) number of variables.

17. The second version of SLICE will also be able to read Blaise data and metadata. In particular, it will be able to read edit rules specified in the Blaise language. SLICE 2.0 is planned to be released mid 2002. It is planned to replace SLICE 1.0 in the IMPECT project.

18. Users and potential users of SLICE have already expressed their interest in SLICE 2.0. In particular, the fact that simultaneous treatment of categorical and continuous data in SLICE 2.0 allows users to specify edit rules that could not be specified in SLICE 1.0 is considered important. Also, the possibility to directly use edit rules specified in the Blaise language is considered important. Somewhat surprising to developers at MID, it is not considered important that SLICE 2.0 is likely to be clearly faster than SLICE 1.0. For users at CBS speed is apparently not an issue.

19. Several improvements on SLICE are planned to be implemented after release of version 2.0. The error localisation module of SLICE 2.0 is guaranteed to find all optimal solutions to the error localisation problem, given that a certain maximum number of variables are allowed to be changed. This maximum number of variables allowed to be changed was introduced for two different reasons, a methodological reason and a practical one. The methodological reason is that if many fields need to be changed in order to let a record pass all edits, the quality of the collected data in the record is generally too low to be useful. The practical reason is that the required computing time to find all optimal solutions for records containing many errors becomes (too) high. This computing time increases exponentially in the maximum number of fields that are allowed to change.

20. However, for certain, clearly non-influential records users sometimes want to localise errors automatically even though these records may contain many errors. As the records concerned contain many errors, it seems not very important to determine optimal solutions during the error localisation phase. For instance, it seems not very important that 11 fields are modified for a record for which the optimal solution says that actually “only” 10 fields need to be modified. This allows us to abandon the aim to determine optimal solutions to the error localisation problem for records containing many errors. For such records we will instead develop a heuristic that determines good, but possibly suboptimal, solutions to the error localisation problem. Such a heuristic seems quite easy to construct, using parts of the error localisation algorithm of SLICE 2.0. Details on our proposed heuristic can be found in De Waal (2001b).

21. The error localisation algorithm and consistent imputation algorithm of SLICE 2.0 can handle a mix of categorical and continuous data. Integer-valued data have to be treated as being continuous. This may occasionally lead to incorrect results, in the sense that an integer-valued variable involved in a “solution” to the error localisation problem can only be imputed consistently by giving it a continuous, non-integer, value. In principle, the error localisation algorithm and consistent imputation algorithm can be extended so that integer-valued data are correctly dealt with (see De Waal, 2001a). Such an extension is rather complicated to implement in full generality, however. We will therefore probably take a more conservative approach that guarantees that each solution found can be imputed consistently while taking into account that some variables should attain an integer value, but that does not ensure that all optimal solutions involving integer-valued variables are found. In most practical cases, the conservative approach is likely to give the same results as the more complex exact approach. In De Waal (2002c) some details of our planned approach are sketched.

B. Strategic Project on Automatic E&I

22. Several projects at CBS have been appointed as so-called strategic projects. These strategic projects are not aimed to solve certain practical problems immediately. Instead, these projects are expected to yield positive results for a wide range of surveys after four or five years. For strategic projects substantial funding is available. This funding can be used for collaboration with Dutch universities.

23. Part of our research on automatic E&I forms such a strategic project. Within this strategic project we aim to develop methodology and software based on the idea underlying NIM (see e.g. Bankier, 1999; Bankier et al., 2000). From an abstract point of view one could say that the idea underlying NIM is that the imputations should be the basis for deciding which fields are correct and which are incorrect rather than the edit rules.

24. The implementation of the methodology and software by Statistics Canada is, unfortunately, not suited for CBS as this methodology and software mainly focuses on demographic data from population censuses. Population censuses have not been held in the Netherlands since several decades now, and it is not likely that a population census will be held within the next decade.

25. This means that new methodology and software should be developed that suits the situation at CBS. For this development CBS will co-operate with at least three Dutch universities, the University of Utrecht, the University of Groningen, and the University of Amsterdam, with considerable expertise in imputation. For social data, hot-deck donor imputation methods and quite possibly imputation methods based on Bayesian models are planned to be examined. For economic data, regression imputation methodology is planned to be developed that takes into account that some of the auxiliary variables may be incorrect. The main contributions of CBS to the project will be our expertise on handling edits and our data sets. It is expected that two persons will write a Ph.D. dissertation on automatic E&I as part of this strategic project.

26. The development of methodology for the strategic project is still in a preliminary stage, and our final methodology may differ substantially from the methodology we envisage at this moment. Currently, we plan to approach the problem in the following way. First, we plan to use the error localisation algorithm that is currently being implemented in SLICE 2.0. This error localisation searches a large binary tree for solutions to the error localisation problem. When we find a solution, we plan to determine 'predicted' values for the variables involved in this solution. To determine these predicted values we plan to use the above-mentioned imputation models that are being developed in collaboration with Dutch universities.

27. We hope to be able to take edit rules into account while determining imputations, at least to a limited extent. However, because taking general edit rules into account during the imputation phase is quite difficult, we anticipate that some edit rules will be violated by the proposed imputation values. Another reason why edits may be violated after imputation is that in some cases the statistical properties of the data would be altered too much if data were forced to satisfy all edits. When edits are violated by the proposed imputation values, we plan to examine the trade-off between adjustment of the predicted values versus imputation of additional fields, i.e. versus identifying more fields as being incorrect. The reason for examining this trade-off is that sometimes better values can be imputed, in the sense that totals and distributions of the variables involved can be preserved better, if more than a minimum (weighted) number of fields is identified as being erroneous. In other words, we acknowledge that the Fellegi-Holt paradigm does not always correctly classify the fields as being erroneous or being correct.

28. For the above comparison, we can e.g. generate a similar, but smaller, binary tree as in the error localisation algorithm. Each node of this small binary tree corresponds to a variable. In each node two branches are constructed. For each node corresponding to a variable involved in the solution to the error localisation problem one branch is constructed where the predicted value is filled in, and another branch where the predicted value has to be modified. For each node corresponding to a variable not involved in the solution to the error localisation problem one branch is constructed where the original value is filled

in, and another branch where the original value has to be modified. By growing such a tree we can in principle compare all options of modifying the predicted values to imputing some additional variables.

29. As an alternative to growing a small binary tree, we may apply techniques known from Operations Research, which were developed to solve similar problems. In particular, algorithms developed for the so-called linear fixed charge problem appear to be useful in this context. For each variable in the linear fixed charge problem we have a fixed charge for changing its value and continuous costs per unit of change. In our case, we could introduce fixed charges for identifying additional fields as being incorrect, and continuous costs for changes in the predicted values of the variables already identified as being incorrect. De Waal (2002c) provides some more details on formulating our problem of comparing adjustment of the predicted values with imputation of additional fields as a linear fixed charge problem. For more information on the linear fixed charge problem in general we refer to Hirsch and Dantzig (1968), and McKeown and Ragsdale (1990).

30. In any case, after the comparison, we plan to store the best solution, and proceed with searching for better solutions to the error localisation problem in the large binary tree. This process goes on until the entire large binary tree has been searched.

31. The approach sketched above is more general than an approach based on the Fellegi-Holt paradigm. In fact, in case no predicated values are used to identify errors the approach reduces to the standard Fellegi-Holt approach. Some more details on our preliminary thoughts on how to compare modification of imputed values to imputation of additional fields can be found in De Waal (2002c).

C. Further Development of WAID

32. WAID is a software package for imputation of missing data that has been developed as part of the AUTIMP project. It is based on Automatic Interaction Detection (AID) trees, cf. Sonquist, Baker and Morgan (1971). Because the developed algorithm gives lower weights to outliers while constructing regression trees, the technique is referred to as Weighted Automatic Interaction Detection (WAID). The methodology of generating WAID-trees is described in a paper by Tsai and Chambers, which is contained in the compilation report by Chambers et al. (2001).

33. At the moment WAID (version 4.1) is being compared to an imputation package developed by a commercial vendor. Partly depending on the outcome of this research CBS will make a decision to what extent CBS will further develop WAID.

34. Assuming the above-mentioned comparison will not lead to a drastic adaptation of our plan, development of WAID at CBS is restricted to some simple improvements in the current software package. No major changes in the methodology or software are envisaged. Examples of the planned improvements are: fixing some bugs in WAID 4.1, storing information about the generated WAID-trees, storing information about the imputations that have been carried out, and implementing a simple form of multivariate imputation for missing numerical data using a single donor record.

35. Although CBS have not further developed the methodology of WAID and do not plan to do so at this moment, methodological development of WAID has not stopped. At the University of Southampton, Prof. Chambers and his collaborators are currently developing the methodology of WAID further. If this leads to important methodological improvements, CBS will probably incorporate these improvements in their versions of WAID.

D. Alternative Algorithms for Automatic Error Localisation

36. Besides the further development of SLICE, the further development of WAID and the strategic project mentioned above, some more work on automatic E&I is planned. This work is of a more experimental nature and is planned to be carried out by students in mathematics or econometrics from Dutch universities doing a work placement at CBS. This kind of research does not have to lead to results that can directly be applied in practice. The main aim of this kind of research is improvement of our

contact with the scientific community in the Netherlands. Moreover, excellent students are sometimes offered a job at CBS. A secondary aim is increasing our own knowledge.

37. For the near future we plan to experiment with a so-called cutting plane approach for solving the error localisation problem. Such a cutting plane approach for categorical data was first proposed by Garfinkel, Kunnathur and Liepins (1986). In 1988 the same authors (see Garfinkel, Kunnathur and Liepins, 1988) also proposed a cutting plane algorithm for continuous data. Ragsdale and McKeown (1996) proposed improvements on that cutting plane algorithm. De Waal (2002d) proposes another cutting plane algorithm that can be considered as a mix of the algorithms proposed by Garfinkel, Kunnathur and Liepins, Ragsdale and McKeown, and the error localisation algorithm of SLICE 2.0. The proposed algorithm can be applied to a mix of categorical and numerical data. It seems worthwhile to study the performance of such an algorithm on realistic data.

V. GRAPHICAL MACRO-EDITING

38. For several years MID have been developing MacroView, a computer program for graphical macro-editing (see e.g. De Waal, Renssen and Van de Pol, 2000). In 2001 a production version of MacroView, version 1.0, has indeed been released. This version of MacroView has been written in Visual C++. Unfortunately, the functionality offered by MacroView 1.0 was deemed insufficient for the IMPECT project.

39. After much deliberation it was decided to temporise the further development of MacroView in favour of a graphical macro-editing approach based on SPSS in combination with components written in Visual Basic. The software architecture of the present versions of SPSS is such that it can easily be integrated with software components written in another language. Nowadays, it is, e.g., quite easy to add some buttons to the task bar of SPSS. These buttons can, once they are clicked, execute Visual Basic programs. The user of SPSS is not aware that actually Visual Basic programs are called and executed. For the user, the new buttons are similar to the other buttons on the task bar.

40. By extending the functionality of SPSS with some special purpose modules written in Visual Basic, the graphical macro-editing functionality required by the IMPECT project could be achieved within a very short period of time. At this moment, it is examined whether the developed Visual Basic components can be made more general so they can also be used for other surveys that are not included in the IMPECT project. Depending on the outcome, a decision about the future of MacroView will be made: will we proceed with the development of this package, or will we rely on the functionality offered by SPSS in combination with some additional modules written in Visual Basic?

V. PARTICIPATION IN EUREDIT

41. On behalf of CBS, MID participate in the EUREDIT project. Within this project we have developed the basic algorithms, and related prototype software, for automatic error localisation (see De Waal and Pannekoek, 2002; De Waal, 2002a) and modification of imputed values (see De Waal, 2002b) that are currently being implemented in SLICE 2.0. We have also implemented and tested standard imputation methods (see Pannekoek and Van Veller, 2002; Pannekoek, 2002). Finally, we have developed methodology to deal with representative outliers in edited sample data (see Renssen, Smeets and Krieg, 2002a and 2002b).

42. The development phase of EUREDIT is now over, and we are in the middle of the evaluation phase. At the moment the developed prototype software and methods for error localisation and imputation are being applied to various evaluation data sets of the EUREDIT projects. This will proceed until September this year.

43. After September the role of CBS in EUREDIT is almost over. From September this year till March next year, the results of the various evaluation experiments performed under the EUREDIT project will be compared. During that period we will compare our results to those of others, and will report our findings. In March next year, the EUREDIT will officially end.

VI. CONCLUSIONS

44. As will be clear from this paper MID are actively involved in research and software development with respect to E&I. Although some of activities, such as our involvement in the EUREDIT project, are about to end soon, other activities, in particular activities planned for the strategic project, will consume a substantial part of our time and energy.

45. Currently, we are in the fortunate position that both (potential) users and management at CBS are supporting our activities with respect to development of E&I methodology and software. Hopefully, this will allow us to do a lot of useful work within the next few years.

References

Bankier, M., 1999, Experience with the New Imputation Methodology Used in the 1996 Canadian Census with Extensions for Future Censuses. UN/ECE work Session on Statistical Data Editing, Rome.

Bankier, M., P. Poirier, M. Lachance and P. Mason, 2000, A Generic Implementation of the Nearest-Neighbour Imputation Methodology (NIM). Paper presented during the 2nd International Conference on Establishment Surveys, 19-21 June, Buffalo.

Blaise Reference Manual, 1998, Department of Statistical Informatics, Statistics Netherlands, Heerlen.

Blaise Developer's Guide, 1998, Department of Statistical Informatics, Statistics Netherlands, Heerlen.

Chambers, R.L., T. Crespo, S. Laaksonen, P. Piela, P. Tsai and T. De Waal, 2001, The AUTIMP-project: Evaluation of WAID. Report, Statistics Netherlands, Voorburg.

De Jong, A., 2002, The Implementation of a New Uniform Edit Strategy for Structural Business Statistics at Statistics Netherlands. Paper to be presented at the UN/ECE Work Session on Statistical Data Editing, Helsinki, 27-29 May 2002.

De Waal, T., 2000a, SLICE: Generalised Software for Statistical Data Editing and Imputation. In: *Proceedings in Computational Statistics* (ed. J.G. Bethlehem and P.G.M. Van der Heijden). Physica-Verlag.

De Waal, T., 2000b, New Developments in Automatic Edit and Imputation at Statistics Netherlands. Paper presented during the UN/ECE Work Session on Statistical Data Editing, Cardiff, 18-20 October 2000.

De Waal, T., 2001a, Automatic Edit and Imputation in Categorical, Continuous and Integer Data. Report, Statistics Netherlands, Voorburg.

De Waal, T., 2001b, Potential Improvements in LEO/Cherry Pie and EC System. Report, Statistics Netherlands, Voorburg.

De Waal, T., 2002a, An Algorithm for Solving the Error Localisation Problem in Mixed Data. *Deliverable 4.1.1 of the EUREDIT project.*

De Waal, T., 2002b, An Algorithm for Solving the Consistent Imputation Problem in Mixed Data. *Deliverable 5.1.1 of the EUREDIT project.*

De Waal, T., 2002c, Algorithms for Automatic Error Localisation and Modification. Paper to be presented during the DATACLEAN conference, 29-31 May 2002, Jyväskylä, Finland.

De Waal, T., 2002d, An Algorithm for Solving the Error Localisation Problem in General Data Based on Cutting Planes. Report, Statistics Netherlands, Voorburg.

De Waal, T., 2002e, Imputation for Missing and Erroneous Data under Edit Restrictions (work in progress). Report, Statistics Netherlands, Voorburg.

De Waal, T. and J. Pannekoek, 2002, Error Localisation Methodology at CBS. *Deliverable 4.1.1 of the EUREDIT project.*

De Waal, T., R. Renssen and F. Van de Pol, 2000, Graphical Macro-Editing: Possibilities and Pitfall. In: *Proceedings of the second International Conference on Establishment Surveys*, 579-588.

Fellegi, I.P. and D. Holt, 1976, A Systematic Approach to Automatic Edit and Imputation. *Journal of the American Statistical Association*, 71, 17-35.

Garfinkel, R.S., A.S. Kunnathur and G.E. Liepins, 1986, Optimal Imputation of Erroneous Data: Categorical Data, General Edits. *Operations Research*, 34, 744-751.

Garfinkel, R.S., A.S. Kunnathur and G.E. Liepins, 1988, Error Localization for Erroneous Data: Continuous Data, Linear Constraints. *SIAM Journal on Scientific and Statistical Computing*, 9, 922-931.

Hirsch, W.M. and G.B. Dantzig, 1968, The Fixed Charge Problem. *Naval Research Logistics Quarterly*, 9, 413-424.

Hoogland, J., 2002, Selective Editing by Means of Plausibility Indicators. Paper to be presented at the UN/ECE Work Session on Statistical Data Editing, Helsinki, 27-29 May 2002.

Kartika, W., 2001, Consistent Imputation for Categorical and Numerical Data. Report, Statistics Netherlands, Voorburg.

McKeown, P.G. and C.T. Ragsdale, 1990, A Computational Study of Using Preprocessing and Stronger Formulations to Solve Large General Fixed Charge Problems. *Computers & Operations Research*, 17, 9-16.

Pannekoek, J., 2002, Multivariate Regression and Hot Deck Imputation Methods. *Deliverable 5.1.1 of the EUREDIT project.*

Pannekoek, J. and M.G.P. Van Veller, 2002, Evaluation of Multivariate Regression and Hot Deck Imputation Methods. *Deliverable 5.1.2 of the EUREDIT project.*

Quere, R. and T. De Waal, 2000, Error Localization in Mixed Data Sets. Report, Statistics Netherlands, Voorburg.

Ragsdale, C.T. and P.G. McKeown, 1996, On Solving the Continuous Data Editing Problem. *Computers & Operations Research*, 23, 263-273.

Renssen, R., M. Smeets and S. Krieg, 2002a, Dealing with Representative Outliers in Survey Sampling: Methodology. *Deliverable 4/5.2.1 of the EUREDIT project.*

Renssen, R., M. Smeets and S. Krieg, 2002b, Dealing with Representative Outliers in Survey Sampling: Algorithms. *Deliverable 4/5.2.1 of the EUREDIT project.*

Sanders, S., 2002, Selective Editing by Means of Classification and Regression Trees (in Dutch). Internal report, Statistics Netherlands, Voorburg.

Sonquist, J.N., E.L. Baker and J.A. Morgan, 1971, Searching for Structure. Institute for Social Research, University of Michigan.

Van Langen, S., 2002a, Towards a Generic Approach to Selective Editing (in Dutch). Internal report, Statistics Netherlands, Voorburg.

Van Langen, S., 2002b, An Exploration of Selective Editing by Means of Logistic Regression (in Dutch). Internal report, Statistics Netherlands, Voorburg.