Automatic Edit and Imputation for Business Surveys: The Dutch Contribution to the Euredit Project

Discussion paper 03011

Jeroen Pannekoek and Ton de Waal

Statistics Netherlands

Voorburg/Heerlen, November 2003

Explanation of symbols

.	= data not available
*	= provisional figure
x	= publication prohibited (confidential figure)
–	= nil or less than half of unit concerned
–	= (between two figures) inclusive
0 (0,0)	= less than half of unit concerned
blank	= not applicable
2002–2003	= 2002 to 2003 inclusive
2002/2003	= average of 2002 up to and including 2003
2002/'03	= crop year, financial year, school year etc. beginning in 2002 and ending in 2003

Due to rounding, some totals may not correspond with the sum of the separate figures.

Publisher
Statistics Netherlands
Prinses Beatrixlaan 428
2273 XZ Voorburg
The Netherlands

Printed by
Statistics Netherlands - Facility Services

Cover design
WAT ontwerpers, Utrecht

Information
E-mail: infoservice@cbs.nl

Where to order
E-mail: verkoop@cbs.nl

Internet
http://www.cbs.nl

ISSN: 1572-0314
Key figure: X-10
Production code: 6008303011

Statistics Netherlands

AUTOMATIC EDIT AND IMPUTATION FOR BUSINESS SURVEYS: THE DUTCH CONTRIBUTION TO THE EUREDIT PROJECT

Summary: Statistics Netherlands participated in the EUREDIT project, a large international research and development project on statistical data editing and imputation that lasted from March 2000 till February 2003. The main goals of this project were the development and evaluation of new and currently used methods for data editing and imputation. In this paper we describe the general approach applied by Statistics Netherlands on the two business surveys used in the EUREDIT project. We also describe the development of our edit and imputation strategy and give results supporting the choices we have made. Finally, we provide results of our approach on the two evaluation data sets, and compare these results to the results of the other institutes participating in EUREDIT.

Keywords: deductive imputation, error localisation, hot-deck imputation, Fellegi-Holt paradigm, multivariate regression imputation

1. Introduction

The EUREDIT project was a large international research and development project on statistical data editing and imputation. This project was partly funded by the Information Society Technologies Programme of the Framework Programme 5 of the European Union. It involved 12 institutes from seven different countries. Six of those institutes were national statistical institutes, namely Office for National Statistics UK (overall project co-ordinator), Statistics Finland, Swiss Federal Statistical Office, Istituto Nazionale Di Statistica, Statistics Denmark, and Statistics Netherlands (CBS). Four universities participated in the project: Royal Holloway and Bedford New College, University of Southampton, University of York, and University of Jyväskylä. Finally, two commercial companies, the Numerical Algorithm Group Limited and Quantaris, were involved in the project. The project lasted from March 1, 2000 till February 28, 2003.

The aims of the project were:

1. To establish a standard collection of data;
2. To develop a methodological evaluation framework designed to ensure the quality of experimental results, and to allow meaningful and valuable comparisons to be made between different methods for editing and imputation;
3. To evaluate current "in-use" methods for data editing and imputation and to develop and evaluate a selected range of new or recent techniques for data editing and imputation;
4. To compare all methods tested and develop a strategy for users of edit and imputation leading to a "best practice guide";

5. To disseminate selected methods on a project-wide basis by developing prototype software;
6. To exploit the results of the project.

CBS has focussed its attention on goals 3 and 4 mentioned above.

This paper describes the approach applied by CBS on the two business surveys used in the EUREDIT project. This approach mimics part of the currently used approach at CBS for editing and imputing data of annual structural business surveys. We describe the development of our edit and imputation strategy and give results supporting the choices we have made. We also compare our results to the results of the other institutes. This comparison is for a substantial part based on evaluation studies performed by Chambers and Xinqiang (2003a and 2003b) in the EUREDIT project.

In the literature, there is quite a scarcity of articles on the combined application of editing and imputation techniques in practice. The only article similar to the present paper we are aware of is the one by Little and Smith (1987) in the Journal of the American Statistical Association. That article focuses on outlier detection and outlier robust imputation techniques for a relatively small and simple survey. The present paper focuses on automatic editing and imputation techniques for two surveys that are considerably more complex than the one considered by Little and Smith. Moreover, whereas the edit and imputation techniques applied by Little and Smith do not ensure internal consistency of individual records, such as component variables summing up to a total, our procedures ensure such consistency.

Section 2 describes how the evaluation experiments were carried out within the EUREDIT project. The two data sets we consider in this paper are discussed in Section 3. Section 4 sketches the edit and imputation methodology applied by CBS to these data sets. The general outline of our approach is the same for both data sets. Section 5 describes the development of our edit and imputation strategy, and how we have tried to optimise various aspects of this strategy. The same section also compares our evaluation results to the results of the other institutes. Section 6 ends this paper by drawing some conclusions.

2. The evaluation experiments

For each evaluation data set used in the EUREDIT project three different versions were, in principle, constructed: a data set with 'true' values (the Y^* data), a data set with missing data but no errors (the Y_2 data), and a data set with both missing data and errors (the Y_3 data). Constructing a data set with only errors but no missings, a Y_1 data set, was considered to be a too unrealistic scenario.

The Y_2 data and the Y_3 data were sent to all participants in the EUREDIT project. These participants then applied their methods to these data sets. Not all methods were applied to all data sets. A method was not applied to a particular data set if the method was not designed for such a data set, or if it would require too much work to

apply the method. The Y_2 data only had to be imputed, the Y_3 data, in principle, had to be both edited and imputed. However, because certain edit or imputation methods were (too) difficult to apply to the Y_3 data, some institutes only imputed, respectively only edited these data.

The 'true' data are data that the provider of the data set considered to be satisfactorily cleaned according to their edit and imputation procedures. The 'true' data were not sent to the participants in the project. These data were retained by the co-ordinator of the project, the Office for National Statistics (UK), for evaluating the data sets 'cleaned' by the various methods applied.

Along with the Y_2 and Y_3 data sets, metadata related to these data sets, such as edit rules (or *edits* for short) and data dictionaries were delivered to the participants. Besides, development data sets were sent to all participants. Each development data set was available in the three versions, Y^*, Y_2, and Y_3, for all participants. The development data set could be used to train neural networks or to parameterise statistical methods, for instance. The development data sets represent the fact that in a real-life situation one can learn from past experiences. The Y^*, Y_2, and Y_3 data sets can respectively be interpreted as cleaned data, edited but not yet imputed data, and raw data from a previous year. The development data sets contained different records and different information than the evaluation data sets. In this paper we will denote the development Y_2 and Y_3 data sets as $Y_{2,D}$, respectively $Y_{3,D}$, and the evaluation Y_2 and Y_3 data sets as $Y_{2,E}$, respectively $Y_{3,E}$.

To construct the Y_2 and Y_3 evaluation and development data sets, the errors and missingness have been synthetically introduced in the Y^* data set. The errors are not actual errors; neither are the missing data the actual missing data. However, care has been taken to ensure that the errors and missingness are as realistic as possible.

Each participant in the project was allowed to submit several cleaned versions of the same data set, where for each version other parameters or another method was used. The results of the evaluation experiments, i.e. the quality of the cleaned data sets, were assessed by applying a large number of evaluation criteria. These evaluation criteria measured many different aspects of an edit and imputation approach, such as its ability to identify errors, to identify the large errors, to accurately impute individual values, to preserve the distributional aspects of the data, and to estimate publication totals and averages. In the section on the results of our approach, Section 5, we describe a number of such evaluation criteria. We refer to Chambers (2003) for more details regarding the evaluation criteria.

3. The data sets

3.1 UK Annual Business Inquiry

The UK Annual Business Inquiry (ABI) data is an annual business survey containing commonly measured continuous variables such as turnover and wages.

The data sets (Y_2 and Y_3) contain 30 variables. The development data sets contain 4,325 records and the evaluation data sets 6,233. A long and a short version of the questionnaire have been used in the data collection. As a consequence, in the evaluation data sets for 3,970 businesses scores on only 17 variables are available (the short version) and for 2,263 records scores on 32 variables (the long version).

Besides the variables *ref* (an identification number for the business at hand), *weight* (a raising weight for calculating population estimates), and *formtype*, pointing out which version of the questionnaire was used, three other variables are included containing neither missing values nor errors. The variables *class* (anonymised industrial classification), *turnreg* (registered turnover), and *empreg* (registered employment size group) are not obtained from the questionnaires but from completely observed registers. So, in the long form 26 variables contain errors or had missing values, and in the short form only 11 variables.

The variables can be subdivided according to a three-level hierarchy. The first level consists of the key economic variables (*turnover*, *emptotc*, *purtot*, *taxtot*, *assacq* and *assdisp*) and an employment variable (*employ*). The six key ABI economic variables are highly skewly distributed. The second level consists of secondary variables (*stockbeg*, *stockend* and *capwork*) measuring business activity. For the long form, the third level consists of variables corresponding to components of three key economic variables, namely the components of *purtot*, *taxtot*, and *emptotc*. For the short form, the third level consists of two component variables for *purtot*, but no components for the other key economic variables.

With the data edits were provided. These edits can be subdivided into *hard* (or *logical*) edits and *soft* ones. The hard edits by definition hold true for correctly observed records. The soft edits hold true for a high fraction of correctly observed records. The fraction of correctly observed records failing a soft edit is non-zero, however. Hard edits for the ABI data are rules stating that certain variables should attain non-negative values (non-negativity edits), and edits stating that certain variables should sum up to an observed total (balance edits). In total there are 20 non-negativity edits, and 4 balance edits. Some of these hard edits are only applicable for the long questionnaire, some others only for the short questionnaire, the rest for both types of questionnaires. Soft edits for the ABI data are rules stating that the ratio of two variables is less than a specified maximum (ratio edits), and rules stating an upper or lower bound on the value of a certain variable (upper/lower bound edits). In total there are 12 ratio edits, and 13 upper/lower bound rules. Some soft edits are conditional on the type of questionnaire and/or on the values of certain variables. An example of a conditional edit is

$$\text{if } employ > 0 \text{ then } emptot \,/\, employ \geq 4.$$

The edit is satisfied if *employ* is not larger than zero, irrespective of the value of *emptot*. Both the value of *employ* as well as the value of *emptot* may be incorrect. It is possible that an observed positive value of *employ* should in fact be zero.

3.2 Swiss Environmental Protection Expenditures data

The Swiss Environmental Protection Expenditures (EPE) data consists of information on expenditure related to environmental issues. The data are the responses to an environmental questionnaire plus additional general business questions, distributed to enterprises in Switzerland in 1993.

The data sets contain 71 variables in total. The development data sets contain 1,039 records, and the evaluation data sets 200. As for the ABI data, the variables in the EPE data sets can be subdivided according to a three-level hierarchy. The first level consists of four key economic variables: *totinvto*, *totexpto*, *subtot* and *rectot*. These variables are highly skewly distributed. The second level consists of 20 component variables corresponding to these four total variables, namely the components of *totinvto* (*totinvwp*, *totinvwm*, *totinvap*, *totinvnp*, *totinvot*), the components of *totexpto* (*totexpwp*, *totexpwm*, *totexpap*, *totexpnp*, and *totexpot*), components of *subtot*, and the components of *rectot*. Finally, the third level consists of 30 variables that correspond to the components of *totinvwp*, *totinvwm*, *totinvap*, *totinvnp*, *totinvot*, *totexpwp*, *totexpwm*, *totexpap*, *totexpnp*, and *totexpot*.

All edits specified for the EPE data are hard ones. In total there are 54 non-negativity edits, and 23 balance edits. Two of the balance edits can be deleted as they are logically implied by the other balance edits. So, there are 21 non-redundant balance edits. The balance edits follow a complex pattern, basically consisting of two two-dimensional tables and two one-dimensional tables of which the internal cell values have to add up to the marginal totals. The two two-dimensional tables are shown in Tables 3.1 and 3.2. For each table, a column with component variables has to add up to a subtotal variable. For instance, in Table 3.1 *eopinvwp*, *pininvwp* and *othinvwp* have to add up to *totinvwp* (edit (i)). A row with component variables has to add up to a subtotal variable. For instance, in Table 3.1 *eopinvwp*, *eopinvwm*, *eopinvap*, *eopinvnp* and *eopinvot* have to sum up to *eopinvtot* (edit (vi)). All component variables, all column subtotal variables and all row subtotal variables have to add up to a total variable (e.g. *totinvto* in Table 3.1). This is indicated in the tables by C (sum of column totals; e.g., edit (ix) in Table 3.1), R (sum of row totals; e.g., edit (x)) and T (sum of component variables; e.g., edit (xi)).

The two one-dimensional tables say that the components of *subtot* have to sum up to *subtot*, respectively that the components of *rectot* have to sum up to *rectot*.

Table 3.1. Edits that apply to investments for the EPE data.

Investments	Water protection	Waste treatment	Air protection	Noise protection	Other	(Sub)total
End of pipe	*eopinvwp*	*eopinvwm*	*eopinvap*	*eopinvnp*	*eopinvot*	*eopinvtot* (vi)
Process integrated	*pininvwp*	*pininvwm*	*pininvap*	*pininvnp*	*pininvot*	*pininvtot* (vii)
Other	*othinvwp*	*othinvwm*	*othinvap*	*othinvnp*	*othinvot*	*othinvtot* (viii)
(Sub)total	*totinvwp* (i)	*totinvwm* (ii)	*totinvap* (iii)	*totinvnp* (iv)	*totinvot* (v)	*totinvto* (ix) C (x) R (xi) T

Table 3.2. Edits that apply to expenditures for the EPE data.

Expenditures	Water protection	Waste treatment	Air protection	Noise protection	Other	(Sub)total
Current expenditures	*curexpwp*	*curexpwm*	*curexpap*	*curexpnp*	*curexpot*	*curexptot* (xvii)
Taxes	*taxexpwp*	*taxexpwm*	*taxexpap*	*taxexpnp*	*taxexpot*	*taxexptot* (xviii)
(Sub)total	*totexpwp* (xii)	*totexpwm* (xiii)	*totexpap* (xiv)	*totexpnp* (xv)	*totexpot* (xvi)	*totexpto* (xix) C (xx) R (xxi) T

4. Applied methodology

4.1 Overview

In this section a number "standard" edit and imputation methods that were applied by CBS to the ABI and EPE data are briefly described. We have subdivided the edit and imputation problem into three separate problems:

- the error localisation problem: given a data set and a set of edits, determine which values are erroneous or suspicious, and set these values to missing;
- the imputation problem: given a data set with missing data, impute these missing data in the best possible way;
- the consistent imputation problem: given an imputed data set and a set of edits, adjust the imputed values such that all edits become satisfied.

For the first and the last problem, algorithms and prototype software have been applied that were developed at CBS and that were extended as part of the EUREDIT project. For the imputation problem we have used a combination of regression and hot-deck methods implemented in S-Plus scripts. At CBS, we aim to let edited and

imputed data sets satisfy all specified edits. The edits therefore play a prominent role in our methods.

4.2 Error localisation

In this section we describe our methodology for localising the errors in a data set. We distinguish between the localisation of systematic errors and random errors, as these kinds of errors require a different treatment.

4.2.1 Finding systematic errors

A systematic error is an error reported consistently between (some of the) responding units. It can, for instance, be caused by the consistent misunderstanding of a question on the survey by (some of) the respondents. Examples are when gross values are reported instead of net values, and particularly when values are reported in units instead of, for instance, the requested thousands of units (the so-called "thousand-errors"). Since such errors occur in groups of related variables such as all financial variables or all variables related to purchases, they often do not violate edits and can therefore not be found by the Fellegi-Holt based methods (to be discussed in the Subsection 4.2.2).

Thousand-errors can often be detected by comparing a respondent's present values with those from previous years, or by comparing the responses to questionnaire variables with values of register variables. For the experiments in the EUREDIT project only the second option is a possibility. Using the ABI development data, it appeared that a considerable amount of thousand-errors occurred in all financial variables. Most of these errors could be found by calculating the ratio of *turnover* (the reported turnover) to *turnreg* (the turnover value from the register) and deciding that a thousand-error was present if this ratio was larger than 300. All financial variables in such records were then divided by 1,000. In the EPE development data no thousand-errors were detected.

4.2.2 Using the Fellegi-Holt paradigm

Besides systematic errors, observed data also contain non-systematic, random errors. These errors are not caused by a systematic reason, but by accident. An example is an observed value where a respondent by mistake typed in a digit too many. To identify such non-systematic errors we have used the (generalised) Fellegi-Holt paradigm. This paradigm says that the data in a record should be made to satisfy the specified edits by changing the fewest possible (weighted) number of fields. To each variable a non-negative weight, the so-called reliability weight, is assigned that indicates the reliability of the values of this variable. The higher the weight of a variable, the more reliable the corresponding values are considered to be. If all weights are equal, the generalised Fellegi-Holt paradigm reduces to the original Fellegi-Holt paradigm (see Fellegi and Holt, 1976).

An algorithm that implements the Fellegi-Holt paradigm has been developed by De Waal and Quere (2003). In the EUREDIT project we have applied a prototype

program called Leo based on that algorithm. The production version of that prototype is nowadays referred to as *Cherry Pie* and we will use this latter name in the rest of the paper. The most important output of *Cherry Pie* consists of a file that contains for each record a list of all optimal solutions to the error localisation problem, i.e. all possible ways to satisfy the edits by changing a minimum (weighted) number of fields. One of these optimal solutions is selected for imputation (see Subsection 4.2.3). The variables involved in the selected optimal solution are set to missing and are subsequently imputed by the methods described in Subsection 4.3. In general, *Cherry Pie* also generates a file with records for which it could not find a solution, because more fields in these records would have to be modified than a user-specified maximum allows. In our experiments, however, we used Cherry Pie to determine the errors in all records.

4.2.3 Selection of Cherry Pie solutions

In practice it is quite common that application of the Fellegi-Holt paradigm yields several optimal solutions. *Cherry Pie* simply returns all these solutions. To select one of these solutions we have implemented a relatively simple approach. The idea is to calculate first a crude prediction for each variable in all solutions generated by *Cherry Pie*. These predictions are based on register variables only since these are (assumed to be) without errors. Subsequently, distances are calculated between the observed values in a record and the corresponding predicted values in all solutions for that record. The optimal solution returned by *Cherry Pie* for which this distance is maximal is the one involving variables that deviate most from their predicted values. Therefore the maximal distance solution is selected as the solution that will be processed further, i.e. the variables in this solution will be set to missing and these missing values will subsequently be imputed.

The distance function used is the sum of normalised absolute differences between the observed values and the predicted values in a record, i.e.

$$D_k = \sum_{i \in I_k} \left| \frac{y_{ij} - \hat{y}_{ij}}{\sqrt{\hat{\mathrm{var}}(e_{ij})}} \right|$$

where y_{ij} denotes the observed value of variable *j* in record *i*, $\hat{y}_{ij}$ the corresponding predicted value, I_k the index set of the variables in the *k*-th optimal solution returned by *Cherry Pie*, and $\hat{\mathrm{var}}(e_{ij})$ an estimate for the variance of the prediction error. The predictions that we used in applying this approach were ratio-type estimators of the form

$$\hat{y}_{ij} = x_{ij} \frac{\bar{y}_j}{\bar{x}_j}$$

where x_{ij} is the value of the (register) predictor variable for variable y_j in record *i*, $\bar{y}_j$ is the mean over all clean records (records that do not violate any of the edits) of variable y_j, and $\bar{x}_j$ is the mean over the same clean records of x_j. Actually, we used separate ratio estimators within strata but for notational simplicity we only describe the unstratified case here.

4.3 Imputation

In this section we sketch the imputation methods we have applied. For more details we refer to Pannekoek (2003a) and Pannekoek and Van Veller (2003).

4.3.1 Deductive imputation

For a number of missing values in the ABI and EPE data, the value can be determined unambiguously from the edits provided for these data sets. If only a single variable in a balance edit is missing, its value can be derived from the other variables involved in the edit. For non-negative variables we also have that if the total variable of a balance edit equals the sum of the non-missing subtotal variables, the missing subtotal variables are all zero, and similarly for subtotal and component variables. Moreover, if the (sub)total variable has a zero value all missing subtotal (component) variables are zero. Such "deductive" imputations are performed as a first step. For the remaining missing values the methods described below are used.

4.3.2 Multivariate regression imputation

A standard technique for imputing continuous variables is to employ a linear regression model to derive predictions for the missing values. Often, some of the predictor variables also contain missing values and these predictors are then also candidates for imputation. In such cases, there is no distinction between predictor variables and target variables. Let the vector with all variables under consideration be denoted by $\mathbf{y}$ and the value of unit *i* on $\mathbf{y}$ by $\mathbf{y}_i$. For each unit the vector $\mathbf{y}_i$ can be partitioned in an observed part $\mathbf{y}_{i,o}$ and a missing part $\mathbf{y}_{i,m}$. Regression imputation can be based, in this case, on the multivariate regression model that relates each of the missing variables to all of the observed variables:

$$\mathbf{y}_{i,m} = \mu_{i,m} + \mathbf{B}_{m.o(i)}(\mathbf{y}_{i,o} - \mu_{i,o}) + \varepsilon_{i,m} \qquad (1)$$

where $\mu_{i,m}$ and $\mu_{i,.o}$ are the expected values of $\mathbf{y}_{i,m}$ and $\mathbf{y}_{i,o}$, respectively, and $\mathbf{B}_{m.o(i)}$ is the $q_i \times p_i$-matrix with regression coefficients for the multivariate regression of the q_i variables that are missing for unit *i* on the p_i (predictor) variables that are observed for unit *i*. The coefficient matrix $\mathbf{B}_{m.o(i)}$ depends on *i* in the sense that the predictor variables and variables to be predicted may differ between units, but the coefficients are equal for units that have the same missing data pattern.

Estimates of the parameters of (1) can be obtained by using an EM-algorithm. This algorithm is an iterative procedure for obtaining maximum likelihood (ML) estimates (assuming multivariate normality) of the expected value vector and covariance matrix of a set of variables based on data with missing values. This procedure is described by, e.g., Little and Rubin (1987) and Schafer (1997).

Let the ML-estimates of the expected value and covariance matrix of all variables be denoted by $\hat{\mu}$ and $\hat{\Sigma}$, respectively. An estimate $\hat{\mu}_{i,m}$ for $\mu_{i,m}$ can then be obtained by collecting the q_i components of $\hat{\mu}$ corresponding to the missing variables for unit *i* and an estimate $\hat{\mu}_{i,o}$ for $\mu_{i,.o}$ can similarly be obtained by collecting the other p_i components of $\hat{\mu}$. The coefficient matrix can be estimated by

$$\hat{\mathbf{B}}_{m.o(i)} = \hat{\Sigma}_{oo(i)}^{-1}\hat{\Sigma}_{om(i)},$$

where $\hat{\Sigma}_{oo(i)}$ is the submatrix of $\hat{\Sigma}$ containing the estimated variances and covariances of the variables observed for unit *i* and $\hat{\Sigma}_{om(i)}$ is the submatrix of $\hat{\Sigma}$ containing the estimated covariances of the variables observed for unit *i* with the variables missing for unit *i*.

Using these estimates, regression imputations for the missing variables in a record *i* can be obtained by

$$\hat{\mathbf{y}}_{i,m} = \hat{\mu}_{i,m} + \hat{\mathbf{B}}_{m.o(i)}(\mathbf{y}_{i,o} - \hat{\mu}_{i,o}).$$

The regression methods are based on a linear additive model for the data. When such a model is not a realistic approximation for the data, regression imputation may give poor results. In the ABI and EPE data there are a number of non-negative variables with many zero values (often 50% or more). For such variables, the assumption of a linear model for a continuous dependent variable is problematic. The regression imputations will never be zero (unless all predictor variables are) and negative predictions will often occur. With only a few exceptions, these variables are component variables that should satisfy certain balance edits; a requirement that will not be satisfied by regression imputed values. For these variables nearest neighbour hot deck methods have been applied that I) will not impute negative values, II) will impute zero values, and III) ensure that at least some of the balance edits are satisfied by the imputed values. These methods are detailed below.

4.3.3 Hot deck imputation methods

Nearest neighbour hot deck methods use a distance function to measure the distance between records. For each record with missing values (the *receptor record*) on some variables (the *target variables*) a donor record is selected that I) has no missing values on the auxiliary variables and the target variables, and II) has the smallest distance to the receptor record. Imputation is then performed by replacing the

missing values of the target variables in the receptor record by the values of these variables from the donor record.

The distance function used in the applications is the distance function suggested for this purpose by Little and Rubin (1987, p. 66). Before applying a distance function it is customary to scale the auxiliary variables such that they have zero mean and unit standard deviation. This prevents implicit weighting of the variables, in particular if they are measured in different units. Let the values of the scaled auxiliary variables in a record i be denoted by z_{ij} ($j=1...J$), then the distance between records i and i' is defined by

$$d(i,i') = \max_j \left| z_{ij} - z_{i'j} \right| .$$

A donor record is thus chosen such that the maximal absolute difference between the auxiliary variables of the donor and the receptor is minimal.

For variables that are part of a balance edit such as subtotals or component variables we have applied a modified version (which we refer to as *ratio hot deck*) of the "standard" nearest neighbour hot-deck method. This method begins by calculating the difference between the total variable (which is either observed or imputed by regression) and the sum of the observed components. This difference equals the sum of the missing components. The sum of the missing components can then be distributed over the missing components using ratios (of the missing components to the sum of the missing components) from a donor record. In this way the level of the imputed components is determined by the total variable but their ratios (to the total of the missing values) are determined by the donor record. This method ensures that that the imputed and observed components will add up to the total.

Note that if only one of the components is missing, the ratio equals 1, so no donor information is used and the method reduces to a deductive imputation rule derived from the additivity constraint. Also, if the sum of the observed components equals the total, the sum of the missing components is 0 and again a deductive imputation rule, derived from additivity and non-negativity, results. It can happen that a donor is chosen for which all the missing components are zero. Then, the ratio's are undefined (reflecting the fact that such a donor does not contain information on how to distribute the sum over the missing components) and another donor (the next nearest neighbour) is used.

For the ABI data this ratio hot deck method ensures that all hard edits are satisfied because they are either balance edits or non-negativity edits, and each variable occurs only once in each balance edit. The situation is different for the EPE data where many variables are part of more than one balance edit. This is illustrated in Table 3.1 of Section 3.2. Suppose that the subtotals of Table 3.1, i.e. *totinvwp*, *totinvwm*, *totinvap*, *totinvnp*, *totinvot*, *eopinvtot*, *pininvtot*, and *othinvtot* are observed or already imputed. Then we can use the ratio hot deck method and the subtotals *totinvwp*, *totinvwm*, *totinvap*, *totinvnp*, and *totinvot* to impute all component variables in which case these imputed values will not necessarily sum up

to the subtotals *eopinvtot*, *pininvtot*, and *othinvtot*, or vice versa. In such cases where the imputation method does not ensure that edits are satisfied we have used an additional step to adjust the imputed values such that they do satisfy all edits.

4.4 Adjustment of imputed values

Adjustment of imputed values to satisfy the edits is done such that the adjustments are as small as possible. This goal is achieved by defining a distance function measuring the distance between the original record (with imputed values that may not satisfy all edits) and an adjusted record (where imputed values have been changed such that all edits are satisfied). The minimal adjustment problem can then be defined as: minimise the distance function subject to the constraint that the imputed value are modified so that all edits become satisfied. The prototype program we have implemented assumes that the imputed values can indeed be modified such that all edits become satisfied, which is the case if the fields to be imputed were determined by using a Fellegi-Holt system like *Cherry Pie*.

Our minimal adjustment algorithm considers distance functions of the type

$$\sum_{i=1}^{n} w_i \mid y_i - \tilde{y}_i \mid , \tag{2}$$

where the y_i and $\tilde{y}_i$ denote the original and adjusted values of the continuous variables in a record, respectively. The w_i (i=1,…,n) are non-negative user-specified weights that are used to compare a change in one variable to a change in another variable. The higher the weight of a variable, the more serious a change in value is considered to be. These weights differ from the reliability weights used in *Cherry Pie*. Note that $\tilde{y}_i = y_i$ for variables that have not been imputed, because we only modify imputed values. The problem of minimising (2) subject to the constraint that all edits become satisfied is a linear programming problem, and can, for instance, be solved by means of the simplex algorithm (see Chvátal, 1983).

De Waal (2003) considers the more complicated problem of adjusting imputed values in a mix of categorical and numerical data, and proposes a heuristic to solve this problem.

5. Results

In this section we present results of our approach. The performance of our approach as applied to the evaluation data was measured by a number of evaluation criteria, developed in the EUREDIT project. Subsection 5.1 describes some of these criteria that will be used in the subsequent three subsections. In Subsection 5.2 we present some results using the development ($Y_{2,D}$ and $Y_{3,D}$) data (for which the true data are available) that were used to decide on questions such as: how to detect systematic errors, which stratification to use for imputation within strata and which imputation

method to use (regression, hot-deck, ratio hot-deck) for which variables. The result of these choices was a final edit and imputation strategy to be applied to the ABI and EPE evaluation data sets. In Subsection 5.3 we present evaluation results for our methods and in Subsection 5.4. the evaluation results for our strategies are compared with those from other partners in the EUREDIT project.

We present only a limited number of statistical results in this section. For many more results we refer to Pannekoek and Van Veller (2003) for the $Y_{2,D}$ data, Pannekoek (2003b) for the $Y_{2,E}$ data, for Vonk, Pannekoek and De Waal (2003a) for the $Y_{3,D}$ data, and Vonk, Pannekoek and De Waal (2003b) for the $Y_{3,E}$ data. For a detailed comparison to other methods we refer to Chambers and Xinqiang (2003a and 2003b).

5.1 Evaluation criteria

To evaluate the editing and imputation methods, a set of evaluation criteria has been developed within the EUREDIT project. Most of the evaluation criteria can be grouped into five groups (a) to (e). These groups are referred to in some of the tables presented in Subsection 5.4.

Each group consists of several individual measures. Group (a) focuses on pure error finding performance measures. It consists of an alpha, a beta and a delta measure. The alpha measure equals the proportion of cases where the value for the variable at hand is incorrect, but is still judged acceptable by the editing process. It is an estimate for the probability that an incorrect value for variable *j* is not detected by the editing process. The beta measure is the proportion of cases where a correct value for the variable at hand is judged as suspicious by the editing process, and estimates the probability that a correct value is incorrectly identified as suspicious. The delta measure is an estimate for the probability of an incorrect outcome from the editing process for the variable at hand, and measures the inaccuracy of the editing procedure for this variable. Group (b) focuses on significant error finding performance. It includes measures such as the "relative average error", i.e. the weighted difference between the true and the edited values of a variable divided by the weighted total of the true values, and the "relative error range", i.e. the range (maximum – minimum) of the non-zero differences between the true and edited values of a variable divided by the interquartile distance of the true values for all cases. Group (c) focuses on the preservation of moments, and includes, for instance, the "absolute relative error of the mean", i.e.

$$\left| \frac{\sum_{i \in G} w_i Y_i \Big/ \sum_{i \in G} w_i}{\sum_{i=1}^{n} w_i Y_i^* \Big/ \sum_{i=1}^{n} w_i} - 1 \right|,$$

where Y_i^* is the true value of the variable under consideration in record *i*, Y_i the observed value, w_i the raising weight of record *i*, *n* the total number of records, and *G* denotes the set of records for which the observed value Y_i is considered correct.

Group (d) focuses on the distance between the imputed and the true values, and consists of (weighted) distance measures, such as the d_{L1}-measure that is defined as

$$d_{L1} = \sum_{i \in M} w_i \left| \hat{Y}_i - Y_i^* \right| \Big/ \sum_{i \in M} w_i \ ,$$

where $\hat{Y}_i$ is the imputed value in record *i* of the variable at hand, and *M* denotes the set of records with imputed values for variable *Y*. Finally, group (e) focuses on the preservation of distributions. It consists of the Kolmogorov-Smirnov distance, and variations thereof.

Besides the criteria that can be classified into these five groups, a number of other evaluation measures have been used. The most important one of these is the m_1 measure, which measures the preservation of the first moment of the empirical distribution of the true values. In particular,

$$m_1 = \left| \sum_{i \in M} w_i (Y_i^* - \hat{Y}_i \ ) / \sum_{i \in M} w_i \right| .$$

For more information on the evaluation measures used within EUREDIT we refer to Chambers (2003).

Besides the evaluation measures used in the EUREDIT project, we have also used the relative difference in means (*rdm*) defined by

$$rdm = \frac{\sum_{i \in M} \hat{Y}_i - \sum_{i \in M} Y_i^*}{\sum_{i \in M} Y_i^*} .$$

5.2 Developing a strategy

In this subsection we present some results and general considerations that motivated our choices on the following issues: 1) a threshold value to detect thousand-errors, 2) whether or not to use soft edits in the error localisation step, 3) an effective stratification for regression imputation and 4) hot-deck versus regression imputation for component variables and variables with many zero values. At the end of this subsection we also use the development data to demonstrate the influence of errors on imputation performance.

5.2.1 Detecting thousand-errors

We developed our strategy for detecting thousand-errors using the development data. With the true values available, these errors were detected by dividing all perturbed values by their true values. When these ratios are close to 1000, they point to thousand-errors. In 191 records of the ABI $Y_{3,D}$ data, thousand-errors were made in *all* financial variables. As mentioned before in Subsection 4.2.1, we consider a record to contain a thousand-error if the ratio between *turnover* and *turnreg* is larger than 300. This threshold value has been determined by minimising the number of

misclassifications. For a threshold value of 300, 187 thousand-errors were correctly detected, 4 thousand-errors were not detected, 5,905 records were correctly considered not to contain a thousand-error, and for 3 records it was incorrectly concluded that they contain a thousand-error. The number of misclassifications is small, especially if we take into consideration that two thousand-errors could never be detected given their values of zero on *turnreg*.

5.2.2 Edits

Our approach explicitly uses edits specified by subject-matter specialists. The performance of the approach is therefore directly dependent on the quality of the specified edits. If the edits are too strict, (too) many correct records will be identified as containing an error and will be incorrectly modified. On the other hand, if the edits are not strict enough, the edits will indicate that (too) many erroneous records are error-free. The errors in these records will hence not be corrected. In practice, a balance has to be found. Preferably, the edits should be specified so that both the number of correct records that are identified as being erroneous and the number of erroneous records that are identified as correct should be as small as possible.

As discussed in Subsection 3.1, the edit rules for the ABI Y_3 data consist of hard (logical) edits and soft edits. The data should at least satisfy all hard edits, but it is likely that a considerable number of errors remain undetected when using these hard edits only. On the other hand, the soft edits are designed by subject-matter specialists for interactive editing and may be too strict for automatic editing, possibly resulting in a considerable number of correct records that are identified as incorrect. For the application to the $Y_{3,E}$ data we have chosen not to make a selection of edit rules that we expect to perform best but to run two experiments: one that uses all edit rules for error localisation and one that uses only the hard edit rules. For the EPE data, no soft edits were specified by the subject-matter specialists.

5.2.3 Different stratifications for the multivariate regression procedure

As is common for business surveys, the ABI data include an indicator for the type of industry: the variable *class*. Imputation procedures (as well as other estimation procedures) for business surveys are often applied separately for different types of industry, thus allowing the parameters of the imputation model to vary between different types of industry. For the ABI data we considered multivariate regression imputation within 14 strata based on the variable *class*. As an alternative, we also considered a stratification suggested by ISTAT as a result from their experiments with the ABI data (Di Zio, Guarnera, and Luzi, 2003). This stratification is based on the register variables *turnreg* and *empreg* and consists of the following three strata for each formtype: (1) $turnreg < 1{,}000$, (2) $turnreg \geq 1{,}000$ and $empreg \leq 3$, (3) $turnreg \geq 1{,}000$ and $empreg > 3$. The resulting number of strata is 6 for variables that are on both forms and 3 for variables that are only part of either the long form or the short form. This last stratification variable will be referred to as *strat*.

Table 5.1. Preservation of mean values for two types of stratification used with multivariate regression imputation for the ABI $Y_{2,D}$ data.

Variable	True values	Stratification by *class*		Stratification by *strat*	
	mean	mean	*rdm*	mean	*rdm*
turnover	2738.0	2642.5	-0.035	2552.2	-0.068
emptotc	414.3	405.1	-0.022	411.1	-0.008
pursale	3128.7	3359.9	0.074	3231.5	0.033
purtot	2292.1	2266.7	-0.011	2272.1	-0.009
taxtot	45.8	53.5	0.168	51.8	0.131
stockbeg	546.2	418.4	-0.234	420.0	-0.231
stockend	1895.6	1652.7	-0.128	1617.0	-0.147
employ	40.0	48.9	0.223	43.9	0.098

Table 5.1 presents results obtained by applying multivariate regression imputation with the two different stratifications to 8 variables of the ABI $Y_{2,D}$ data set. This table shows the mean of the true values and the mean of the regression imputed values for the two alternative stratifications. All means are calculated over the values that were missing in the $Y_{2,D}$ data set only. In addition this table also shows the relative difference between the mean of the imputed values and the mean of the corresponding true values (*rdm*; see Section 5.1). The results show that stratification by *strat* leads to a better preservation of the mean than stratification by *class* for 6 of the 8 variables, even though the number of classes is much less. Based on these results, stratification by *strat* was used in the evaluation experiments.

The EPE data include a variable *act* (industrial activity) which is comparable in meaning to the variable *class* in the ABI data as well as a variable *emp* (number of employees) without missing values. However, since the number of records for the EPE data is much smaller than for the ABI data, the possibilities for stratification are much more limited. As an alternative to full stratification we have included *emp* and 8 dummy variables for the categories corresponding to the first digit of *act* in the multivariate regression imputation procedure. In this way the regression model used for imputation always includes additive effects of *emp* and *act* (along with other predictor variables, depending on their availability for a particular record) thus providing a differentiation in imputations between industry type and numbers of employees.

5.2.4 Hot deck imputation versus regression imputation

One of the methods considered for component variables was the ratio hot-deck method, but for some component variables we investigated the performance of regression imputation as well.

Table 5.2. Preservation of mean values for component purchase variables in ABI $Y_{2,D}$ data for ratio hot deck and regression imputation.

Variable	True values	Ratio hot-deck		Regression	
	mean	mean	*rdm*	mean	*rdm*
purenoth	345.05	545.18	0.58	272.59	-0.21
purhire	58.51	80.16	0.37	105.32	0.80
purins	148.07	99.21	-0.33	182.13	0.23
purtrans	1350.34	1228.81	-0.09	3551.39	1.63
purtele	111.39	38.99	-0.65	108.05	-0.03
purcomp	28.67	98.05	2.42	168.58	4.88
puradv	3118.52	2806.67	-0.10	3243.26	0.04

In Table 5.2 the means of the imputed values for the ratio hot deck imputation and for the multivariate regression imputation of the component purchase variables of the ABI $Y_{2,D}$ data are compared with the means of the corresponding true values. The differences between the imputed means and true means are presented in terms of the relative difference in means (*rdm*). From these results it appears that for some variables regression is better (*purenoth, purtele, puradv)* but for other variables (*purhire, purtrans, purcomp*) ratio hot deck is better.

These results do not point strongly to one of the imputation methods as the method of choice. The regression method has some disadvantages not shared by the ratio hot deck method. In particular, some imputed values are negative while the corresponding variables should only assume non-negative values, due to convergence problems one of the component variables (*purothse*) had to be left out and other solutions are therefore needed for this variable and, contrary to the ratio hot deck method, the regression imputed component variables will not satisfy the corresponding balance edit. Similar experiments were carried out on the EPE data with comparable results. For these reasons we decided to use ratio hot deck imputation for all component variables.

Some variables such as *assacq* and *assdisp* are not component variables and can therefore not be imputed by the ratio hot deck method, but regression imputation is not well suited either because they contain a large number of zero values. To decide on an imputation strategy for the variables *assacq* and *assdisp*, we have considered several alternatives. The first alternative is a standard nearest-neighbour hot deck method. The second method is a two-step approach, using a hot-deck method to impute whether or not the missing value is zero and subsequently a regression approach to impute only the non-zero values. Negative imputations by the regression step of this method are set to zero. The third method is essentially the same as the second method but uses a log-transform of the target variable to prevent the regression part of this method to impute extra zeros. The fourth method is the multivariate regression method with negative values set to zero. Table 5.3 shows, for the variables *assacq* and *assdisp*, the mean of the true values corresponding to the

missing values, the mean of the imputed values, and *rdm* for the four different imputation methods.

Table 5.3. Preservation of mean values for assacq *and* assdisp *for 4 imputation methods for the ABI $Y_{2,D}$ data.*

Variable	True values	Hot deck		Hot deck/ regression				Regression	
	mean	mean	*rdm*	mean[a]	*rdm*[a]	mean[b]	*rdm*[b]	mean	*rdm*
assacq	620.7	572.0	-0.08	662.4	0.07	617.8	-0.00	682.0	0.10
assdisp	33.4	9.9	-0.70	16.9	-0.49	0.1	-1.00	54.3	0.63

[a]No log-transform, [b]Log-transform.

From Table 5.3 it appears that the smallest difference between the mean of the imputed values and the mean of the corresponding true values for *assacq* is obtained by the combined method with log-transform, but this method is the worst performing method for *assdisp*. For *assdisp*, the method that best preserves the mean is the combined method without log-transform. On the basis of these results, this last method may also be the method of choice if one method is selected for both these variables. However, the regression part of the combined method should only impute positive values (zero values have already been imputed by the hot deck part of the method), and without log-transform many negative imputations arise that are set to zero. This situation is illustrated in Table 5.4 where the number of imputed zero values for the different imputation methods is compared with the true number of zero values.

Table 5.4. Imputation of zeros for assacq *and* assdisp *for 3 imputation methods for the ABI $Y_{2,D}$ data.*

Variable	# Missing values	True values #zeros	Hot deck #zeros	Hot deck/regression[a] #zeros	Regression #zeros
assacq	235	98	124	155	66
assdisp	215	151	169	186	43

[a] No log transform

This table shows that based upon the number of zeros that are imputed, hot deck imputation or the hot deck/regression combination with log transformation of the target variable (which leads to the same number of zero imputed values) should be preferred. However, as we saw before the combined approach with log-transform performed very poorly with respect to the preservation of the mean of *assdisp* (see Table 5.3). If it is important to have the number of firms that have non-zero values for assets disposed (*assdisp*) or assets acquired (*assacq*) about right and at the same time preserve the mean reasonably well, the hot deck method seems to be a good compromise.

5.2.5 Influence of errors on imputation performance

So far, the development data have been used to decide on an edit and imputation strategy to be applied to the evaluation data. These data, for which the true values for both missing values and erroneous values are available, also give us the opportunity to explore the effect of errors on the imputation performance. We will look at this briefly before turning to the evaluation criteria and evaluation data.

In Table 5.5 some imputation results are given for the four overall total variables (imputed by multivariate regression and deductive imputation) for both the EPE $Y_{3,D}$ data set and the EPE $Y_{2,D}$ data. These results include the true mean of the imputed values (mean true), the mean of the imputed values themselves (mean imp), the relative difference between these two means (*rdm*) and the number of imputations (# imp).

Table 5.5. Preservation of mean values for the four overall total variables of the EPE development data.

Variable	Errors and missings ($Y_{3,D}$)				Missings only ($Y_{2,D}$)			
	mean true	mean imp	*rdm*	# imp	mean true	mean imp	*rdm*	# imp
totinvto	1872.67	1073.41	-0.43	21	1509.42	1413.06	-0.06	19
totexpto	206.38	46.38	-0.78	36	1083.45	1083.45	0.00	33
subtot	15.00	21.88	0.46	2	15.00	19.48	0.30	2
rectot	743.18	210.80	-0.72	11	743.18	362.90	-0.51	11

Two of the variables in Table 5.5 (*totinvto* and *totexpto*) contain more missing values for the $Y_{3,D}$ data set than for the $Y_{2,D}$ data set because *Cherry Pie* found errors in these variables. Four errors in *totinvto* and *totexpto* were not detected. For the other two variables, no errors are present or detected. The values of *rdm* show that the means for the $Y_{3,D}$ data are less well preserved than for the $Y_{2,D}$ data, for all variables. In general, the quality of imputations of a regression procedure can be influenced adversely by errors because of two reasons: 1) the values of some of the predictor variables in the records with missing values are erroneous and 2) errors in any of the variables in the records with missing values as well as in fully observed records can lead to biased estimates of the regression coefficients. For these data, however, the four undetected errors are all in records with no missing values. Thus, the lesser quality of the imputations for the $Y_{3,D}$ data can be explained entirely by the influence of the errors on the estimated regression coefficients.

5.3 Application to the evaluation data

In this subsection we will discuss results of the application of our edit and imputation strategy to the evaluation data. First we show results related to the edit rules: results that show the effectiveness of deductive imputation (Subsection 5.3.1) and the amount of adjustment that is necessary to let the imputed values satisfy the

edit rules (Subsection 5.3.2). Next, we give some results for the error localisation performance (alpha, beta and delta measures) and imputation performance (d_{L1} and m_1 measures) for the ABI (Subsection 5.3.3) and EPE data (Subsection 5.3.4).

5.3.1 Deductive imputation

We illustrate the effect of deductive imputation by means of Table 5.6. To save space we only show the results for the end-of-pipe investment variables of the EPE evaluation data. Other variables display a similar behaviour. The columns "missing" give the total number of missings in the $Y_{2,E}$ data set, respectively the $Y_{3,E}$ data set, the columns "deductive" the number of deductively imputed missing values, and the columns "non-deductive" the remaining missing values that were imputed otherwise.

Table 5.6. Number of deductively imputed values in the EPE $Y_{2,E}$ and $Y_{3,E}$ data.

Variable	Data set with missings only ($Y_{2,E}$)			Data set with missings and errors ($Y_{3,E}$)		
	missing	deductive	non-deductive	missing	deductive	non-deductive
eopinvwp	55	24	31	59	27	32
eopinvwm	49	28	21	51	28	23
eopinvap	49	24	25	53	28	25
eopinvnp	36	19	17	37	20	17
eopinvot	18	8	10	23	13	10
eopinvtot	60	18	42	69	27	42

In total, about 42% of the values to be imputed in the EPE $Y_{2,E}$ data could be deductively imputed, and about 45% of the values to be imputed in the EPE $Y_{3,E}$ data. So, a substantial amount of the values to be imputed can be deductively imputed by using the edits. These numbers are slightly lower for the ABI $Y_{2,E}$ and $Y_{3,E}$ data, but there too a substantial number of deductive imputations were carried out. Note that the total number of the fields to be imputed in each of the $Y_{3,E}$ data sets (of ABI and EPE) depends on the number of implausible values that have been identified.

5.3.2 Adjustment of imputed values

In Table 5.7 we illustrate the effect of adjusting imputed values so the final records satisfy the specified edits. We only display the results for the end-of-pipe investments of the EPE $Y_{2,E}$ data. As one can see, only a small percentage of the imputed values is later adjusted. The difference between the means of the imputed values and the adjusted values is quite small too. In total, 111 values of the 2,230 imputed values were adjusted, i.e. about 5.0%. The sum (over all variables in the EPE $Y_{2,E}$ data) of the absolute differences of the means of the imputed values and the means of the adjusted imputed values is 70.6, and the sum (over all variables) of the means of the imputed values is 2,855.9. So, the "average" change due to the

adjustment procedure is about 2.5%. For only three variables, the change is larger than 10%, namely *pininvwm* (73.6%), *taxexpot* (113.7%), and *taxexpap* (135.0%).

Table 5.7. Adjustments to imputed values of the end-of-pipe investment variables in the EPE $Y_{2,E}$ data.

Variable	mean of imputed values	mean of adjusted imputed values	difference between these means	number of imputed values	number of adjusted imputed values
eopinvwp	87.48	87.62	-0.14	55	1
eopinvwm	41.64	42.30	-0.65	49	2
eopinvap	183.27	187.61	-4.34	49	3
eopinvnp	38.54	38.54	0.00	36	0
eopinvot	222.69	222.69	0.00	18	1
eopinvtot	192.75	182.29	10.46	60	4

The results for the EPE $Y_{3,E}$ data are similar to the results for the EPE $Y_{2,E}$ data. In fact, the changes due to the adjustment procedure are even smaller for the $Y_{3,E}$ data than for the $Y_{2,E}$ data. In total, 95 values of the 2,362 imputed values were adjusted, i.e. about 4.0%. The sum (over all variables in the EPE $Y_{3,E}$ data) of the absolute differences of the means of the imputed values and the means of the adjusted imputed values is 72.8, and the sum (over all variables) of the means of the imputed values is 6,336.89. So, the "average" change due to the adjustment procedure is about 1.1%. For only two variables, the change is larger than 10%, namely *pininvto* (12.6%), and *taxexpot* (37.4%).

5.3.3 Edit and imputation results for the ABI evaluation data

In Table 5.8 the error localisation results for strategies I and II are presented. The alphas are quite high for both strategies, pointing to a large proportion of undetected errors. Because less edits apply to the variables, it is evident that the alphas are higher for strategy II than for strategy I. Conversely, the betas are lower. Using less edits results in fewer correct values considered implausible by the editing process. Most deltas are similar or lower in strategy II than in strategy I, showing that the amount of misclassification is smaller with fewer edits.

Table 5.8. Error localisation results for ABI evaluation ($Y_{3,E}$) data set using strategy I (all edits) and strategy II (hard edits only).

Variable	Strategy I			Strategy II		
	alpha	beta	delta	alpha	beta	delta
turnover	0.529	0.055	0.096	0.628	0.000	0.054
emptotc	0.378	0.274	0.284	0.613	0.001	0.059
purtot	0.696	0.016	0.117	0.708	0.006	0.111
taxrates	0.585	0.004	0.027	0.654	0.002	0.027
taxothe	0.589	0.000	0.023	0.647	0.000	0.025
taxtot	0.569	0.045	0.107	0.679	0.001	0.082
stockbeg	0.599	0.002	0.059	0.636	0.001	0.062
stockend	0.589	0.002	0.059	0.636	0.001	0.062
assacq	0.630	0.001	0.049	0.662	0.000	0.050
assdisp	0.619	0.001	0.038	0.651	0.001	0.040
capwork	0.559	0.001	0.009	0.559	0.001	0.009
employ	0.678	0.133	0.159	1.000	0.000	0.048

In Table 5.9 imputation results for the ABI evaluation data are presented. These results pertain to the $Y_{3,E}$ data set with errors localised by either strategy I or strategy II and the $Y_{2,E}$ data set (missings only).

Table 5.9. Imputation results for ABI evaluation $Y_{3,E}$ and $Y_{3,D}$ data sets.

Variable	$Y_{3,E}$ data strategy I		$Y_{3,E}$ data strategy II		$Y_{2,E}$ data (no errors)	
	d_{L1}	m_1	d_{L1}	m_1	d_{L1}	m_1
turnover	428.43	169.40	74.81	55.51	126.39	60.47
emptotc	59.39	56.68	42.50	36.29	12.42	3.52
purtot	858.10	834.12	331.30	306.74	4.56	1.96
taxtot	7.92	5.94	40.25	36.17	3.41	0.58
taxrates	6.64	0.77	20.02	15.69	1.20	0.87
taxothe	6.70	5.72	52.49	46.35	0.82	0.71
assacq	36.19	29.57	33.91	27.67	115.37	105.20
assdisp	66.08	60.96	71.08	65.44	3.46	1.94
employ	3.33	0.97	2.66	2.00	4.21	1.02
stockbeg	30.36	14.01	190.97	177.21	45.82	6.07
stockend	25.90	3.13	27.56	15.89	47.16	6.96
capwork	19.40	18.06	19.40	18.06	2.69	2.59

The results show that the largest differences in imputation performance results are between the $Y_{3,E}$ data and the $Y_{2,E}$ data. With a few exceptions, the results are much better for the $Y_{2,E}$ data than for both experiments with the $Y_{3,E}$ data. An exception where the imputations for the $Y_{3,E}$ data are much better than for the $Y_{2,E}$ data occurs for *assacq*. From the results in Table 5.8 it was concluded that the error localisation strategy II performed better than strategy I. The imputation results, however, show that the difference in imputation performance between these two experiments is not

so clear cut. For three of the main variables, *turnover, emptotc* and *purtot* the imputation results are better for strategy II than for strategy I, but for *taxtot* and the components thereof (*taxrates* and *taxothe*) as well as *stockbeg* strategy I is better.

5.3.4 Edit and imputation results for the EPE evaluation data

Results for the error localisation and imputation performance for both EPE evaluation data sets ($Y_{3,E}$ and $Y_{2,E}$) are summarised in Table 5.10. With respect to the error localisation, a striking result is that the alphas are often 1, indicating that none of the errors has been correctly localised. It should be noted however, that there are only a few errors in each variable. But still, the overall error detection performance is not very good; only 11 (20%) of the 54 errors in these variables have been detected correctly.

Table 5.10. Error localisation and imputation results for EPE evaluation $Y_{3,E}$ (errors and missings) and $Y_{2,E}$ (missings only) data sets.

Variable	$Y_{3,E}$ data						$Y_{2,E}$ data	
	#errors	alpha	beta	delta	d_{L1}	m_1	d_{L1}	m_1
totinvto	12	0.83	0.003	0.014	52.14	41.47	57.46	49.99
totexpto	14	0.50	0.009	0.017	30.72	21.43	0.00	0.00
subtot	1	1.00	0.000	0.001	25.01	25.01	9.45	9.45
rectot	1	1.00	0.000	0.001	20.73	11.70	21.14	9.95
totinvwp	5	1.00	0.001	0.006	35.63	6.83	34.53	18.66
totinvwm	8	0.63	0.001	0.006	77.81	52.52	23.33	3.74
totinvap	6	1.00	0.000	0.006	40.03	30.68	40.89	36.57
totinvnp	5	1.00	0.001	0.006	15.01	12.88	16.15	10.31
totinvot	2	0.50	0.000	0.001	52.16	18.29	57.39	14.07

The imputation results (not presented here; see Pannekoek, 2003b, and Vonk, Pannekoek, and De Waal, 2003a) show that the imputation performance is more often better for the $Y_{2,E}$ data than for the $Y_{3,E}$ data. The 175 missing values for *totexpto* could all be imputed deductively, and therefore without error, for the $Y_{2,E}$ data set. But also the imputation of *totinvwm* is considerably better for the $Y_{2,E}$ data than for the $Y_{3,E}$ data.

5.4 Comparison with other approaches on the Y_3 data

In this section we compare the approach of CBS to the approaches used by the other participants in the EUREDIT project. These other approaches are sketched in Subsection 5.4.1. We will only describe the results for the $Y_{3,E}$ data sets. The results for the ABI $Y_{3,E}$ data is given in Subsection 5.4.2, and for the EPE $Y_{3,E}$ data in Subsection 5.4.3. The analysis in these latter subsections is based on the analyses carried out by Chambers and Xinqiang (2003a and 2003b). To summarise the results in a comprehensible manner, Chambers and Xinqiang have normalised and grouped the different measures, using the five groups described earlier. In particular, the

"averaged" performance across variables has been determined in order to see whether certain experiments tended to be better or worse relative to this average. To calculate this average the two-factor ANOVA model

Response = Overall Effect + Experiment Effect + Variable Effect

has been fitted, where *Response* was defined as the logarithm of the sum of a (normalised) performance measure and a small positive constant, chosen so that this logarithm was always defined. This model was then fitted (via least squares) to the values of the performance measure for a group of experiments and a group of variables, and the estimated *Experiment Effects* obtained. By definition, these estimated effects sum to zero. Negative values are better than positive ones for these effects. Several of the tables below therefore show the number of negative values (NNV) for each row in a table. The larger this value, the better the performance of an experiment relative to the variables and/or measures being analysed in that table.

In order to keep the size of the present paper limited, we do not describe the results for the $Y_{2,E}$ data sets here. For these results we refer to the analyses carried out by Chambers and Xinqiang (2003a and 2003b). Here we only note that one of the conclusions of those analyses is that the imputation approach of CBS (see Sections 4.3 and 4.4) did well. With respect to the ABI $Y_{2,E}$ data Chambers and Xinqiang (2003a) note that the best experiment (of the two) carried out by CBS and three other experiments (of the 12 experiments in total) stand out as providing good results across all situations. With respect to the EPE $Y_{2,E}$ data Chambers and Xinqiang (2003b) note that the best experiment (of the two) carried out by CBS does particularly well, and is the best overall experiment.

5.4.1 Brief description of the experiments in the EUREDIT project

Some experiments that have been carried out on the ABI and EPE $Y_{3,E}$ data concentrated exclusively on the editing part, and others exclusively on the imputation part. In the subsequent subsections we will only compare the results of the experiments that carried out both editing and imputation, however. In this section we briefly describe the methods used for these experiments. For the ABI $Y_{3,E}$ data, University of Jyväskylä, Swiss Federal Statistical Office, University of Southampton and CBS performed experiments that included both editing and imputation. For the EPE $Y_{3,E}$ data, the same institutes except University of Jyväskylä performed these kinds of experiment.

The University of Jyväskylä performed two experiments, JA30001 and JA30004, that carried out edit and imputation on the ABI $Y_{3,E}$ data. Both experiments were based on the application of a kind of neural network called a self-organising map (SOM). Application of a SOM results, among others, in a subdivision of the data records in mutually exclusive subsets which are referred to as "nodes" or "clusters" and that are in some sense as homogeneous as possible with respect to the target variables. The SOM parameters (node means and standard deviations) were estimated in a robust way. Outliers were determined on a variable by variable basis as those values that were a number of standard deviations away from the node mean.

Within nodes, imputed values were obtained by drawing from a left and right truncated normal distribution. The main difference between the two experiments is that in experiment JA30001 11 separate SOMs where used for 11 mutually exclusive subsets of the ABI variables whereas in experiment JA30004 only 5 different SOMs where used. For details on these experiments we refer to Koikkalainen (2003) and Koikkalainen, Piela and Laaksonen (2003).

The Swiss Federal Statistical Office performed 6 experiments, SA31101, SA31102, SA31501, SA31502, SA32401 and SA32402, on the ABI $Y_{3,E}$ data, and two experiments, SE30200 and SE32400 on the EPE $Y_{3,E}$ data. All these experiments were based on outlier detection methods followed by an outlier-robust imputation method. Various outlier detection methods were used: the Epidemic Algorithm (SA31101, SA31102, SE30200), the Transformed Rank Correlation method (SA32401, SA32402, SE32400), and the BACON algorithm with the EM-algorithm (SA31501, SA31502). The experiments SA31102, SA32402, and SA31502 use besides outlier detection also edit rules to identify errors. The imputation method for all experiments was a robust version of a nearest neighbour hot-deck method. With this method, a donor is selected that is "close" to the recipient in terms of a robust distance measure and also belongs to the set of non-outlying records. In addition, donor selection can also be restricted to records that do not fail hard or soft edit rules. Details on these methods can be found in Béguin and Hulliger (2003a and 2003b).

The University of Southampton performed 7 experiments, UA30001 to UA30007, on the ABI $Y_{3,E}$ data, and one experiment, UE30001 on the EPE $Y_{3,E}$ data. All these experiments were again based on outlier detection methods. The outlier detection methods used are univariate Forward Search (UA30003 to UA30006), univariate WAID which is based on outlier robust regression trees (UA30001, UA30007 and UE30001), and a multivariate version of WAID (UA30002). The WAID methods split up the data records in a number of mutually exclusive classes or "nodes" which are homogeneous with respect to the target variable(s). Outlying values within nodes are seen as errors and are subsequently imputed by a robust estimate of the node mean. The outliers defined by the Forward Search method were imputed using different robust regression models. For details on these methods we refer to Hentges (2003a and 2003b), Ren and Chambers (2003) and Xinqiang and Chambers (2003).

Finally, CBS used the approach described in Section 4. For the ABI $Y_{3,E}$ data we have carried out an experiment using all edits, experiment CA30001, and an experiment using only the hard edits, experiment CA30002. For the EPE $Y_{3,E}$ data we have carried out only one experiment, CE30001.

5.4.2 Comparison for the ABI $Y_{3,E}$ data

In total 33 experiments have been carried out on the ABI $Y_{3,E}$ data. However, 10 experiments only concentrated on editing, and 6 only on imputation. The remaining 17 experiments carried out both editing and imputation. We examine the results of these latter experiments below. In Table 5.11 we show the results for the six key

economic variables (*turnover*, *emptotc*, *purtot*, *taxtot*, *assacq*, and *assdisp*) on the grouped measures described in Subsection 5.1.

The results in Table 5.11 indicate that the univariate Forward Search experiments show the best overall performance. The univariate WAID experiments UA30001 and UA30007, the SOM experiment JA30001 and the CBS experiment with hard edits only, also perform above average on all 5 criteria. Inspection of the alpha and beta values underlying the results on the pure error localisation criterion *a* in Table 5.11 shows that the Fellegi-Holt based error localisation in the CBS experiment CA30002 does not find as many of the true errors as the other experiments but at the same time the proportion of correct values that is considered erroneous is smaller than for the other experiments.

Table 5.11. Experiment Effects for grouped performance measures in the ABI $Y_{3,E}$ data. Effects were defined relative to all six key ABI economic variables for groups (a) to (c) and relative to the variables turnover*,* emptotc*,* purtot *and* taxtot *for groups (d) and (e). Bold font indicates best performer in each column.*

Experiment	NNV	(a)	(b)	(c)	(d)	(e)
CA30001	4	0.4413	-0.3009	-0.3678	-0.8070	-0.5114
CA30002	5	-0.5530	-0.0687	-0.3422	-0.8372	-0.5638
JA30001	5	-0.5227	-0.2114	-0.3675	-0.2494	-0.4458
JA30004	4	-0.1015	-0.1544	-0.3529	0.3292	-0.5034
SA31101	1	-0.4263	3.5850	3.6945	0.6101	0.9951
SA31102	4	2.0412	-0.1486	-0.3098	-0.1249	-0.3022
SA31501	1	-0.4581	0.3105	0.2371	0.5227	2.5957
SA31502	2	2.0056	**-0.5587**	-0.3814	2.2404	0.1383
SA32401	1	-0.3946	0.2981	0.2371	0.5635	2.2492
SA32402	2	2.0223	-0.5491	-0.3811	2.2224	0.1182
UA30001	5	-0.6225	-0.3328	-0.3764	-0.7382	-0.5701
UA30002	2	-0.4737	0.7934	0.6604	0.0354	-0.3950
UA30003	5	**-0.6290**	-0.5480	**-0.3991**	-0.8431	-0.5684
UA30004	5	**-0.6290**	-0.5480	**-0.3991**	-0.8419	-0.5654
UA30005	5	**-0.6290**	-0.5480	**-0.3991**	**-0.8450**	-0.5446
UA30006	5	**-0.6290**	-0.5480	**-0.3991**	-0.8366	**-0.5945**
UA30007	5	-0.4421	-0.4702	-0.3534	-0.4003	-0.5320

In Table 5.12 we show the results for the secondary variables (*stockbeg, stockend* and *capwork*). Table 5.12 contains fewer experiments than Table 5.11, because many experiments were not carried out for the secondary variables (and the component variables; see Table 5.13). All performed experiments are reported in Table 5.12.

Table 5.12 shows that the error localisation performance (criteria *a* and *b*) of the CBS experiments is the best of these six experiments. Since there were no hard edits for the three variables to which these results apply, the only error localisation

method used in experiment CA30002 was the detection of thousand-errors. This simple method was apparently quite successful. Using soft edits improves the error localisation for these variables whereas for the variables in Table 5.11 the opposite was true. The imputation results on criteria *d* and *e* show that experiment UA30007 stands out and there is little difference between the other experiments.

*Table 5.12. Experiment Effects for grouped performance measures in the ABI $Y_{3,E}$ data. Effects were defined relative to all three secondary variables (*stockbeg, stockend *and* capwork*) for groups (a) to (c) and relative to the variables* stockbeg *and* stockend *for groups (d) and (e). Bold font indicates best performer in each column.*

Experiment	NNV	(a)	(b)	(c)	(d)	(e)
CA30001	3	-0.8863	**-1.1967**	**-0.4480**	0.3944	0.3955
CA30002	3	**-0.9015**	-0.4252	-0.4411	0.3991	0.4065
JA30001	3	-0.4159	-0.2965	-0.2627	0.4122	0.4433
JA30004	2	1.6254	-0.1063	-0.4415	0.4373	0.3926
UA30001	2	-0.1688	0.2255	-0.4426	0.3980	0.4030
UA30007	2	0.7472	1.7992	2.0359	**-2.0410**	**-2.0409**

Finally, in Table 5.13 we present results for the component variables of *taxtot* (*taxrates* and *taxothe*). Results for the component variables of *emptotc* and *purtot* are not given here, but can be found in Chambers and Xinqiang (2003a).

Table 5.13. Experiment Effects for grouped performance measures in the ABI $Y_{3,E}$ data. Effects were defined relative to the two component variables of taxtot *(*taxrates *and* taxothe*) for groups (a) to (c) and relative to* taxrates *for groups (d) and (e). Bold font indicates best performer in each column.*

Experiment	NNV	(a)	(b)	(c)	(d)	(e)
CA30001	4	**-0.9870**	1.2847	**-0.8552**	**-0.5835**	-0.3650
CA30002	4	-0.9299	1.2963	-0.7354	-0.5495	-0.4076
JA30001	2	1.4451	-0.5929	1.5793	-0.0175	2.0397
JA30004	2	0.9728	**-0.6648**	0.9052	1.9992	-0.3789
UA30001	5	-0.2278	-0.6611	-0.5324	-0.4383	**-0.4792**
UA30007	5	-0.2733	-0.6622	-0.3615	-0.4105	-0.4090

From Table 5.13 we see that the two WAID experiments (UA30001 and UA30007) are performing above average on all 5 criteria whereas the CBS experiments are performing above average on only 4 of them. The Fellegi-Holt based error localisation method of the CBS experiments is the best performing method (with respect to criterion *a*). It could be expected that a Fellegi-Holt based error

localisation method performs relatively better for *taxrates* and *taxothe* than for the six key economic variables (see Table 5.11), since a simple edit rule like *taxtot* = *taxrates* + *taxothe* leaves less uncertainty about which variable is in error than edit rules involving larger numbers of variables. With respect to the significant error finding criterion *b,* the CBS experiments were the worst performing.

From Tables 5.11 to 5.13 we can conclude that experiment CA30002, our approach where we have used only the hard edits, had a solid performance. It is, however, slightly outperformed by the experiments UA30001 and UA30007. This is confirmed by other evaluation results (cf. Chambers and Xinqiang, 2003a). The performance of the Fellegi-Holt based error localisation of experiment CA30002 on the pure error localisation criterion *a* is not far behind or even better than the best outlier based methods (see Tables 5.11 to 5.13). The Fellegi-Holt based method seems to be more conservative than the outlier based methods in the sense that fewer errors are detected but also fewer correct values are identified as errors. The outlier based methods score (much) better on the significant error finding criterion *b*, meaning that these methods find relatively more large errors than the Fellegi-Holt based method. This comes as no surprise because methods that rely on outlier detection for error localisation are specifically set up to find large errors. Our multivariate regression imputation procedure (CA30001 and CA30002) performed well for the six key economic variables (see Table 5.11) although the robust regression imputation of experiments UA30003-UA30006 performed a little bit better. The nearest neighbour hot-deck method of the experiments SA31101-SA32402 did not perform as well as the other methods for these variables. Our hot-deck (see Table 5.12) and ratio hot-deck (see Table 5.13) imputation methods performed well for the secondary and component variables, respectively, but the robust node mean imputation methods of experiments UA30001 and UA30007 did about as well or better.

From their analysis of the evaluation results Chambers and Xinqiang (2003a) conclude: "the ABI-based EUREDIT experiments indicate that the outlier robust regression tree-based automatic editing and imputation procedures underpinning the experiments UA30001 and UA30007 are worth developing further as an editing and imputation tool for business survey data. In addition, the linear model-based methods used in the experiments UA30003 – UA30006 and CA30002 and the SOM-based methods used in the experiment JA30001 also performed well and are well worth consideration when setting up an automatic editing and imputation tool for business survey data that has more 'linear structure' (perhaps after transformation)".

5.4.3 Comparison for the EPE $Y_{3,E}$ data

The EPE $Y_{3,E}$ data are very complex; they contain many zeros and have to satisfy complex edits. Therefore, only 7 experiments have been carried out on the EPE $Y_{3,E}$ data. Furthermore, only 4 of those experiments carried out both editing and imputation. We examine the results of these latter experiments below. In Table 5.14 we show the results for the four total variables (*totinvto, totexpto*, *subtot*, and *rectot*).

Table 5.14. Experiment Effects for grouped performance measures in the EPE $Y_{3,E}$ data. Effects were defined relative to all four total variables for groups (a) to (c) and relative to the variables totinvto *and* totexpto *for groups (d) and (e). Bold font indicates best performer in each column.*

Experiment	NNV	(a)	(b)	(c)	(d)	(e)
CE30001	3	**-1.3000**	1.0465	0.8587	**-1.3061**	**-1.2837**
SE30200	1	0.6366	**-0.8713**	0.1906	1.0051	1.0580
SE32400	3	0.9168	-0.8393	**-1.4400**	-0.1948	0.4347
UE30001	2	-0.2534	0.6642	0.3907	0.4958	-0.2090

From Table 5.14 we see that the experiments CE30001 and SE32400 perform better overall than the experiments SE30200 and UE30001. In Tables 5.15 and 5.16 we show the results for the component variables of *totinvto*, respectively *totexpto*.

Table 5.15. Experiment Effects for grouped performance measures in the EPE $Y_{3,E}$ data. Effects were defined relative to all five component variables of totinvto *(*totinvwp*,* totinvwm*,* totinvap*,* totinvnp*, and* totinvot*). Bold font indicates best performer in each column.*

Experiment	NNV	(a)	(b)	(c)	(d)	(e)
CE30001	2	0.2776	1.1468	0.5177	**-1.2705**	**-1.3764**
SE30200	2	**-1.4435**	-0.7607	0.3637	1.0150	0.6516
SE32400	2	0.3005	0.5251	**-1.4922**	-0.2742	0.8198
UE30001	2	0.8654	**-0.9113**	0.6108	0.5296	-0.0950

Table 5.16. Experiment effects for grouped performance measures in the EPE $Y_{3,E}$ data. Effects were defined relative to all five component variables of totexpto *(*totexpwp*,* totexpwm*,* totexpap*,* totexpnp *and* totexpot*). Bold font indicates best performer in each column.*

Experiment	NNV	(a)	(b)	(c)	(d)	(e)
CE30001	4	0.3421	-0.8462	**-1.0503**	**-1.0884**	**-0.8985**
SE30200	1	**-1.354**	0.7638	1.1119	1.2036	1.4078
SE32400	4	-0.0124	0.9624	-0.6046	-0.4819	-0.4561
UE30001	2	1.0244	**-0.8799**	0.543	0.3668	-0.0531

From Table 5.15 we conclude that all four experiments perform comparably with respect to the editing and imputation for the components of *totinvto*, and from Table 5.16 that CE30001 and SE32400 perform best for the components of *totexpto*.

For the EPE $Y_{3,E}$ data, the Fellegi-Holt based error localisation method (our experiment CE30001) performed very well (with respect to the pure error localisation criterion *a*) for the four overall total variables (see Table 5.14) but not

well for the component variables (see Tables 5.15 and 5.16). With respect to the localisation of significant errors (criterion *b*), the Fellegi-Holt method did not perform well for the overall total variables and the component variables of *totinvto*. For the component variables, the experiment SE30200 performed best with respect to criterion *a*, whereas the experiment UE30001 found the largest proportion of significant errors (see Tables 5.15 and 5.16). The imputation methods used by our experiment (multivariate regression for the four total variables and ratio hot-deck for the component variables) performed better than the robust node mean imputation of experiment UE30001 and the robust nearest neighbour hot-deck method of experiments SE30200 and SE32400.

From their analysis of the evaluation results Chambers and Xinqiang (2003b) conclude: "Of the four experiments [involving both editing and imputation[1]] that were carried out, it is clear from the analysis that CE30001 is the overall leader, followed by SE32400, then UE30001 and then SE30200. With the Y_2 version of this data set, the experiment CE20001 stands out. Both CE30001 and CE20001 are made up of an "expertly defined" mix of editing and imputation methods, and the general conclusion therefore seems to be that for data like the EPE data, no one particular type of editing and/or imputation strategy will be effective, and best results are obtained by applying several different methods simultaneously".

6. Conclusions

From the analyses carried out by Chambers and Xinqiang (2003a and 2003b), which are summarised in Section 5.4, we conclude that the approach used by CBS performed well in comparison with the other methods. Not only did our approach rank among the best for each data set, it could also be applied to edit and impute both the ABI and EPE data. In other words, our approach is sufficiently general to be able to process various kinds of business data set. Another strong point of our approach is that it leads to data that satisfy the specified edits. Other approaches that lead to acceptable results for either the ABI or the EPE data, do not guarantee that edits are satisfied by the edited and imputed data sets. Finally, our approach is a very flexible one. Individual steps, such as the detection of obvious errors and the imputation of erroneous and missing values, can, if desired, separately be modified without having to change the other steps in the approach. Furthermore, additional steps can easily be added. For instance, the experiments on the ABI data indicate that for these kinds of data, it is useful to identify outliers and impute them by means of an outlier-robust method (see the results of the experiments UA30001 and UA30007 in Section 5.4). Such an outlier detection step can, for instance, be added to our approach immediately after the detection of obvious errors. The imputation method we have applied can be replaced by outlier-robust versions of the regression and hot-deck methods.

[1] Insertion by the authors of the present paper

Despite the above-mentioned strong points of our approach, we are aware that automatic editing and imputation is a potentially dangerous approach. Our methodology correctly identifies only a low fraction of the errors in the observed data. Moreover, although the imputation performance of our methodology is good for the $Y_{2,E}$ data sets, it is less good for the $Y_{3,E}$ data sets. This leads us to the conclusion that the edit and imputation process should not be fully automated in practice.

We advocate an edit and imputation approach that consists of the following steps:

- correction of obvious mistakes, such as thousand-errors;
- application of selective editing to split the records in a critical stream and a non-critical stream (see Lawrence and McKenzie, 2000, and Hedlin, 2003);
- editing of the data: the records in the critical stream are edited interactively, the record in the non-critical stream are edited and imputed automatically;
- validation of the publication figures by means of (graphical) macro-editing.

The above steps are used at CBS in the production process for structural annual business surveys (see De Jong, 2002). At CBS, so-called plausibility indicators to split the records in a critical stream and a non-critical stream are applied. We refer to Hoogland (2002) for a description of these indicators. Very unreliable or highly influential records lead to a low score on the plausibility indicators. Such records are edited interactively. For some results of the combined use of selective editing and automatic editing on CBS business surveys, we refer to Hoogland and Van der Pijll (2003). The final validation step is performed by statistical analysts, who compare the publication figures based on the edited and imputed data to publication figures from a previous year, for instance. Influential errors that were not corrected during automatic (or interactive) editing can be detected during this final, important, step, which helps to ensure the quality of our data.

We feel that only a combined approach using selective editing, interactive editing, automatic editing, and macro-editing can improve the efficiency of the traditional interactive edit and imputation process, while at the same time maintaining or even enhancing the statistical quality of the produced data. To some extent our intuition is confirmed by our experiences in the EUREDIT project where our approach to automatic edit and imputation, a mix of several different methods for automatic edit and imputation, lead to good results in comparison to the other methods.

References

Béguin, C. and B. Hulliger (2003a), Multivariate Outlier Detection in Incomplete Survey Data: The BEM Algorithm. In: *Methods and Experimental Results from the EUREDIT Project*, (ed. J.R.H. Charlton) (forthcoming volume).

Béguin, C. and B. Hulliger (2003b), Robust Multivariate Outlier Detection and Imputation with Incomplete Data. In: *Methods and Experimental Results from the EUREDIT Project* (ed. J.R.H. Charlton) (forthcoming volume).

Chambers, R. (2003), Evaluation Criteria for Statistical Editing and Imputation. In: *Methods and Experimental Results from the EUREDIT Project* (ed. J.R.H. Charlton) (forthcoming volume).

Chambers, R. and Z. Xinqiang (2003a), Evaluation of Edit and Imputation Methods Applied to the UK Annual Business Inquiry. In: *Towards Effective Statistical Editing and Imputation Strategies – Findings of the EUREDIT Project* (ed. J.R.H. Charlton) (forthcoming volume).

Chambers, R. and Z. Xinqiang (2003b), Evaluation of Edit and Imputation Methods Applied to the Swiss Environmental Protection Expenditure Survey. In: *Towards Effective Statistical Editing and Imputation Strategies – Findings of the EUREDIT Project* (ed. J.R.H. Charlton) (forthcoming volume).

Chvátal, V. (1983), *Linear Programming*. W.H. Freeman and Company, New York.

De Jong, A. (2002). *Unit-Edit: Standardized Processing of Structural Business Statistics in the Netherlands*. UN/ECE Work Session on Statistical Data Editing, Helsinki.

De Waal, T. (2003), *Processing of Erroneous and Unsafe Data*. Ph.D. Thesis, Erasmus University, Rotterdam.

De Waal, T. and R. Quere (2003), A Fast and Simple Algorithm for Automatic Editing of Mixed Data. Submitted to *Journal of Official Statistics*.

Di Zio, M., U. Guarnera, O. Luzi (2003), Application of GEIS to the UK ABI Data: Editing. In: *Methods and Experimental Results from the EUREDIT Project* (ed. J.R.H. Charlton) (forthcoming volume).

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

Hedlin, D. (2003), Score Functions to Reduce Business Survey Editing at the U.K. Office for National Statistics, *Journal of Official Statistics 19*, pp. 177-199.

Hentges, A. (2003a), Robust Multivariate Outlier Detection Based on Forward Search Methods. In: *Methods and Experimental Results from the EUREDIT Project* (ed. J.R.H. Charlton) (forthcoming volume).

Hentges, A. (2003b), Robust Multivariate Outlier Detection via Forward Search: Application to the ABI Data Set. In: *Methods and Experimental Results from the EUREDIT Project* (ed. J.R.H. Charlton) (forthcoming volume).

Hoogland, J. (2002), Selective Editing by Means of Plausibility Indicators. UN/ECE Work Session on Statistical Data Editing, Helsinki.

Hoogland, J. and E. Van der Pijll (2003), *Summary of the Evaluation of Automatic versus Manual Editing of the Production Statistics 2000 Trade and Transport*. UN/ECE Work Session on Statistical Data Editing, Madrid.

Koikkalainen, P. (2003), Description of the Error Localisation Methodology Based on the Tree Structured Self-Organising Map. In: *Methods and Experimental*

Results from the EUREDIT Project (ed. J.R.H. Charlton) (forthcoming volume).

Koikkalainen, P., P. Piela and S. Laaksonen (2003), Description of the Imputation Methodology Based on the Tree Structured Self-Organising Map. In: *Methods and Experimental Results from the EUREDIT Project* (ed. J.R.H. Charlton) (forthcoming volume).

Lawrence, D. and R. McKenzie (2000), The General Application of Significance Editing. *Journal of Official Statistics 16*, pp. 243-253.

Little, R.J.A. and P.J. Smith (1987), Editing and Imputation of Quantitative Survey Data. *Journal of the American Statistical Association 82*, pp. 58-68.

Little, R.J.A. and D.B. Rubin (1987), *Statistical Analysis with Missing Data.* Wiley, New York.

Pannekoek, J. (2003a), (Multivariate) Regression and Hot Deck Imputation Methods. In: *Methods and Experimental Results from the EUREDIT Project* (ed. J.R.H. Charlton) (forthcoming volume).

Pannekoek, J. (2003b), Imputation Using Standard Methods: Evaluation of (Multivariate) Regression and Hot Deck Methods. In: *Methods and Experimental Results from the EUREDIT Project* (ed. J.R.H. Charlton) (forthcoming volume).

Pannekoek, J. and M.G.P. Van Veller (2003), Regression and Hot Deck Imputation Strategies for Continuous and Semi-Continuous Variables. In: *Methods and Experimental Results from the EUREDIT Project* (ed. J.R.H. Charlton) (forthcoming volume).

Ren, R. and R. Chambers (2003), Outlier Robust Methods: Outlier Robust Estimation and Outlier Robust Imputation by Reverse Calibration. In: *Methods and Experimental Results from the EUREDIT Project* (ed. J.R.H. Charlton) (forthcoming volume).

Schafer, J.L. (1997), *Analysis of Incomplete Multivariate Data*. Chapman & Hall, London.

Vonk, M., J. Pannekoek and T. De Waal (2003a), Edit and Imputation Using Standard Methods: Evaluation of the (Automatic) Error Localisation Strategy for the ABI and EPE Data Sets. In: *Methods and Experimental Results from the EUREDIT Project* (ed. J.R.H. Charlton) (forthcoming volume).

Vonk, M., J. Pannekoek and T. De Waal (2003b), *Development of (Automatic) Error Localisation Strategy for the ABI and EPE Data*. Report (research paper 0302), Statistics Netherlands, Voorburg.

Xinqiang, Z. and R. Chambers (2003), Outlier Identification and Imputation Using Robust Regression Trees. In: *Methods and Experimental Results from the EUREDIT Project* (ed. J.R.H. Charlton) (forthcoming volume).