Journal of Official Statistics, Vol. 21, No. 2, 2005, pp. 257–286

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

Jeroen Pannekoek and Ton de Waal[1]

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 article we describe the general approach applied by Statistics Netherlands on the two business surveys used in the EUREDIT project. In the EUREDIT project data for only one year were available. In our edit and imputation methods we therefore could not use data from a previous year and had to restrict ourselves to using only data from the data set to be edited and imputed itself. 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.

Key words: Automatic editing; consistency; deductive imputation; error localisation; hot-deck imputation; Fellegi-Holt paradigm; multivariate regression imputation; nearest neighbour hot-deck imputation; random errors; ratio hot-deck imputation; systematic errors.

1. Introduction

High-quality and timely statistical information on many different aspects of society is a prerequisite for policy-makers in order to make well-informed decisions. National statistical institutes (NSIs) fulfil a prominent role in providing such statistical information. The successful fulfilment of their role is complicated by the fact that data collected by NSIs generally contain errors. In particular the data collection phase is a potential source of error. Errors may have been made by the respondent, who may make errors by mistake or deliberately while filling in the questionnaire, may misunderstand a certain question, or may not know the correct answer to a certain question. Errors may also have been made at the NSI, for instance while transferring the data from the questionnaire to the computer system. In order to be able to publish reliable statistical information the errors in the collected data have to be corrected. This correction process is referred to as statistical data editing. Traditionally, each received questionnaire was checked for errors. Subsequently, subject-matter specialists corrected the detected errors by using their expert knowledge, or by contacting the supplier of the information. This form of statistical data editing, called

[1] Statistics Netherlands, PO Box 4000, 2270 JM Voorburg, The Netherlands. Emails: jpnk@cbs.nl and twal@cbs.nl

Acknowledgment: The research described in the article was carried out as part of the EUREDIT project, which was funded by the European Commission under the Fifth Framework, contract IST-1999-10226. The views expressed in this article are those of the authors and do not necessarily reflect the policies of Statistics Netherlands.

© Statistics Sweden

manual (or interactive) editing, leads to statistical data of good quality, but is very costly and time consuming. Several studies (see, e.g., Granquist 1995, 1997; Granquist and Kovar 1997) have shown that in order to obtain reliable publication figures only the most influential errors have to be edited manually. This is an important observation that allows one to improve the efficiency of the statistical data editing process.

The EUREDIT project (see http://www.cs.york.ac.uk/euredit) was a large international research and development project aimed at improving the efficiency and the quality of the statistical data editing and imputation process at NSIs. It involved 12 institutes from seven countries. Six of those institutes were NSIs, namely the UK Office for National Statistics (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.

For CBS, the main aims of the project were:

- 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;
- To compare all methods tested and develop a strategy for users of edit and imputation leading to a "best practice guide."

The EUREDIT project concentrated on automatic methods for editing and imputation. Other editing methods, such as selective editing (cf. Lawrence and McDavitt 1994; Lawrence and McKenzie 2000; Hedlin 2003) where part of the data are edited manually, were examined only to a very limited extent.

It has been argued that the role of statistical data editing should be broader than only error localisation and correction (cf. Granquist 1995; Granquist, 1997; Granquist and Kovar, 1997; Bethlehem and van de Pol 1998). We fully agree with this point of view, and, for instance, consider the feedback provided by the edit process on the questionnaire design at least as important as error localisation and correction. However, within the EUREDIT project the role of editing was strictly limited to error localisation and correction, and in the present article we will therefore concentrate on this.

The article 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 from 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 involved. This comparison is to a substantial extent based on evaluation studies performed by Chambers and Zhao (2004a and 2004b) in the EUREDIT project.

Although the methods and tools we consider in this article are automatic ones, they require quite a bit of expert knowledge and statistical analysis to set up. In practice, however, the tools have to be set up only once. In future versions of a particular survey, one only needs to update the parameters of the various methods. This updating process can to a substantial extent be automated. So, in future versions of a survey, preparing and using

the methods and tools we consider in this article are almost fully automated, much more than when a survey is conducted for the first time.

In practice, to edit and impute a data set, one often uses corresponding cleaned data from a previous year. In the EUREDIT project, however, data from only one year were available. In our edit and imputation methods we therefore had to restrict ourselves to using only data from the data set to be edited and imputed itself. Our general strategy can in a natural way be extended to the case where cleaned data from a previous year are available.

In the literature, there is quite a scarcity of articles on the combined application of editing and imputation techniques in practice. The only articles similar to the present article we are aware of are the ones by Little and Smith (1987) and Ghosh-Dastidar and Schafer (2003), both published in the *Journal of the American Statistical Association*. The former article focuses on outlier detection and outlier robust imputation techniques for a relatively small and simple survey, and the latter one on outlier detection and multiple imputation based on a regression model. The present article focuses on automatic editing and imputation techniques for two surveys that are considerably more complex than the ones considered by the aforementioned authors. Moreover, whereas the edit and imputation techniques applied by these authors do not ensure internal consistency of individual records, such as component variables summing up to a total, our procedures ensure such consistency.

The remainder of this article is organised as follows. Section 2 describes how the evaluation experiments were carried out within the EUREDIT project. The two data sets we consider in this article 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 provides some conclusions.

2. The Evaluation Experiments

For each data set used in the EUREDIT project six different versions were, in principle, constructed: three evaluation data sets and three development data sets. These six data sets are given in Table 2.1.

The evaluation data sets were used to evaluate the edit and imputation procedures applied by the participants in EUREDIT. All three versions of the development data, i.e., including the "true" data, were sent to all participants. For instance, the development data set could be used to train neural networks or to parameterise statistical methods.

Table 2.1. The six versions of each data set

Type	"true"	with missing data	with missing data and errors
Evaluation	Y_E^*	$Y_{2,E}$	$Y_{3,E}$
Development	Y_D^*	$Y_{2,D}$	$Y_{3,D}$

The development data represent the fact that in a real-life situation one can learn from past experience.

A Y^* data set contains "true" values, the corresponding Y_2 data set the data with missing values but with no errors, and the corresponding Y_3 data set the data with both missing values and errors. 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. The records and information in the development data sets differed from the records and information in the evaluation data sets. Constructing a data set with only errors but no missing values, a Y_1 data set, was considered to be too unrealistic a scenario. A Y_2 data set allows one to evaluate imputation methods, a Y_3 data set allows one to evaluate a combination of editing and imputation methods.

The $Y_{2,E}$ data and the $Y_{3,E}$ data were sent to all participants in the EUREDIT project. These participants then applied their methods to these data sets. The $Y_{2,E}$ data only had to be imputed. The $Y_{3,E}$ data had to be both edited and imputed. The "true" evaluation 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.

In the ideal situation, one would have a data set with true values, a corresponding data set with actual missing values without errors, and a data set with actual missing values and actually observed errors. This would allow one to evaluate edit and imputation methods by comparing edited and imputed data sets to the true data. Unfortunately, data sets with true values are very rare. In the EUREDIT project, data sets with true values were not available. Out of necessity, we defined the "true" data as the data that the provider of the data set considered to be satisfactorily cleaned according to their edit and imputation procedures. The errors in the Y_3 data are not actual errors; neither are the missing values in the Y_2 and Y_3 data the actual missing values. These missing values and errors were synthetically introduced in the corresponding Y^* data set by the coordinator of the EUREDIT project. In this way the mechanisms that generated the missing values and the errors were fully controlled by the coordinator, while remaining unknown to the participants in the EUREDIT project. By the full control of the coordinator over the error generating mechanism and the missing data mechanism, it was possible to ensure that the Y_2 and Y_3 data sets provided sufficient challenges to the participants, while at the same time remaining as realistic as possible. The fact that the error generating mechanism and the missing data generating mechanism were unknown to the participants mimics reality, where these mechanisms are also unknown to the NSIs.

Along with the data sets sent to the participants, the $Y_{2,E}$ data, the $Y_{3,E}$ data and the three development data sets, metadata related to these data sets, such as edit rules (or *edits* for short) and data dictionaries were delivered to the participants. In general, 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 large fraction of correctly observed records, but not necessarily for all correctly observed records. Examples of hard edits are rules stating that certain variables should attain nonnegative values (nonnegativity edits), and edits stating that certain variables should sum up to an observed total (balance edits). Examples of soft edits are rules stating that the ratio of two variables is generally smaller than a specified maximum (ratio edits), and rules stating that

the value of a variable is generally lower or higher than a specific upper bound or lower than a specific lower bound (upper/lower bound edits).

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 general ability to identify errors, to identify 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 (2004) for more details regarding the evaluation criteria.

3. The Data Sets

3.1. UK annual business inquiry

The UK Annual Business Inquiry (ABI) is an annual business survey containing commonly measured continuous variables such as turnover and wages. 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). Three variables, *class* (anonymised industrial classification), *turnreg* (registered turnover) and *empreg* (registered employment size group), were not obtained from the questionnaires but from completely observed registers. These variables could be used to construct suitable imputation strata, for instance. In the long questionnaire 26 variables contained errors or had missing values, and in the short questionnaire only 11 variables. The names and brief descriptions of the main variables are given in Table 3.1.

The variables in the ABI data set 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 the main employment variable *employ*. The six key ABI economic

Table 3.1. The main variables in the ABI data set

Name	Description
turnover	Total turnover
emptotc	Total employment costs
purtot	Total purchases of goods and services
taxtot	Total taxes paid
assacq	Total cost of all capital assets acquired
assdisp	Total proceeds from capital asset disposal
employ	Total number of employees
stockbeg	Value of stocks held at beginning of year
stockend	Value of stocks held at end of year
capwork	Value of work of a capital nature

variables have distributions that are highly skewed. The second level consists of the secondary variables *stockbeg*, *stockend* and *capwork* measuring business activity. For the long questionnaire, 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 questionnaire, the third level consists of two component variables for *purtot*, but no components for the other key economic variables.

For the ABI data both hard and soft edits were provided. In total 24 hard edits are specified for the ABI data: 20 nonnegativity edits, and four balance edits. Some of these hard edits are only applicable for the long questionnaire, some others only for the short questionnaire, and the rest for both types of questionnaire. In total 25 soft edits are specified for the ABI data: 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

$$\text{if } \mathit{employ} > 0 \text{ then } \mathit{emptotc}/\mathit{employ} \geq 4.$$

The edit is satisfied if *employ* is not larger than zero, irrespective of the value of *emptotc*. Both the value of *employ* and the value of *emptotc* 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 consist 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. There are four main groups of financial variables (in thousands of Swiss Francs): variables related to investments, expenditures, subsidies and income. The nomenclature of the variables in these four groups follows a logical structure. The last two letters indicate which aspect of environmental protection the variable refers to. The letters *wp* indicate "water protection," *wm* "waste treatment," *ap* "air protection," *np* "noise protection," *ot* "other," and *tot* (or *to*) "(sub)total." Variables related to subsidies begin with *sub*, variables related to income with *rec* (abbreviation for "receipts"). Variables related to investments and expenditures start with two blocks of three letters each. If the last block of three letters is *inv*, the variable refers to investments. If the last block of three letters is *exp*, the variable refers to expenditures. The first block of three letters subdivides the variable further: *eop* indicates "end-of-pipe," *pin* "process-integrated," *oth* "other," and *tot* "(sub)total." For instance, *eopinvwp* indicates the end-of-pipe investments with respect to water protection, and *eopinvtot* the total end-of-pipe investments. Tables 3.2 and 3.3 below will further clarify the nomenclature of the variables.

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 have distributions that are highly skewed. The second level consists of 20 component variables corresponding to these four total

Table 3.2. 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 3.3. 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

variables, namely the components of *totinvto* (*totinvwp*, *totinvwm*, *totinvap*, *totinvnp*, *totinvot*), the components of *totexpto* (*totexpwp*, *totexpwm*, *totexpap*, *totexpnp*, and *totexpot*), the 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 nonnegativity edits and 23 balance edits. Four of the balance edits can be deleted as they are logically implied by the other balance edits. So there are 19 nonredundant 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 3.2 and 3.3. For each table, a column with component variables has to add up to a subtotal variable. For instance, in Table 3.2 *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 3.2 *eopinvwp*, *eopinvwm*, *eopinvap*, *eopinvnp* and *eopinvot* have to add 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 3.2). This is indicated in the tables by C (sum of column totals; *e.g.,* edit (ix) in Table 3.2), R (sum of row totals; *e.g.,* edit (x)) and T (sum of component variables; *e.g.,* edit (xi)). The two one-dimensional tables state that the components of *subtot* have to add up to *subtot* and that the components of *rectot* have to add up to *rectot*.

4. Applied Methodology

4.1. Overview

In this section a number of "standard" edit and imputation methods that were applied by CBS to the ABI and EPE data are briefly described. We have subdivided the general 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 third problem, algorithms and prototype software developed at CBS have been applied and then 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.

For academic statisticians the emphasis on consistent data, i.e., our wish to let the data satisfy all specified edits, may be difficult to understand. Statistically speaking there is indeed hardly any reason to let a data set satisfy all edits, other than the *hope* that enforcing internal consistency will result in data of higher statistical quality. NSIs, however, have the

responsibility to supply data for many different academic and nonacademic users in society. For the majority of these users, inconsistent data are incomprehensible. They may reject the data as being an invalid source of information or make adjustments themselves. This hampers the unifying role of an NSI in providing data that are undisputed by different parties such as policy makers in government, opposition, trade unions, employer organisations, etc.

In principle, one could ensure consistency between publication figures during the estimation phase rather than during the edit phase, just as one could treat missing values during the estimation phase rather than during an imputation phase. Ensuring consistency during the estimation phase would, however, lead to an extremely complicated estimation problem. For simplicity, we therefore ensure consistency during the edit phase, just as we treat missing values in an imputation phase.

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 different treatments.

4.2.1. Finding systematic errors

A systematic error is an error reported consistently among (some of the) responding units. It can, for instance, be caused by a consistent misunderstanding of a survey question by (some of) the respondents. Examples are when gross values are reported instead of net values, and 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 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 possibile. Using the ABI development data, it appeared that a considerable number 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, random errors. Random errors are not caused by a systematic deficiency, but by accident. An example might be an observed value where a respondent by mistake typed one extra digit. To identify such random 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 nonnegative weight, the so-called

reliability weight, is assigned that indicates the reliability of the values of this variable. The larger 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). To determine all optimal (in the sense of the Fellegi-Holt paradigm) solutions to the error localisation problem, this algorithm generates a binary tree. In each node of this tree a branching variable is selected. After selection of a variable two branches are constructed. In one branch it is assumed that the original value of the selected variable is correct. The original value of this selected variable is filled in into the current set of edit rules. In this way we obtain a set of edit rules for a new node in the tree. In the other branch it is assumed that the original value of the selected variable is incorrect. We eliminate the selected variable from the set of current edit rules to obtain a set of edit rules for a new node in the tree (for more details on the elimination method see de Waal and Quere 2003). The resulting set of edit rules for the new node should be satisfied by the remaining variables. After all variables have been selected, we have reached a terminal node of the tree and we are left with a set of relations involving no unknowns. If and only if this set of relations contains no selfcontradicting ones, the variables that have been eliminated in order to reach the corresponding terminal node of the tree can be imputed consistently, i.e., so that all original edits can be satisfied (cf. Theorems 1 and 2 in de Waal and Quere 2003). We check for each terminal node of the tree whether the variables that have been eliminated in order to reach this node can be imputed consistently. Of all the sets of variables that can be imputed consistently we select the ones with the lowest sum of reliability weights. In this way we find all optimal solutions to the error localisation problem (cf. Theorem 3 in de Waal and Quere 2003).

In the EUREDIT project we have applied a prototype of a program called *Cherry Pie* that is based on the algorithm described above. This program has now evolved into a production version. 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 all errors in each record.

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. Each solution consists of a set of suspicious observed values. To select one of these solutions we have used a relatively simple approach. The general idea is to determine the most suspicious set of observed values. To this end we first calculate a crude prediction for all the variables in the 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 each of the solutions for that record. The optimal solution returned by *Cherry Pie* for which this distance is maximal is the one involving the variables that deviate most from their predicted values. The variables in this maximal distance solution are, in some sense, the variables with the most outlying values, and these values are hence considered to be the erroneous ones. Thus, we use error localisation by outlier detection as a means to single out one of the several solutions to the Fellegi-Holt problem. The maximal distance solution 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.,

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

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 $\hat{\text{var}}(e_{ij})$ an estimate for the variance of the prediction error. As one of the reviewers remarked, more involved distance measures could have been used instead, for instance a Mahalanobis distance that takes the correlations between the residuals into account. 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, which is a richer model that replaces the single parameter estimate $\bar{y}_j/\bar{x}_j$ by similar estimates for each stratum separately, but for notational simplicity we only describe the unstratified case here. In the applications the predictor used was the only relevant continuous register variable (registered turnover) in combination with stratification by industry type.

4.3. Imputation

In this section we sketch the imputation methods we have applied. For more details we refer to Pannekoek (2004a) and Pannekoek and van Veller (2004).

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 nonnegative variables we also notice that if the total variable of a balance edit equals the sum of the nonmissing 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. It should be noted that deductive imputations will be in error if the observed values from which these imputations are derived contain errors. Nevertheless, these deductive imputations are the only values that are consistent with the edit rules. So, given that the possibilities for finding and correcting errors are exhausted, deductive imputation is a logical first imputation step. In Section 5.2.5 we discuss the influence of errors on other (nondeductive) imputation methods.

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 into 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:

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

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 nonnegative 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 (a) has no missing values on the auxiliary variables and the target variables, and (b) 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.

A distance measure that is often used for this purpose by NSIs is the minimax distance (see e.g., Sande 1983; Little and Rubin 1987, p. 66). That is the measure chosen for the nearest neighbour module in the edit and imputation software system of Statistics Canada (Statistics Canada 1998). This measure is motivated by the idea that it is important that the donor record is similar to the receptor record on all matching variables simultaneously: potential donors with a large difference on *any* of the matching variables will have a large value for $d(i,i')$ and will therefore not be selected.

Before applying a distance function it is customary to scale the auxiliary variables so 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 so that the maximal absolute difference between the auxiliary variables of the donor and the receptor is minimal. This way of selecting a donor ensures that even the most differing matching variable of the donor record is close to the receptor record. The method is therefore robust against the presence of outliers.

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 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 nonnegativity, results. It can happen that a donor is chosen for which all the missing components are zero. Then, the ratios 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 nonnegativity edits, and each variable occurs only once in a 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 3.2 of Section 3.2. Suppose that the subtotals of Table 3.2, *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 so that the adjustments are as small as possible. This goal is achieved by minimising a distance function measuring the distance between the imputed record, which may not satisfy all edits, and an adjusted record, where imputed values have been changed so that all edits are satisfied. We assume that only linear numerical edits are specified. For convenience, we write the set of linear numerical edits as

$$\mathbf{A}\mathbf{y} \geq \mathbf{b} \tag{2}$$

where $\mathbf{y}$ denotes the vector of values in the record under consideration, $\mathbf{A}$ a matrix and $\mathbf{b}$ a vector. Together $\mathbf{A}$ and $\mathbf{b}$ define the set of edits. Note that the system (2) can include equations as any equation can be written as two inequalities. We partition the vector $\mathbf{y}$ for the record under consideration into the imputed variables $\mathbf{y}_{\text{imp}}$ and the nonimputed, i.e., observed, variables $\mathbf{y}_{\text{non}}$. For notational convenience, we assume that the vector $\mathbf{y}$ starts with all imputed variables followed by all nonimputed variables, i.e., $\mathbf{y} = (\mathbf{y}_{\text{imp}}, \mathbf{y}_{\text{non}})$. We partition $\mathbf{A} = (\mathbf{A}_{\text{imp}}, \mathbf{A}_{\text{non}})$ accordingly. We denote the (possibly) imputed values in the record under consideration by $\mathbf{y}^0 = (\mathbf{y}^0_{\text{imp}}, \mathbf{y}^0_{\text{non}})$. We fill in the values for the nonimputed fields of the record under consideration into (2), and obtain a set of linear constraints for the fields that have been imputed. This set of constraints is given by

$$\mathbf{A}_{\text{imp}}\mathbf{y}_{\text{imp}} \geq \mathbf{b} - \mathbf{A}_{\text{non}}\mathbf{y}^0_{\text{non}} \tag{3}$$

The right-hand side of (3) is constant. The vector $\mathbf{y}_{\text{imp}}$ consists of all unknowns. Note that if the fields to be imputed are determined by a Fellegi-Holt system like *Cherry Pie* the imputed values can indeed be modified so that all edits become satisfied, because the solutions found by a Fellegi-Holt system can, by definition of a Fellegi-Holt system, be imputed consistently.

To measure the distance between an imputed record with known values $\mathbf{y}^0 = (\mathbf{y}^0_{\text{imp}}, \mathbf{y}^0_{\text{non}})$ and an adjusted record with unknown values $\mathbf{y} = (\mathbf{y}_{\text{imp}}, \mathbf{y}^0_{\text{non}})$ we use

$$\sum_{i \in \text{imp}} w_i \left| y^0_{\text{imp},i} - y_{\text{imp},i} \right| \tag{4}$$

where $i \in$ imp indicates that the summation is taken over all imputed values. The w_i ($i = 1, \ldots, n$, where n denotes the number of variables) are nonnegative user-specified weights that are used to compare a change in one variable to a change in another variable. The larger 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*, because here the weights reflect the effects of deviating from the imputed values, whereas the weights used in *Cherry Pie* reflect the level of confidence in the variables. In our application we simply chose all $w_i = 1 (i = 1, \ldots, n)$.

We now seek an adjusted vector $\tilde{\mathbf{y}}_{\text{imp}}$ that minimises (4) subject to the constraints given by (3). This is a linear programming problem, which can, for instance, be solved by means of the well-known simplex algorithm (see Chvátal, 1983). The adjusted, final record $\tilde{\mathbf{y}} = (\tilde{\mathbf{y}}_{\text{imp}}, \mathbf{y}^0_{\text{non}})$ satisfies all edits (2).

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 that 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 (2004) for the $Y_{2,D}$ data, Pannekoek (2004b) for the $Y_{2,E}$ data, Vonk, Pannekoek, and de Waal (2003) for the $Y_{3,D}$ data, and Vonk, Pannekoek, and de Waal (2004) for the $Y_{3,E}$ data. For a detailed comparison with other methods we refer to Chambers and Zhao (2004a and 2004b).

5.1. Evaluation criteria

To evaluate the editing and imputation methods, we use a limited set of evaluation criteria in this article. To measure the error-finding performance of our approach we use an alpha, a beta and a delta measure. The alpha measure equals the proportion of cases where the value for the variable under consideration is incorrect but is still judged acceptable by the editing process. It is an estimate of 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 under consideration 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 of the probability of an incorrect outcome from the editing process for the variable under consideration, and measures the inaccuracy of the editing procedure for this variable.

To measure the imputation performance we use a d_{L1}, an m_1 and an *rdm* measure. The d_{L1} measure is the average distance between the imputed and true values 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 under consideration, Y_i^* is the corresponding true value, M denotes the set of records with imputed values for variable Y and w_i is the raising weight for record i.

The m_1 measure, which measures the preservation of the first moment of the empirical distribution of the true values, is defined as

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

Finally, the *rdm* (relative difference in means) measure is defined as

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

It is important to note that these imputation performance measures are only used in a relative way, i.e., to compare different imputation methods in an experimental setting. Smaller values of the measures indicate better imputation performance. These measures are not necessarily appropriate or sufficient to measure the effect of imputation on the quality of survey estimates in general. For an actual production process it depends on the intended use of the data whether record level accuracy (d_{L1}) or more aggregate measures of imputation bias like m_1 or *rdm* are more important. Furthermore, to assess the importance of bias caused by imputation it should be related to other quality aspects such as sampling variance.

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 1,000, 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 of 300 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 three 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.

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 (Strategy I) and one that uses only the hard edit rules (Strategy II). For the EPE data, no soft edits were specified by the subject-matter specialists.

5.2.3. Different stratifications for the multivariate regression imputation 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 of their experiments with the ABI data

(Di Zio, Guarnera, and Luzi 2004). This stratification is based on the register variables *turnreg* and *empreg* and consists of the following three strata for each type of questionnaire: (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 six for variables that are on both questionnaires and three for variables that are only part of either the long questionnaire or the short questionnaire. This last stratification variable will be referred to as *strat*.

In order to decide which stratification to use, the multivariate regression imputation method was applied to the variables *turnover, emptotc, purtot, taxtot, stockbeg,* and *stockend* (see Table 3.1 for a description of these variables) and *pursale* (purchases of goods bought for resale) of the ABI $Y_{2,D}$ data set, using each of these stratifications. To compare the results we computed, for each variable, the relative difference between the mean of the imputed values and the mean of the corresponding true values. The results showed that stratification by *strat* leads to a better preservation of the mean than stratification by *class* for six of the eight variables, even though the number of classes is much smaller. 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 eight 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 number of employees.

5.2.4. Hot-deck imputation versus regression imputation

One of the imputation methods considered for component variables was the ratio hot-deck imputation method, but for some component variables we investigated the performance of regression imputation as well. Application of these two methods to the six component purchase variables, i.e., the six component variables of *purtot*, of the ABI $Y_{2,D}$ data showed that, with respect to the *rdm* criterion, multivariate regression imputation is better for three variables but for the other three variables ratio hot-deck is better.

These results do not point strongly to one of the imputation methods as the method of choice. The regression imputation method has some disadvantages not shared by the ratio hot-deck imputation method. In particular, some imputed values are negative while the corresponding variables should only assume nonnegative values 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 imputation method, but regression imputation is not well suited either, because these variables contain a large number of zero

values. For these variables a standard nearest neighbour hot-deck imputation method was used with a distance function based on the variables *turnreg* (registered turnover) and *empreg* (registered number of employees), and stratification by *class*.

Several alternatives to this hot-deck imputation method were investigated and evaluated on two criteria:

- The relative difference between the mean of the imputed values and the mean of the true values for the missing data;
- The difference between the number of imputed zero values and the true number of zero values among the missing data.

One alternative was a two-step approach, using a hot-deck method to impute whether or not the missing value is zero and subsequently a regression imputation approach to impute only the nonzero values. Negative imputations by the regression step of this method were set to zero. The preservation of the mean value for this approach was a little bit better than for the hot-deck imputation method. The number of zero values, however, appeared to be much too large because of the extra zeroes introduced by the regression part of this method (besides the zeroes that had already been imputed by the hot-deck). To prevent these extra zeroes the regression imputation was also applied with a log transformation of the target variable. This resulted in the same, rather accurate, number of zero values as the hot-deck method but the performance with respect to the preservation of the mean was much worse. So, if it is important to have the number of firms that have nonzero values for assets disposed (*assdisp*) or assets acquired (*assacq*) about right and at the same time preserve the means reasonably well, the hot-deck imputation 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 development 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 5.1 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).

Two of the variables in Table 5.1 (*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 for two reasons. First, the values of some of the predictor variables in the records with missing values can be erroneous. Second, 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. In our case, the four undetected errors are all in records with no missing

Table 5.1. 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	1,872.67	1,073.41	−0.43	21	1,509.42	1,413.06	−0.06	19
totexpto	206.38	46.38	−0.78	36	1,083.45	1,083.45	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

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 the 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

In total, about 42% of the values to be imputed in the EPE $Y_{2,E}$ data could be deductively imputed, and for the EPE $Y_{3,E}$ data the figure is approximately 45%. 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 fields to be imputed in each of the $Y_{3,E}$ data sets (ABI and EPE) depends on the number of implausible values that have been identified.

5.3.2. Adjustment of imputed values

As mentioned in Subsection 4.3.3, the imputation methods for the ABI data already take the hard edits into account and adjustment of imputed values is therefore not necessary. For the EPE data sets not all hard edits (see Section 3.2 for a brief description of these edits) are taken into account and the imputed values have been adjusted so that the final records satisfy all hard edits. But since most of the hard edits for the EPE data sets are taken into account, the effect of adjusting imputed values is limited. For the EPE $Y_{2,E}$ data only 111 of the 2,230 imputed values are 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 to the imputed values owing to the adjustment procedure is about 2.5%. For the $Y_{3,E}$ data, 95 values of the 2,362 imputed values were adjusted, i.e., about 4.0%, and the average change is 1.1%.

5.3.3. Edit and imputation results for the ABI evaluation data

In Table 5.2 the error localisation results for Strategies I and II (see Subsection 5.2.2) are presented.

Taxrates (amounts paid for national nondomestic rates) and *taxothe* (other amounts paid for taxes and levies) in Table 5.2 are the two component variables of *taxtot*.

The alphas are quite high for both strategies, pointing to a large proportion of undetected errors. Because fewer edits apply to the variables, it is evident that the alphas are larger for Strategy II than for Strategy I. Conversely, the betas are smaller, because using less edits

Table 5.2. 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

results in fewer correct values considered implausible by the editing process. Most deltas are similar or smaller for Strategy II than for Strategy I, showing that the amount of misclassification is smaller with fewer edits.

In Table 5.3 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).

The results show, with a few exceptions, that 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 5.2 it was concluded that the error localisation Strategy II performed better than Strategy I. The imputation results in Table 5.3, however, show that the difference in imputation performance between these two experiments is not so clear cut. For three of the

Table 5.3. Imputation results for ABI evaluation $Y_{3,E}$ and $Y_{2,E}$ 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

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 5.4. 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.

The imputation results (not presented here; see Pannekoek 2004b, and Vonk, Pannekoek, and De Waal 2004) show that the imputation performance is better more often for the $Y_{2,E}$ data than for the $Y_{3,E}$ data.

5.4. Comparison with other approaches regarding the $Y_{2,E}$ and $Y_{3,E}$ data

An important part of the EUREDIT project was the comparative analysis of the results of the different experiments. Chambers (2004) developed a number of measures to assess and compare the edit and imputation performance of the experiments. Based on these criteria, Chambers and Zhao (2004a and 2004b) performed a comprehensive analysis of all the results of the EUREDIT experiments. Our edit and imputation approach to the ABI and EPE data has been compared to an approach based on self-organising maps (cf., Koikkalainen 2004; Koikkalainen, Piela, and Laaksonen 2004), and to several outlier detection techniques and outlier-robust imputation methods (cf., Béguin and Hulliger 2004a; 2004b; 2004c; Chambers, Hentges, and Zhao 2004; Hentges 2004a; 2004b; Ren and Chambers 2004). In order to give an impression of the performance of our approach compared to the performance of the approaches of other participants in the EUREDIT project, we will summarise some of the general conclusions of Chambers and Zhao. In addition, we identify some situations where other methods performed better than our methods in order to indicate directions for improvements of our approach.

Table 5.4. 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

Some general conclusions for data sets without errors (missings only) are as follows. With respect to the ABI $Y_{2,E}$ data Chambers and Zhao (2004a) 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 Zhao (2004b) note that the best experiment (of the two) carried out by CBS does particularly well, and is the best overall experiment. For the data sets with errors Chambers and Zhao formulate the following conclusions. With respect to the ABI $Y_{3,E}$ data they conclude that outlier-robust regression tree-based automatic editing and imputation procedures are the best for this data set, and are worth developing further as an editing and imputation tool for business survey data. In addition, they conclude that the linear model-based methods used, for instance in an experiment by CBS, 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). With respect to the EPE $Y_{3,E}$ data Chambers and Zhao conclude that of the four experiments involving both editing and imputation that were carried out, it is clear from the analysis that the experiment by CBS is the overall leader.

According to these conclusions our approach is among the best-performing ones for the EPE $Y_{2,E}$ and $Y_{3,E}$ data, as well as for the ABI $Y_{2,E}$ data. For the ABI $Y_{3,E}$ data our approach performed reasonably well but there are methods that perform better with respect to error localisation as well as with respect to imputation. Methods that performed relatively well (compared to other methods) for error localisation for all variables and for imputation of all variables except the four economic total variables *turnover*, *emptotc*, *purtot*, and *taxtot* are methods based on robust regression trees. A regression tree is a nonparametric regression modelling procedure that uses a sequence of binary splits of the data set resulting in subgroups or "nodes" that become increasingly homogeneous with respect to the target variable. The splits are defined in terms of the values of a set of covariates. For the applications considered here, the covariates were the register variables *turnreg* and *empreg*. The robustness of the procedure was obtained by using an outlier-robust measure of heterogeneity. Outlier values with respect to this tree-model are defined, and treated as errors. These errors as well as the missing values are subsequently imputed by a robust estimate of the node mean. For a more detailed explanation of this approach, see Chambers, Hentges, and Zhao (2004), Hentges (2004a; 2005a), and Zhao and Chambers (2004).

For the ABI $Y_{3,E}$ data set, this outlier-based error detection procedure performed better than our approach. For instance, for the 12 variables in Tables 5.2 and 5.3 this experiment resulted in average values of alpha (fraction undetected errors), beta (fraction of correct values incorrectly detected as errors) and delta (fraction erroneous decisions) of 0.56, 0.011 and 0.051, respectively, whereas the corresponding values for the CBS experiment were 0.67, 0.001 and 0.059. This shows that the outlier-based error detection method finds more errors than the combination of the Fellegi-Holt based method and the detection of thousand-errors. But there is a trade-off; the smaller values for alpha go together with larger values for beta: more correct values are "detected" as errors. Nevertheless, the fraction of erroneous decisions is larger for the CBS experiment. Similar conclusions hold for other experiments based on outlier detection methods.

With respect to the imputation performance, the results vary according to type of variables involved. For four total variables, *turnover*, *emptotc*, *purtot,* and *taxtot*, a robust parametric regression imputation method using *turnreg* as the only predictor variable performed somewhat better than our multivariate regression approach. This is remarkable because the multivariate regression approach used more predictor variables, namely each of the variables *turnreg*, *turnover*, *emptotc*, *purtot*, *taxtot*, *stockbeg*, *stockend* (see Table 3.1 for a description of these variables), as well as *pursale* (purchases of goods bought for resale) when they were observed. Here it seems clear that there is an advantage to using outlier robust imputation models. For the components of *emptotc*, *purtot*, and *taxtot*, as well as for the variables with many zeroes (*stockbeg*, *stockend*, and *capwork*) a robust regression tree imputation method performed better than our (ratio) hot-deck methods. Again, the robustness of the method may be the reason for this better performance.

6. Conclusions

From the analyses carried out by Chambers and Zhao (2004a; 2004b), which are summarised in Section 5.4, we conclude that the approach used by CBS performed well in comparison with the other methods. Our approach could be applied to edit and impute both the ABI and EPE data, something that many edit and imputation approaches evaluated under the EUREDIT project were unable to do. 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 systematic errors and the imputation of erroneous and missing values, can, if desired, be modified separately without having to change the other steps of the approach. Furthermore, more 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. Such an outlier detection step can, for instance, be added to our approach immediately after the detection and correction of systematic errors. The imputation method we have applied can be replaced by outlier-robust versions of the regression and hot-deck imputation methods.

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 small 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 systematic mistakes, such as thousand-errors;
- Application of selective editing to split the records into a critical stream and a non-critical stream (see Lawrence and McDavitt 1994; Lawrence and McKenzie 2000; Hedlin 2003);
- Editing of the data: the records in the critical stream are edited interactively, the records in the noncritical stream are edited and imputed automatically;
- Validation of the publication figures by means of (graphical) macroediting.

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 (cf., Hoogland 2002) that split the records into a critical stream and a noncritical stream are applied. Very unreliable or highly influential records lead to a low score on the plausibility indicators. Such records constitute the critical stream, and are edited interactively. The other records, i.e., the records in the noncritical stream, are edited automatically. Each year we edit and impute the same business surveys. To apply our automated approach to a new version of a business survey, we therefore only have to update the parameters. This updating process is to a substantial extent automated too. Edit and imputation of the records in the noncritical stream hence requires hardly any human intervention. This is in stark contrast with our experiences in the EUREDIT project, where we had to develop edit and imputation strategies for the ABI and EPE data sets from scratch. For some evaluation results on 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, for instance, compare the publication figures based on the edited and imputed data to publication figures from a previous year. In this final step the focus is more on the overall results than on the correctness of individual records. 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.

At CBS, outlier detection techniques are used during the selective editing step and the macro-editing step. Large errors that were undetected by our approach in the EUREDIT project would in the CBS production process probably be detected in either the selective editing or the validation step. In contrast to our approach in EUREDIT, where we had to restrict ourselves to edit and imputation methods using only data from the data set to be edited and imputed itself, in our production process for structural annual business surveys we use cleaned data from a previous year, for instance during selective editing, manual editing, automatic editing and the validation step.

One could argue that with selective editing the automatic editing step is superfluous. At CBS, we strongly advocate the use of automatic editing, even when selective editing is used. We mention three reasons. First, the sum of the errors in the noncritical records may have an influential effect on the publication figures, even though each error itself may be noninfluential. Provided that the set of edits used is sufficiently powerful, application of the Fellegi-Holt paradigm generally results in data of higher quality. This is confirmed by various evaluation studies such as Houbiers, Quere, and de Waal (1999) and Hoogland and van der Pijll (2003). Second, many noncritical records will be internally inconsistent if they are not edited, which may lead to problems when publication figures are calculated. Finally, automatic editing provides a mechanism to check the quality of the selective editing procedures. If selective editing is well-designed and well-implemented, the records that are not selected for manual editing need no or only slight adjustments. Records that are substantially changed during the automatic editing step therefore point to an incorrect design or implementation of the selective editing step. We feel that automatic editing, when used in combination with other editing techniques, can only improve the quality of the data, not deteriorate it.

We also feel that only a combined approach using selective editing, interactive editing, automatic editing and macroediting 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, led to good results in comparison with the other methods.

7. References

Béguin, C. and Hulliger, B. (2004a). Multivariate Outlier Detection in Incomplete Survey Data: The Epidemic Algorithm and Transformed Rank Correlation. Journal of the Royal Statistical Society, Series A, 167, 275–294.

Béguin, C. and Hulliger, B. (2004b). Multivariate Outlier Detection in Incomplete Survey Data: The BEM Algorithm. Methods and Experimental Results from the EUREDIT Project, J.R.H. Charlton (ed.). (http://www.cs.york.ac.uk/euredit/).

Béguin, C. and Hulliger, B. (2004c). Robust Multivariate Outlier Detection and Imputation with Incomplete Data. Methods and Experimental Results from the EUREDIT Project, J.R.H. Charlton (ed.). (http://www.cs.york.ac.uk/euredit/).

Bethlehem, J.G. and van de Pol, F. (1998). The Future of Data Editing. In Computer Assisted Survey Information Collection, M. Couper, R. Baker, J. Bethlehem, C. Clark, E. Martin, W. Nicholls, and J. O'Reilly (eds). New York: Wiley, 201–222.

Chambers, R. (2004). Evaluation Criteria for Statistical Editing and Imputation. Methods and Experimental Results from the EUREDIT Project, J.R.H. Charlton (ed.). (http://www.cs.york.ac.uk/euredit/).

Chambers, R., Hentges, A., and Zhao, X. (2004). Robust Automatic Methods for Outlier and Error Detection. Journal of the Royal Statistical Society, Series A, 167, 323–339.

Chambers, R. and Zhao, X. (2004a). Evaluation of Edit and Imputation Methods Applied to the UK Annual Business Inquiry. Towards Effective Statistical Editing and Imputation Strategies – Findings of the EUREDIT Project, J.R.H. Charlton (ed.). (http://www.cs.york.ac.uk/euredit/).

Chambers, R. and Zhao, X. (2004b). Evaluation of Edit and Imputation Methods Applied to the Swiss Environmental Protection Expenditure Survey. Towards Effective Statistical Editing and Imputation Strategies – Findings of the EUREDIT Project, J.R.H. Charlton (ed.). (http://www.cs.york.ac.uk/euredit/).

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

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 Quere, R. (2003). A Fast and Simple Algorithm for Automatic Editing of Mixed Data. Journal of Official Statistics, 19, 383–402.

Di Zio, M., Guarnera, U., and Luzi, O. (2004). Application of GEIS to the UK ABI Data: Editing. Methods and Experimental Results from the EUREDIT Project, J.R.H. Charlton (ed.). (http://www.cs.york.ac.uk/euredit/).

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

Ghosh-Dastidar, B. and Schafer, J.L. (2003). Multiple Edit/Multiple Imputation for Multivariate Continuous Data. Journal of the American Statistical Association, 98, 807–817.

Granquist, L. (1995). Improving the Traditional Editing Process. In Business Survey Methods, B.G. Cox, D.A. Binder, N. Chinnappa, A. Christianson, and P. Kott (eds). New York: John Wiley and Sons, 385–401.

Granquist, L. (1997). The New View on Editing. International Statistical Review, 65, 381–387.

Granquist, L. and Kovar, J. (1997). Editing of Survey Data: How Much is Enough? In Survey Measurement and Process Quality, L. Lyberg, P. Biemer, M. Collins, E. De Leeuw, C. Dippo, N. Schwarz, and D. Trewin (eds). New York: John Wiley and Sons, 415–435.

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

Hentges, A. (2004a). Robust Multivariate Outlier Detection Based on Forward Search Methods. Methods and Experimental Results from the EUREDIT Project, J.R.H. Charlton (ed.). (http://www.cs.york.ac.uk/euredit/).

Hentges, A. (2004b). Robust Multivariate Outlier Detection via Forward Search: Application to the ABI Data Set. Methods and Experimental Results from the EUREDIT Project, J.R.H. Charlton (ed.). (http://www.cs.york.ac.uk/euredit/).

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

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

Houbiers, M., Quere, R., and de Waal, T. (1999). Automatically Editing the 1997 Survey on Environmental Costs. Internal report (BPA number: 4917-99-RSM), Statistics Netherlands, Voorburg.

Koikkalainen, P. (2004). Description of the Error Localisation Methodology Based on the Tree Structured Self-Organising Map. Methods and Experimental Results from the EUREDIT Project, J.R.H. Charlton (ed.). (http://www.cs.york.ac.uk/euredit/).

Koikkalainen, P., Piela, P., and Laaksonen, S. (2004). Description of the Imputation Methodology Based on the Tree Structured Self-Organising Map. Methods and Experimental Results from the EUREDIT Project, J.R.H. Charlton (ed.). (http://www.cs.york.ac.uk/euredit/).

Lawrence, D. and McDavitt, C. (1994). Significance Editing in the Australian Survey of Average Weekly Earnings. Journal of Official Statistics, 10, 437–447.

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

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

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

Pannekoek, J. (2004a). (Multivariate) Regression and Hot-Deck Imputation Methods. Methods and Experimental Results from the EUREDIT Project, J.R.H. Charlton (ed.). (http://www.cs.york.ac.uk/euredit/).

Pannekoek, J. (2004b). Imputation Using Standard Methods: Evaluation of (Multivariate) Regression and Hot-Deck Methods. Methods and Experimental Results from the EUREDIT Project, J.R.H. Charlton (ed.). (http://www.cs.york.ac.uk/euredit/).

Pannekoek, J. and van Veller, M.G.P. (2004). Regression and Hot-Deck Imputation Strategies for Continuous and Semi-Continuous Variables. Methods and Experimental Results from the EUREDIT Project, J.R.H. Charlton (ed.). (http://www.cs.york.ac.uk/euredit/).

Ren, R. and Chambers, R. (2004). Outlier Robust Methods: Outlier Robust Estimation and Outlier Robust Imputation by Reverse Calibration. Methods and Experimental Results from the EUREDIT Project, J.R.H. Charlton (ed.). (http://www.cs.york.ac.uk/euredit/).

Sande, I.G. (1983). Hot-Deck Imputation Procedures. In Incomplete Data in Sample Surveys, Vol. III: Symposium on Incomplete Data, W.G. Madow and I. Olkin (eds). New York: Academic Press.

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

Statistics Canada (1998). GEIS: Functional Description of the Generalized Edit and Imputation System.

Vonk, M., Pannekoek, J., and de Waal, T. (2003). Development of (Automatic) Error Localisation Strategy for the ABI and EPE Data. Report (Research Paper 0302), Statistics Netherlands, Voorburg.

Vonk, M., Pannekoek, J., and de Waal, T. (2004). Edit and Imputation Using Standard Methods: Evaluation of the (Automatic) Error Localisation Strategy for the ABI and EPE Data Sets. Methods and Experimental Results from the EUREDIT Project, J.R.H. Charlton (ed.). (http://www.cs.york.ac.uk/euredit/).

Zhao, X. and Chambers, R. (2004). Outlier Identification and Imputation Using Robust Regression Trees. Methods and Experimental Results from the EUREDIT Project, J.R.H. Charlton (ed.). (http://www.cs.york.ac.uk/euredit/).

Received November 2003
Revised September 2004