The EUREDIT project: activities and results

Giulio Barcaroli [1]

Abstract

The EUREDIT project was carried out from 2000 to 2003 under the 5th Framework Program of European Research, with the aim of developing and evaluating new methods for data editing in official statistics, in particular with respect to the phases of (i) error localisation and (ii) imputation of errors and missing values. Multi-Layer Perceptrons, Self-Organising Maps, Correlation Matrix Memories and Support Vector Machines have been considered as new methods for error localisation and imputation, together with robust methods for outlier identification and treatment. In addition, standard methods (edit rules based and model-based) have been taken into consideration as a benchmark to evaluate the performance of the new methods. For this purpose, a set of performance indicators were defined, together with an experimental plan making use of different datasets, covering the different typologies of data currently treated in official statistical production processes. Results of experiments are reported and discussed, together with methodological indications on how to optimally conduct evaluation tasks of this kind.

Keywords: statistical data editing, error localisation, imputation, outliers detection, quality evaluation.

1 Istituto Nazionale di Statistica.

Introduction

The EUREDIT project ("The development and evaluation of new methods for editing and imputation") was carried out under the 5th Framework Program of European Research from March 2000 to February 2003. To the project participated six national institutes, four universities and two private firms: ISTAT[2], the UK Office for National Statistics (as coordinator of the project), Statistics Denmark, Statistics Netherlands, the Swiss Federal Statistical Office, Statistics Finland, the universities of Jyvaeskylae (Finland), Southampton, York, and Royal Holloway College (UK), Qantaris GmbH (Germany) and Numerical Algorithms Group (UK). Objective of the project was the application and evaluation of new approaches and algorithms to the problems of (i) error localisation in data, and (ii) imputation of errors and missing values. The most important new approach to be evaluated was identified basically in the family of techniques developed for pattern recognition and based, strictly or loosely, on artificial neural networks paradigm. But not only that: for particular problems, as robust estimation and treatment of time series, also ad hoc methods were ideated and evaluated.

For the evaluation of these new methods, it was necessary to consider and make available the following elements:

1. a set of standard methods, already defined and currently in use, whose performance might be considered as a benchmark for the evaluation of the new ones;
2. a conceptual framework (a set of indicators) for the compared evaluation of the quality of the different methods;
3. a set of different datasets, each of them representing a given typology of data, so as to cover the range of possible situations for a statistical user;
4. an experimental plan for the compared evaluation of all methods.

The basic output of the project is given by a set of reports containing the description and the evaluation of the methods, together with guidelines indicating the conditions for the optimal use of each of them. As additional output, a number of software tools were produced and made available to the project partners[3].

In paragraph 1, a description of *new methods* that were investigated is given. They have been subdivided in those belonging to the class of neural network methods, and those classifiable as robust methods. Methods peculiar for time series are directly dealt with in paragraph 5.

2 Members of the EUREDIT ISTAT project group are: Giulio Barcaroli, Giorgio Della Rocca, Marco Di Zio, Ugo Guarnera, Orietta Luzi, Antonia Manzari, Emanuela Scavalli, Angela Seeber.

3 The results of the project are fully described in deliverable 6.1 ("Methods and Experimental Results from the Euredit Project") and deliverable 6.2 ("Towards effective statistical editing and imputation strategies - Finding of the Euredit Project").

In paragraph 2, *standard methods* are described. Also these are distinguished in two classes, accordingly to their belonging to the Fellegi-Holt family (rules based methods), or to the model-based methods family.

In paragraph 3, the *criteria for the evaluation* of error localisation and imputation performance are introduced.

In paragraph 4 a description is given of the six *datasets* chosen for experimenting the various methods, together with the planned experiments.

In paragraph 5 *results of experiments* are analysed and evaluated for each dataset and, more synthetically, for best methods.

1. The new methods

As already said, we can distinguish two main classes of new methods: those belonging to the wide class of pattern recognition computer intensive methods (mainly *neural networks* methods), and the others related to the particular problem of *outlier* detection and *robust imputation*.

1.1 Neural network methods

In general, a neural network is composed by a set of elementary units (neurones) linked by weighted connections (Bishop 1995, Ripley 1996). Neurons are organised in layers: at least one input layer and one output layer must be present in a net. One or more hidden layers are optional. Weights are determined by using training datasets, in case of *supervised* methods, or on the basis of available data in case of *unsupervised* methods.

Apart from any other possible characterisations, an important feature of neural networks is that they are *non-parametric* methods that can capture not only linear relationships between variables, but also *non-linear*. Another crucial element, very important in the phase of error localisation, is that they do not require the explicit knowledge represented by edit rules, but they rather need a set of cases (training datasets or available data) from which implicit knowledge required to operate is acquired: in other terms, these methods can *learn*.

This is the general approach. Actually, a variety of methods were considered in this class, each of them with relevant peculiarities. In the following, a synthetic description for each will be given.

1.1.1 Multi-Layer Perceptrons (MLPs)

A Multi-Layer Perceptron is a neural network characterised by at least one hidden *layer*. One layer is composed by elementary units called *neurons*, each neuron is linked to others neurons by weighted *connections*. For any given neuron x_j, the input is given by the weighted sum of the outputs of linked neurons, while its output is the result of the application of a non-linear function $f(x_j) = f(a_j + \sum_{i=1}^{k} w_{ij} y_i)$, where f is typically the *sigmoid* function (logistic or tangent hyperbolic).

Weights of the MLP, initially defined on a random basis, are sequentially adjusted by submitting a set of individual cases, with known values, for any of which predictions are made. The adjustment of weights, carried out so as to obtain best possible predictions, is based on different possible algorithms (all of the type *feed forward*), and proceeds until convergence, i.e. when the accuracy of predictions can no longer be significantly increased. To prevent over-fitting[4] and ensure generalisation, during this process a validation set is also used.

A very important aspect of MLP construction is in the choice of input variables to the network. Redundant information may produce noise that limits predictive capability of the net. A number of techniques to select relevant variables were defined and tested in EUREDIT project (Scavalli,2002).

The application of MLP's to the error localisation task can occur in two basic ways (Nordbotten 1995 and 1996):

- in a subset of cases where it is known when an error occurs in a given variable, and is therefore possible to define an *error flag*, a neural network is trained to predict the value (0 or 1) of the error flag. When applied to the complete set of cases, the neural net outputs (or in other terms, *predicts*) the values of the error flag. Values closest to '1' indicate the presence of errors. It is necessary to define a threshold value above which corresponding values can be judged as erroneous: this is generally done by minimising the total amount of misclassifications (false positive and false negatives);
- in a subset of cases that can be reasonably judged as "error free", a neural network is trained for each variable, so as to predict its values. When the neural net is applied to the complete set of cases, predicted values of each variable become available. The distance between current value and predicted value is an indicator of the presence of errors. Also in this case, a threshold value should be defined in order to assess when a value is erroneous or not.

4 When over-fitting occurs, a solution is found that minimises errors in prediction or classification of training data, but does not perform well on other datasets, i.e. it lacks in generalisation.

With regard to the imputation task, a straightforward solution is the following: for each variable, the subset of cases with no missing errors are considered, and in this subset a neural network is trained to predict values for that variable. The neural net is then applied to the subset of cases with missing values for that variable, and predicted values are imputed to the variable.

This solution is acceptable when the missing mechanism is judged to be MCAR (missing completely at random) or at least MAR (missing at random). In case of NMAR (not missing at random) a different solution should be followed, based on the availability of a subset of cases in which, in correspondence of each missing value for a given variable, also its true value is available.

1.1.2 Tree-Structured Self-Organising Maps (TS-SOMs)

A Self-Organising Map (SOM) is a neural network that approximates a first principal curve, that is a low-dimensional representation (typically one or two dimensions) of a multivariate distribution (Kohonen1997).

The Tree-Structured Self-Organising Map (TS-SOM) algorithm combines the representation capability of the SOM and a tree-search of the best matching unit (Koikkalainen 1999).

When training TS-SOM, several SOMs with different resolution (i.e. with a different number of nodes or neurons or data clusters) are trained and are organised in a tree structure, starting from the simplest SOM at the root, ending to the most complex SOMs at the leaves. The more complex is a SOM, the higher is its capability to represent non-linear relationships in data.

As SOMs are unsupervised neural networks, the training does not require the availability of a subset of true data (as in the case of MLPs). To train a TS-SOM it is necessary to define the following parameters:

- the number of layers in the tree: this parameter defines the complexity of the net: the higher the number of the layers, the higher the complexity of the SOMs in the final layers;
- the robustness of the training algorithm: for continuous variables, the observations that are k times the value of the standard deviation in the nodes are considered to be outliers, while for categorical variables a “cut probability” is defined in order to mark the observation as an outlier.

Once a TS-SOM has been trained, it is possible to use it to localise errors in data. This is done by (i) searching in the tree the best matching SOM for the current unit, and (ii) by considering the differences between the SOM model projections and the observed values. Potential errors are those that show the largest differences. A threshold is defined in order to choose actual errors.

To perform imputation, TS-SOM can be used in a similar way. For any observation with missing value, the best SOM is searched in the tree, and a conditional distribution is therefore available for the missing values. There are different possible imputation procedures:

- the mean in the cluster;
- a random draw from a probability density function;
- a random donor;
- a nearest neighbour donor;
- a MLP regression model specific for the node.

The differences between observations and predicted values are computed in terms of Euclidean distance between vectors of values. This requires that data are preventively pre-processed in order to perform equalisation of variable ranges, normalisation of scales, log-transformations and dummy coding of categorical variables.

1.1.3 Correlation Matrix Memories (CMMs)

A Correlation Matrix Memory is a particular type of neural network that is trained to associate pairs of patterns (an input pattern and an output pattern). It requires only a single cycle through the training data in order to learn the association of a pair of patterns, while the majority of neural networks require many training cycles, necessary to fit a non linear-regression model to data, where also implicit relationships between variables are represented. CMM create an explicit associative mapping between input and output patterns, instead of regression-type models.

The use of CMM for error localisation involves the following steps. First, a pre-processing of data provides to convert them into binary format. A training of a CMM using the resulting binary representation of every record of data is performed. Then, for each record in the dataset, the trained CMM is applied to find the j best matches. Similarities between records (or "patterns") are determined by considering their Hamming distance. For these j best matches, the k-NN ("k nearest neighbours") subset is considered, and a DKN (the distance from the record from its k-th neighbour) value is computed in the following way:

- for each of the j matched records (neighbours), the Euclidean distance between them and the current record is computed;
- neighbours are sorted accordingly to their Euclidean distance;
- the DKN is retained for the current record.

All records are sorted accordingly to their DKN values. Given a threshold cut-off distance, all records exceeding this cut-off are considered as erroneous. The error-status of each variable in an erroneous record is determined on the basis of individual contribution to the DKN value.

The use of CMM for imputation is quite straightforward: once the k-NN subset for a given record with missing values (or variables flagged for imputation), has been determined the values to be imputed are determined by using one of five possible methods: nearest-neighbour, random donor, median, mean and weighted mean.

In other words, both for error localisation and imputation, CMM is used only in a first step in order to find a set of closest records, that are used differently accordingly to the specific imputation method.

1.1.4 Support Vector Machines (SVMs)

A Support Vector Machine is an algorithm for defining a smoothing function that predicts the values of a set of target variables from a set of explanatory variables (Vapnik 1995). There are two forms of SVM, one for the prediction of continuous variables (SVM for regression, or SVR), the other for binary categorical variables (SVM for classification), both able to learn non-linear functions from data. SVM, originated in the so-called "machine-learning community", can be grouped with other semi-parametric approaches like Multi-Layer Perceptrons and Radial-Basis Functions: semi-parametric in the sense that they offer the efficient training characteristics of parametric techniques, but have the capability to learn non-linear dependencies as non-parametric methods do.

Another possible definition of SVM is the following: a non-linear generalisation of linear techniques (Cristianini and Shawe-Taylor 2000). Covariate data is projected onto a higher dimensional space ("features space"), and then inserted in a linear algorithm: the parameters of the linear model learned in the higher-dimensional space describe a non-linear model in the original space. The advantage of this approach is that the objective function minimised during training is convex quadratic, and therefore the problem of local minima is avoided. SVM learning also avoids over-fitting by introducing a penalisation factor (regularisation) of over-complex models.

1.2 Robust methods

The problem of outliers identification and subsequent treatment is very important, especially in business data. Unidentified outliers can seriously compromise the accuracy of estimates and the validity of standard analyses of data. Not all outliers are errors: they can be characterised as *representative* outliers (corrects values) or *non representative* outliers (errors) (Chambers, 1986). The treatment subsequent to the identification should take into account this distinction, that conversely is not important in the phase of their detection: even if an outlier is not an error, it is nonetheless crucial to detect it in order to give it a special treatment.

Detection requires first of all a metric able to measure the "outlyingness" of a value. Metrics are usually derived by the adoption of models and measures of the discrepancy between real and predicted values (Barnett and Lewis 1994). A very common metric for continuous data is the Mahalanobis distance.

A problem to be dealt with is that the estimation of model parameters can be influenced by those outliers that should be detected by using the model. Robust methods for outlier detection are based in turn on robust estimation of models and distances.

In the following, a number of methods for outlier detection are synthetically illustrated. As for imputation, most of them make use of a particular software, POEM (*imPutation for Outliers, Edit failures and Missing values*), that is a robust nearest neighbour imputation algorithm, while the last one, WAID, has an embedded function that allows not only to detect but also to impute outliers.

1.2.1 Outlier detection: Robust distance via Transformed Rank Correlation (TRC)

The basic idea of Transformed Rank Correlations (Gnanadesikan and Kettenring 1972) is to compose a pseudo covariance matrix $\widetilde{\mathbf{S}}$ using robust bivariate covariances.

This matrix is built by using the standardised Spearman rank correlation, multiplied by the standardised median absolute deviation of the variables involved. Data are then transformed into the space of principle axis derived from the pseudo covariance matrix: the transformation matrix **B** is defined by the equation $\widetilde{\mathbf{S}} = \mathbf{B}\,\mathbf{\Lambda}\,\mathbf{B}^{T}$, with **B** orthogonal and Λ diagonal. The matrix of data **X** is transformed into **Y=XB**, and medians m' and median absolute deviations s' are transformed back to m=**B**m' and **B** diag(s') $\mathbf{B}^{T}$. Finally, the Mahalanobis distance $d^2 = \mathbf{(x_i - m)^T S(x_i - m)}$ is computed for each point, and its ratio to the median of Mahalanobis distances is compared to an F distribution to determine outliers.

1.2.2 Outlier detection: Forward Search Algorithms (BACON)

Forward Search Algorithms start from an initial subset of data that is judged as being free of outliers (Hadi and Simonoff 1993; Riani and Atkinson 2000). In the case of BACON algorithm ("*Blocked Adaptive Computationally Efficient Outlier Nominators*") (Billor et al 2000), this subset of data, of dimension cp, where p is the dimension of data (number of variables) and c is a constant chosen by the statistician (usually c=3), can be built in two different ways: by considering (i) observations with the smallest Mahalanobis distance to mean, or (ii) observations with the smallest Euclidean distance to the median point in the space; of course, the second alternative offers a more robust starting point. This initial subset is used to estimate a regression

model for the variables of interest. The algorithm then calculates Mahalanobis distances for all observations, based on mean and covariance estimated by the model. The next step is to redefine the "clean" subset by considering these new Mahalanobis distances. The procedure is iterated until (i) distances of observation outside the clean set are too large and the clean set does not vary anymore, or (ii) all observations are inside the clean set.

The BACON algorithm has been adapted in order to take into account missing values: in each iteration, the EM algorithm is applied before BACON, under the assumption of a multivariate normal distribution.

1.2.3 Outlier detection: Epidemic Algorithm (EA)

The Epidemic Algorithm (Beguin and Hulliger 2003) simulates an epidemic whose starting point is in the centre of a data scatter and spreads from it stepwise. As first step, Euclidean distances between all observations are calculated. The centre of the distribution (*sample spatial median*) is the point that has the least sum of distances from all other points. The epidemic is such that the probability that an "infected" point i transmits the "disease" to a non infected point j at the next step is inversely proportional to the distance:

$$P(j|i) = h(d_{ij})$$

where d_{ij} is the distance between observations i and j, and h varies from 1 to 0). The form of the transmission function determines the behaviour of the algorithm: for the EUREDIT project, the *inverse power function* has been chosen:

$$h(d) = ((\alpha^{-\frac{1}{p}} - 1)\frac{d}{d_0 + 1})^{-p} \quad \{d \le d_0\}$$

where the constant d_0 is called *reach of the transmission*, and can be determined over the observations as the maximum of the distances to the nearest neighbour. Given a subset I of infected points, the *total infection probability* of an observation j is $P(j|I) = 1 - \prod_{i=1}^{I} (1 - h(d_{ij}))$.

The algorithm starts at the sample spatial median, and at each step the total infection probability of each uninfected point is evaluated. The expected number of new infected points is calculated, and points with the largest total infection probabilities are infected. The algorithm stops when no new infection occurs, or there are no more uninfected points. Infection times of observations are taken as a measure of outlyingness, and simple univariate decisions can be applied to identify outliers. Doubtless outliers are those observations that have never been infected.

1.2.4 Outlier Reverse Calibration Imputation

We assume that a reliable population total estimate $\hat{t}_y = \sum_{i \in s} w_i^* Y_i$ can be obtained by some outlier-resistant procedure: here, weights w_i^* are weights w_i (inverses of inclusion probabilities, or calibration weights) that have been corrected in such a way so to ensure this resistance. We can also say that $\hat{t}_y = \sum_{i \in s} w_i^* Y_i = \sum_{i \in s} w_i Y_i^*$.

In other terms, we can obtain the same outlier-resistant estimate by imputing values instead of modifying weights. Once outliers have been detected, let s_2 be the subsample of outliers, and s_1 the subsample of inliers. Then, the problem is to define a set of imputed values $(Y_i^*; i \in s_2)$ under the constraint:

$$\hat{t}_y - \hat{t}_{1y} = \hat{t}_y - \sum_{i \in s1} w_i Y_i = \sum_{i \in s2} w_i Y_i^*$$

The imputed values $Y_i^*; i \in {}_2$ should remain as close as possible to the original ones, subject to this constraints. This problem is equivalent to the calibration problem, where the survey variable Y plays the role of the sample weight and the sample weight plays the role of the survey variable. The distance here considered is:

$$d(Y^*, Y) = \sum_{i \in s2} (Y_i^* - Y_i)^2 / 2 q_i Y_i$$

where the q_i 's are constants that are chosen by the statistician. So, it follows (Deville and Sarndal, 1992):

$$Y_i^* = Y_i \left[1 + q_i w_i \frac{\hat{t}_{2y} - \sum_{j \in s2} w_j Y_j}{\sum_{j \in s2} q_j w_j^2 Y_j} \right]$$

1.2.5 Outlier detection and imputation: Robust Tree Modelling (WAID)

In the EUREDIT project, the regression tree modelling software WAID has been used[5]. The basic idea of regression tree models (Breiman *et al* 1984) is to sequentially divide the dataset into subgroups (nodes) that are more and more homogenous with respect to the values of a response variable.

5 WAID regression tree modelling software operates under R (a public domain statistical software) and is an extension of WAID software for missing data imputation developed under the AUTIMP project (Chambers *et al* 2001).

The *univariate* version of WAID allows only one continuous response variable Y and p covariates X_1, ..., X_p, all categorical. The tree modelling is robust in the sense that outliers are down-weighted when the measure of each internal node heterogeneity is calculated: weights are in this case based on *outlier robust influence functions*.

At any iteration, each node k is evaluated in order to decide if it should be split, on the basis of a measure of the heterogeneity given by the weighted sum of square residuals:

$$WSSR_k = \sum_{i=1}^{n_k} w_i (y_i - \bar{y}_{wk})^2$$

where $\bar{y}_{wk}$ is the weighted mean of Y in node k, and the weight w_i is calculated as the ratio $w_i = \frac{\psi(y_i - \bar{y}_{wk})}{(y - \bar{y})}$, where $\psi(x)$ is a given influence function, whose default is the *bi-weight influence function*:

$$\psi(t) = t(1 - (t^2/c^2))^2$$

So, each time a current node is split to create two children nodes, a new set of robust weights is created: outliers receive weights close to zero, while inliers receive weights close to one.

The algorithm defines as outliers those observations that in the overall process of nodes splitting are characterised by an average weight below a specified threshold. The optimal threshold is the one that allows to identify successfully outliers minimising the number of misclassifications.

The only difference of the *multivariate* version of WAID is in the evaluation of the heterogeneity. In particular, one of the possible options defines the weight associated to observation i in candidate node h at stage k as

$$w_i^{(k)} = \frac{\psi\left(\left\| \mathbf{y}_i - \bar{\mathbf{y}}_{wh}^{(k)} \right\|_{wh}\right)}{\left\| \mathbf{y}_i - \bar{\mathbf{y}}_{wh}^{(k)} \right\|_{wh}}$$

where $\mathbf{y}_i$ is the p-vector of response values, $\bar{\mathbf{y}}_{wh}^{(k)}$ is the p-vector of means, and

$$\left\| \mathbf{y}_i - \bar{\mathbf{y}}_{wh}^{(k)} \right\|_{wh} = \sqrt{\sum_{j=1}^{p} s_{whj}^{-2} (y_{ij} - \bar{y}_{whj}^{(k)})^2}$$

where s_{whj}^{-2} is the variance within the candidate node h. Of course, in this case weights

have to be calculated iteratively.

Once a subset of observations have been declared as outliers, the robust tree structure generated by WAID can be used to impute them. There are two possible alternatives: (i) the outlier value is replaced by the weighted mean of the terminal node to which the observation belongs, or (ii) a random donor inside the terminal node is searched for.

2. Standard methods

Standard methods have been considered in the EUREDIT project so as to offer a benchmark for the evaluation of new methods.

These standard methods can be grouped in two different classes:

- methods that are *edit-rule based* and, more specifically, follow the optimal editing approach defined by Fellegi and Holt (Fellegi and Holt, 1976);
- methods that are *model-based*.

2.1 Fellegi-Holt methods (F-H)

These methods are currently being used by a variety of National Statistical Institutes. The set of edit rules is used both for error localisation and for imputation. For error localisation, the subset of edits activated by a given record is processed in order to individuate the subset of variables most likely to contain the errors that caused the activation of those edits. The F-H error localisation algorithm is based on the *minimum change principle*, i.e. the number of variables judged to be erroneous must be the minimum under the constraint to explain all edit failures. A variant to this approach is given by the Nearest-neighbour Imputation Methodology (NIM). Accordingly to NIM, the error localisation is no longer based on the minimum change principle, but on the consideration of the differences between the current record (with edit failures) and a potential donor (a neighbour with no edit failures): this approach can be defined as *data driven*, while the F-H methodology is purely *edits driven*.

For imputation, a range of possible values to impute is first determined in order to avoid values that might cause additional failures of edit rules; then, actual values can be assigned by using a number of different methods, from nearest neighbour to regression imputation. In particular, we include in this category the imputation methods based on the *donor search*, as opposite to regression imputation considered in the model-based methods.

A number of systems incorporating F-H methods have been developed by Statistics Canada, Statistics Netherlands, ISTAT and ONS, and applied in the

EUREDIT project. In the following, a short description for each will be given.

2.1.1 CANCEIS and SCIA for editing and imputation of categorical variables

The CANadian Census Edit and Imputation System (CANCEIS) has been developed by Statistics Canada to be applied to the last Population Census. It fully incorporates the Nearest-neighbour Imputation Methodology (NIM) (Bankier *et al* 2000).

The basic steps of NIM is (i) to search, for each record with edit failures, a set of *nearest neighbours* and, (ii) for each couple recipient-donor, to calculate the minimum number of *imputation actions*, so as to let the recipient failing no edits. As already said, this approach is not strictly adherent to the minimum change principle that characterises the Fellegi-Holt methodology, but has a number of advantages that makes it preferable in some applications. One of them is the editing and imputations of complex hierarchical structures, such as households. NIM allows to consider an entire household as the record to be edited, and experiences carried out made it clear that its performance is higher than that of pure F-H systems or other systems. The NIM approach also allows to handle contemporarily both continuous and categorical data, but so far the only applications we know refer to households categorical data, namely the variables that are linked by constraints that involve more than one member of the household.

On the contrary, SCIA (*Sistema per il Controllo e l'Imputazione Automatici*), developed by the Italian Statistical Institute, is a pure Fellegi-Holt system for the edit and imputation of categorical data (Riccini, 2002). Initially, the set of edit rules defined by the user is analysed and checked for contradictions and redundancies, and the complete set of rules, including implicit edits, is generated. These are applied to each record, and for those failing at least one edit, the minimal set of variables to be changed is determined, on the basis of the coverage of failed edits. Range of acceptable values are also determined for each variable. Then, the imputation step is performed, by searching first a unique donor for all imputations, on the basis of the values of the matching variables. If no such donor can be found, a sequential imputation is tried (one donor for each variable to be imputed). The final option is to impute values on the basis of the marginal distributions.

Experience showed that SCIA performs well for variables that are not subject to hierarchical constraints. Then, a typical edit and imputation application concerning a survey on households will consist firstly of an application of CANCEIS to variables whose edit rules mainly refer to the household composition and constraints (relation to head, sex, marital status and age), and secondly of an application of SCIA involving only individual variables (for instance, level of instruction, social condition, etc.) (Manzari, 2002).

2.1.2 GEIS for editing and imputation of continuous variables

The Generalised Edit and Imputation System (GEIS), developed by Statistics Canada (Kovar *et al* 1988), allows to apply the Fellegi-Holt methodology to continuous data. Only linear edits on non-negative variables are admissible. GEIS enables the user to analyse initial edits, identifying inconsistencies and redundancies. Error localisation is carried out on the basis of the minimum change principle: as in the case of categorical variables, for each record with edit failures the minimum set of variables covering all failed edits is identified and flagged for imputation. It is also possible to apply methods, as the Hidiroglou-Berthelot procedure (Hidiroglou and Berthelot, 1986), for outlier detection (Di Zio *et al* 2002a).

Imputation can be carried out in three different ways (Di Zio *et al* 2002b):

i. *deterministic* imputation, when for a given variable there exists one and only value that once assigned to the variable allows the record to pass the edits;
ii. *nearest neighbour* imputation: among all the units passing the edits, a potential donor with minimum distance is searched and its values, if acceptable, assigned to the recipient variables that require imputation;
iii. *estimated value imputation*: variables are imputed sequentially by using estimates based on different functions (means, ratios, historical trends).

2.1.3 CHERRY-PIE and E-C system for editing and imputation of continuous variables

CHERRY-PIE is another implementation of Fellegi-Holt methodology, that allows the user to handle jointly both categorical and continuous data (De Waal 2002). The output of CHERRY-PIE for each record that fails at least one edit is the list of variables that must be imputed as they have been flagged as erroneous.

The user can adopt whatever imputation method. In EUREDIT experiments a number of them have been used:

- *deductive* imputation (analogous to the GEIS deterministic imputation);
- *multivariate simultaneous regression* imputation: a multivariate regression model is estimated using fully observed predictors, and its predicted values assigned to missing/erroneous values;
- *ratio hot-deck* imputation: in case of balance edits, where many variables are sub-totals referred to a total, regression imputation is not adequate, since imputed variables are never zero and can be also negative; it is therefore better first to impute (by regression or deductively) the total, and then to search a donor (nearest neighbour with respect to the total), and allocate the differences between the variable total and the computed total (as sum of subtotals), by using ratios of subtotals to total in the donor (Pannekoek 2002).

The imputations carried out as outlined above, can lead to additional edit failures, because these imputation methods do not take into account edits. A particular procedure is available, the EC System, that allows to adjust the final values in order to satisfy all rules. Adjustments are made by using the *simplex method*, so as to minimise the distance between imputed and final values, under the constraint that final values satisfy edits.

2.1.4 DIS for imputation of continuous and categorical variables

The Donor Imputation System (DIS) has been developed by the Office for National Statistics to be used in the 2001 UK Censuses. It implements the joint imputation method proposed by Fellegi and Holt in 1976. DIS searches for a donor in three different stages. First, a donor is searched having the same values of the recipient on a set of matching variables (exact match). If no such donor can be found, then categories of each categorical matching variable are collapsed, and the search is repeated. If a donor still cannot be found, less significant matching variables are removed until at least one donor is found. If more than one donor is found, a random selection can be performed. A penalty function is applied in order to avoid imputations of the same donor to many recipients (Yar 1988).

2.2 Model-based methods

The basic idea is to define and fit a (parametric and linear) model for every variable involved in the process of edit and imputation. This model will be used both for error localisation and imputation.

Error localisation is carried out with the following steps:

1. for each variable, an expected value is calculated, conditional on a set of covariates;
2. the actual value is compared to the expected value, and if the two values diverge too much, the actual value can be considered erroneous.

Obviously, problems arise when adopting this approach. Firstly, also covariates can contain errors (or missing values). Secondly, what metric should be adopted in evaluating closeness of actual and expected values, and how to define thresholds beyond which data have to be considered as errors?

As for the imputation, on the basis of a given model the expected value is assigned to missing and erroneous data. Also in this case we have to deal with some problems. First, as in the case of error localisation, we should consider the possibility that covariates may contain errors: if so, also the predicted value will be different from the true one. Second, imputation can be *deterministic* (the predicted value is directly imputed), and in this case first order estimates are generally best preserved,

but further data analysis can be biased by a reduced variability; or imputation can be *stochastic* (the imputed value is drawn by from a conditional distribution), with a reduced preservation of means and totals. Third, imputations carried out in this way generally do not take into account the coherence of imputed values with other values in the record, and edit failures are therefore possible after imputation.

2.2.1 Expectation-Maximisation Algorithm (EM)

EM algorithm is a method for estimating distribution parameters in the presence of missing data, under a specified super-population model and an ignorable non-response mechanism (Dempster *et al* 1977) .

In the presence of missing data the complete data score function, i.e. the first derivative of the logarithm of $L(\theta | Y)$, is not easily computable, so an iterative algorithm is preferred to the analytical solution. The algorithm consists in repeatedly applying standard complete data methods to incomplete data, by iterating the following steps:

1. impute missing data Y_{miss} using current estimates of unknown parameter θ (*expectation* step);
2. re-estimate θ using Y_{obs} and imputed Y_{miss} (*maximisation* step).

The procedure is iterated until convergence to the unique maximum-likelihood estimate of θ.

Two methods of imputation can be used:

- each missing value is imputed with its best prediction $E(Y_{miss}|Y_{obs}, \hat{\theta})$ (the conditional expectation given the observed data and the current estimates of the model parameters);
- the imputation is carried out by drawing randomly from the conditional distribution of missing data given the observed data $P(Y_{miss}|Y_{obs}, \hat{\theta})$.

The first method should be chosen if primary estimates of interest are total or means, while the second is preferable to preserve variability in data.

The convergence of EM algorithm is not ensured if the assumption of multi-normality does not hold, and also imputation is performed on the basis of a multi-normal model. So, real applications do require (i) analysis of data to individuate strata in which multi-normality assumption holds and (ii) transformations of variables (usually, logarithmic transformations).

2.2.2 Integrated Modelling Approach to Imputation (IMAI)

The IMAI approach has been developed at Statistics Finland, and can be used both for error localisation and for missing/erroneous data imputation. It is based on the following different steps:

1. selection of training data and auxiliary variables for any given variable of interest;
2. construction of an error localisation model for the prediction of an error indicator for any given variable of interest, and/or an imputation model for the direct prediction of variables of interest;
3. choice of the criteria for error localisation: in particular, it is necessary to decide a proper cut-off probability for errors;
4. choice of the criteria for data imputation: if the predicted value (with or without an error term) is directly used to impute, then the imputation method is *model-donor*; on the contrary, if the predicted value is used to find a nearest neighbour, the method is *real-donor* (Regression Based Nearest Neighbour, RBNN, see Laaksonen 2000).

3. The evaluation criteria

One of the most important objectives of the EUREDIT project was to individuate best methods for given typologies of data and errors. So, the determination of the evaluation criteria was a crucial task that engaged the first phase of the project.

Different sets of evaluation criteria were defined for error localisation and for imputation. All of them imply that knowledge concerning true values is entirely available. In other words, quality indicators, to be calculated, need to know the true value Y_{ij}^{*} of the j-th variable in the i-th unit in the dataset, the corresponding observed (or raw) value Y_{ij}, and the possibly imputed value. In the following we will introduce separately indicators for the evaluation of error localisation methods and indicators for the evaluation of imputation methods.

3.1 The evaluation criteria for error localisation

When considering an error localisation method, we are interested in evaluating two different performances, namely:

- the *efficient error detection*, i.e. the capability of a method to correctly classify errors and true values in data, or, conversely, its capability to minimise misclassifications (*false negatives*, errors judged as true values, and *false positives*, true values judged as errors);
- the *influential error detection*, i.e. the ability to detect the most influential errors, those with the highest impact on final estimates.

3.1.1 Efficient error detection

After the application of a given method for error localisation, for every variable j of interest in the dataset, the following table can be defined:

	$E_{ij} = 1$ (value judged as correct)	$E_{ij} = 0$ (value judged as erroneous)
$Y_{ij} = Y_{ij}^{*}$ (correct value)	n_{aj}	n_{bj}
$Y_{ij} \neq Y_{ij}^{*}$ (erroneous value)	n_{cj}	n_{dj}

It is evident that frequencies on the main diagonal refer to correct classifications, while in the other two cells misclassifications are contained.

We can define the following indicators:

$$\alpha_j = \frac{n_{cj}}{n_{cj} + n_{dj}} \qquad (1)$$

that is the *false negative rate*, i.e. the proportion of errors that have not been recognised as such by the method, and

$$\beta_j = \frac{n_{bj}}{n_{aj} + n_{bj}} \qquad (2)$$

that is the *false positive rate*, i.e. the proportion of true values that have been erroneously recognised as errors by the method.

Finally,

$$\delta_j = \frac{n_{bj} + n_{cj}}{n} \qquad (3)$$

is the *total misclassification rate*, i.e. an estimate of the probability of an incorrect outcome from the error localisation method.

3.1.2 Influential error detection

It is worth while to measure not only the efficiency of the error localisation method in finding errors, but also its capability to find *influential* errors, in other words the errors that more than others could influence the estimates of interest.

To measure this capability, we introduce the concept of *post-edited* value $\hat{Y}_{ij} = E_{ij} Y_{ij} + (1 - E_{ij}) Y^*_{ij}$. If the measured value Y_{ij} is erroneous, and the method can recognise it as an error, then the post-edited value is assumed to be set to the true value. On the contrary, if the method fails in recognising the error, the post-edited value remain erroneous.

For continuous variables an important quantity is $D_{ij} = \hat{Y}_{ij} - Y^*_{ij} = E_{ij}(Y_{ij} - Y^*_{ij})$, i.e. the difference between the post-edited value and the true value. A desirable property of an error localisation method is that the two distributions of true values and post-edited values are as close as possible.

To measure this closeness, we can define the *relative average error*:

$$RAE_j = \sum_{i=1}^{n} w_i D_{ij} / \sum_{i=1}^{n} w_i Y^*_{ij} \qquad (4)$$

that indicates the mean difference between undetected errors and true values.Values w_i indicates sampling weights, and are obviously used only in case of sample surveys. If variable j can assume also negative values, a more suitable indicator is the relative *root average square error*:

$$RRASE_j = \sqrt{\sum_{i=1}^{n} w_i D^2_{ij}} \Big/ \sum_{i=1}^{n} w_i Y^*_{ij} \qquad (5)$$

A useful measure of how much differences between undetected errors and true values are spread, is given by the *relative error range*:

$$RER_j = R_j (D)/IQ_j(Y^*) \qquad (6)$$

where $R_j(D)$ is the range (maximum - minimum) of the non-zero D_{ij} values, and $IQ_j (Y^*)$ is the inter-quartile distance of the true values.

For categorical (nominal or ordinal) variables, a different indicator has to be defined. Considering the joint distribution of post-edited and true values, we have to take into account the number of cases not lying in the principal diagonal (where $\hat{Y}_{ij} = a$ and $Y^*_{ij} = b$, with $a \neq b$), each of them with an associated distance d(a,b). In case of nominal variables, d(a,a)=0, and d(a,b)=1 for any a,b. In case of ordinal variables, d(a,b) is given by the number of categories that lie between a and b, plus one. So, we can define the *influential error detection performance for a categorical variable*:

$$DCAT_j = \frac{1}{n} \sum_{a=1}^{p_j} \sum_{b \neq a} d(a,b) \sum_{i \in j(ab)} w_i \qquad (7)$$

Another useful measure of the performance of an error localisation method refers to the impact of remaining errors in post-edited data to the variance of the estimator in a sample survey. We can estimate this variance by means of the jackknife formula:

$$v_w(Y) = \frac{1}{n(n-1)} \sum_{i=1}^{n} \left(\left\{ n \sum_{i=1}^{n} w_i Y_i - (n-1) \sum_{k \neq i}^{n} w_k^{(i)} Y_k \right\} - \sum_{i=1}^{n} w_i Y_i \right)^2 \quad (7)$$

where $w_k^{(i)} = w_k \left(\sum_{q=1}^{n} w_q / \sum_{q \neq i}^{n} w_q \right)$. In other words, variance is calculated from survey data each time excluding the i-th unit, and rescaling weights to take into account this exclusion.

Then the indicator

$$t_j = \sum_{i=1}^{n} w_i D_{ij} / \sqrt{v_w(D_j)} \quad (8)$$

is a standardised measure of the *effect of error localisation method on the variance of the estimator*. Values of t_j greater than 2 indicate a significant failure of the error localisation method.

Finally, we can compare the moments and the distributions of the *outlier-free* data value is retained, and values with corresponding moments and distributions of the true values, in order to evaluate the capability of a method to detect outliers. Remembering that $E_{ij} = 1$ if the $E_{ij} = 0$ otherwise, for positive continuous variables we can define the *absolute relative error for the k-Mean*:

$$AREmk_j = \left| \frac{\sum_{i=1}^{n} w_i E_{ij} Y_{ij}^k / \sum_{i=1}^{n} w_i E_{ij}}{\sum_{i=1}^{n} w_i Y_{ij}^{*k} / \sum_{i=1}^{n} w_i} \right| \quad (9)$$

where this indicator is typically calculated for k=1 and k=2 (to compare first and second moments of the two distributions).

3.2 The evaluation criteria for imputation

An imputation procedure should be evaluated with respect to the following properties:

i. *predictive accuracy*: an imputation method should preserve single values, i.e. imputed values should be the same than true values (for categorical variables), or as close as possible to the true values (for continuous variables);
ii. *distributional accuracy*: the imputation procedure should preserve the distribution of true data;

iii. *estimation accuracy*: the imputation method should reproduce as much as possible the lower order moments of the distribution of true data (at least first and second moments).

An additional desirable property is that imputed values should be "plausible", i.e. coherent with other data and not failing any edit rule.

3.2.1 Performance measures for the preservation of true values (predictive accuracy)

Given a *categorical nominal* variable Y with c+1 categories, and be Y_i^* its true value and $\hat{Y}_i$ its imputed value, both in i-th observation, a measure of how well an imputation method preserves true vales is given by

$$D = \frac{1}{n}\sum_{i=1}^{n} I(\hat{Y}_i - Y_i^*) \qquad (10)$$

that is the proportion of off-diagonal entries in the square table of order c+1 obtained by cross-classifying true and imputed values.

In case of a *categorical ordinal* variable, we can define a more general version of D to take into account the distance between true and imputed values.

We can test whether D is significantly greater than a small positive constant s that is an acceptable proportion of incorrect imputation. If $D > \varepsilon + 2\sqrt{\hat{V}(D)}$, where $\hat{V}$ (D) is an estimate of the variance of D, we can say that the imputation method does not preserve true values. We can set

$$\varepsilon^* = \max(0, D - 2\sqrt{\hat{V}(D)}) \qquad (11)$$

The smaller this value, the better is the performance of the method in preserving true values.

In case of a *categorical ordinal* variable, we can define a more general version of D to take into account the distanced $(\hat{Y}_i, Y_i^*)$ between true and imputed values:

$$Dgen = \frac{1}{n}\sum_{i=1}^{n} d(\hat{Y}_i, Y_i^*) \qquad (12)$$

In case of *continuous* variables, a completely different approach is followed. If an imputation methods preserves true values, $\hat{Y}_i$ should be close to Y_i^* for all cases where imputations have been made. A first measure of this closeness can be *weighted Pearson moment correlation* r between $\hat{Y}_i$ and Y_i^*. This measure is not recommended for highly skewed data.

Another approach is based on regression: first, a linear model of the form

$$Y^* = \beta \hat{Y} + \varepsilon \qquad (13)$$

is fitted to the subset of imputed data, and then a test is made whether the *slope* β is equal to 1. If the test does not reveal a significant difference (non significant p-value), then a measure of the *regression mean square error* can be computed:

$$\mathrm{MSE} = \frac{1}{n-1}\sum_{i=1}^{n} w_i (Y_i^* - \beta \hat{Y}_i)^2 \qquad (14)$$

Another regression-based measure is the value R^2, the *proportion of the variance in* Y^* *explained by the variation in* $\hat{Y}$.

The preservation of values can also be directly evaluated by calculating the distance $d(\hat{Y},Y^*)$ between the vector of imputed values and the vector of corresponding true values:

$$d_{L\alpha}(\hat{\mathbf{Y}}, \mathbf{Y}^*) = \left\{ \sum_{i=1}^{n} w_i \left| \hat{Y}_i - Y_i^* \right|^{\alpha} / \sum_{i=1}^{n} w_i \right\}^{1/\alpha} \qquad (15)$$

where typical values of α are 1 and 2.

3.2.2 Performance measures for the preservation of distributions (distributional accuracy)

For a *categorical* variable with c+1 categories, the distributional preservation capability of an imputation can be evaluated by calculating the following Wald-type statistic:

$$W=(\mathbf{R}-\mathbf{S})^t[\mathrm{diag}(\mathbf{R}+\mathbf{S})-\mathbf{T}-\mathbf{T}^t]^{-1}(\mathbf{R}-\mathbf{S}) \qquad (16)$$

where **R** is the c-vector of frequencies of imputed values for the first c categories, **S** is the c-vector of frequencies of true values for these categories, and **T** is the square matrix of order c corresponding to the cross-classification of true and imputed values for these categories. Distribution of W is chi-square with c degrees of freedom, and statistical tests concerning distributional preservation can be carried out.

For *continuous* variables, we introduce the weighted empirical functions for true and imputed values:

$$F_{Y^*}(t) = \sum_{i=1}^{n} w_i \, I(Y_i^* < t) / \sum_{i=1}^{n} w_i$$

$$F_{\hat{Y}}(t) = \sum_{i=1}^{n} w_i \, I(\hat{Y}_i < t) / \sum_{i=1}^{n} w_i$$

We can now measure the distance between the two functions using the *Kolmogorov-Smirnoff distance*:

$$KS = d_{KS}(F_{Y^*}, F_{\hat{Y}}) = \max_t \left(\left| F_{Y^*}(t) - F_{\hat{Y}}(t) \right| \right) \qquad (17)$$

An alternative is the integrated distance

$$KS(\alpha) = d_{\alpha}(F_{Y^{*}n}, F_{\hat{Y}n}) = \frac{1}{t_{2n} - t_0} \sum_{j=1}^{2n} (t_j - t_{j-1}) \left| F_{Y^{*}n}(t_j) - F_{\hat{Y}n}(t_j) \right|^{\alpha} \quad (17b)$$

where t_0 is the largest integer smaller than or equal to t_1. Larger values of α give more importance to larger differences. Usual values of α are α = 1 and α = 2.

3.2.3 Performance measures for the preservation of aggregates (estimation accuracy)

For *continuous* variables, we consider the problem of preserving the raw moments of the distribution. We can measure this preservation by using the indicator:

$$m_k = \left| \sum_{i=1}^{n} w_i (Y_i^{*k} - \hat{Y}_i^{k}) / \sum_{i=1}^{n} w_i \right| = \left| m(Y_i^{*k}) - m(\hat{Y}_i^{k}) \right| \quad (18)$$

with typical assignments of 1 and 2 to k.

4. Datasets and planned experiments

Six different datasets were chosen in order to represent a variety of data (continuous and categorical) and surveys (census and sample surveys; enterprises and households; cross-section and panel) typologies. The characteristics of the datasets have been reported in the following table.

Dataset	Type of dataset	Type of variables	Number of variables	Number of records
Danish Labour Force Survey (DLFS)	Administrative records	Continuous, nominal, ordinal	14	15,579
UK Annual Business Inquiry (ABI)	Quarterly Sample Survey	Continuous	26	6,233
Sample (1%) of Anonymised Records of UK 1991 Population Census (SAR)	Population Census	Nominal, ordinal	35	494,024
Swiss Environment Protection Expenditures (SEPE)	Yearly Sample Survey	Continuous	54	1,039
German Socio-EconomicPanel	Panel Sample Survey	Nominal, ordinal,	30	5,383
Survey (GSOEP)		Continuous		
Time Series for Financial Instruments (Shares and bonds, Options)	Time Series	Continuous	87 daily time series from 1995 to 1999	

For ABI, SARS, SEPE and time series , three different evaluation versions have been made available:

- Y^* containing true data, i.e. the dataset assumed to be complete and without errors;
- Y_2 containing data with missing values, but without errors;
- Y_3 containing data with both missing values and errors.

For DLFS and GSOEP only Y^* and Y_2 were produced.

Only versions Y_2 and Y_3 were given to partners for carrying out experiments.

Versions Y^*, considered as the "target" data, were not distributed by the project coordinator (ONS), with the exception of small subsets of data (near 10% of each dataset), necessary for some methods, as neural networks, that require "training" datasets to estimate internal parameters.

Together with datasets, also edit rules currently used by owners were disseminated to partners.

Versions Y_2 and Y_3 were produced by perturbing original Y^* in the following way:

a. missing values were generated by adopting a missing completely at random (MCAR) non-response mechanism;
b. errors were generated trying to simulate the way they occur during the compilation of the questionnaire or the data entry operations.

The percentages of missing values and errors for each variable were determined as much as possible on the basis of real situations verified in previous experiences.

Also *development* datasets were given to partners, in order to let them produce by themselves perturbed versions (also perturbation software was available), apply methods and evaluate their performance, to get valuable experience before the application to the evaluation datasets.

Partners applied each suitable method to different dataset according to the following rules:

1. each *error localisation method* was to be applied only to versionsY_3 of datasets (with both errors and missing values), while *imputation methods* were to be applied to both versions Y_2 and Y_3[6];
2. each partner could use the available subset of Y^* to train neural networks, or to estimate the parameters of a statistical model; for imputation methods, partners could use the complete subset of Y_2;
3. edit rules, given together datasets, could be (i) used by partners without modifications, (ii) with modifications, (iii) not used at all.

Once the different methods for error localisation and/or imputation of data were applied to the datasets, the corresponding outputs were given back to ONS, that provided to calculate the set of performance indicators illustrated in paragraph 3.

6 The rationale for the application of imputation methods to both versions was to test their robustness in the presence of errors.

5. Evaluation results

The experiments that were carried out are analysed from a double point of view: (i) for each dataset, the performance of the various methods that were applied to it are compared, and (ii) methods that revealed to be the best are highlighted.

5.1 Evaluation results bydataset

For any dataset, the different performance indicators will be grouped so as to analyse the following quality indicators:

a. "pure" *error localisation* performance: indicators from (1) to (3);
b. *influential error detection* performance: indicators from (4) to (8);
c. *difference between moments* of true and edited data: indicator (9);
d. *predictive accuracy*: indicators from (10) to (15);
e. *distributional accuracy*: indicators from (16) to (17b);
f. *estimation accuracy*: indicator (18).

5.1.1 Evaluation results in Annual Business Inquiry (ABI)

ABI dataset contain 26 variables organised in a three-level hierarchy: at the top level there are six economic variables and one employment variable. Each of these variables breaks down in a number of elements; for some of the latter there is another level with component variables. Most of the analysis that was carried out refers to the first level, including the six most important economic variables. A high number of error localisation and/or imputation methods were applied to this dataset. Up to 33 experiments involving Y_3 version, and 24 related to Y_2 version, were conducted by applying:

1. CHERRY-PIE plus multivariate regression and hot-deck imputation to Y_3 (CP-MRH);
2. multivariate regression and hot-deck imputation (MRH) to Y_2;
3. MLP to both Y2 and Y3;
4. Integrated Modelling Approach to Imputation (IMAI) to both Y_2 and Y_3;
5. Generalised Edit and Imputation System (GEIS) to both Y_2 and Y_3;
6. Self-Organising Maps (SOM) plus random draw form normal Probability Density Function (PDF), or MLP, or nearest neighbour (NN), or mean (MEAN) to both Y_2 and Y_3;
7. Donor Imputation System (DIS) to both Y_2 and Y_3;
8. Epidemic Algorithm plus POEM (EA-POEM) to Y_3;
9. Bacon plus EM algorithm plus POEM (BEM-POEM) to Y_3;
10. Transformed Rank Correlation plus POEM (TRC-EM-POEM) to Y_3;

11. Univariate robust tree modelling (UWAID with node mean or node nearest neighbour imputation) to Y_3;
12. Multivariate robust tree modelling (MWAID) to Y_3;
13. Univariate Forward Search plus Reverse Calibration Imputation (UFS-RCI), or Nearest Neighbour Imputation (UFS-NNI), or Linear and Log-linear Imputation (UFS-REG and UFS-LREG) to Y_3;
14. Correlation Matrix Memory (CMM) plus weighted mean or median to Y_3;
15. Support Vector Machines (SVM) to Y_2.

Once having standardised the quality indicators, if we consider the six most important economic variables in Y_3, and the *only error localisation* experiments, three methods obtained good values in all the three error localisation groups of indicators (a), (b) and (c), namely MLP, GEIS and SOM. In particular, MLP experiments obtained best results in groups (a) and (c) (pure error localisation performance and differences between moments), while SOM was the best in group (c) (influential error detection). If we consider the *only imputation* experiments, best performance in groups (d) and (e), i.e. predictive and distributional accuracy, was revealed by MLP. Finally, considering *both error localisation and imputation experiments*, good values in all the five groups (a)-(e) were shown by CP-MRH, SOM, UWAID and the set of UFS methods with the various imputation methods (RCI, NNI, REG and LREG). This latest set seems to achieve the absolute best values.

If we consider the Y_2 dataset, again for the six upper level economic variables, two methods rank above the others, namely MLP and MHR. In particular, MLP is the only method that achieves good results for all the considered indicators.

5.1.2 Evaluation results in UK Sample of Anonymised Records (SARs)

The evaluation here concentrated on six key variables, four concerning *individuals* (relation to head, marital status, sex and age), and two the *households* (number of rooms and presence of bathroom).

The methods that were applied to both Y_2 and Y_3 are CANCEIS-SCIA, MLP, SVM and, SOM, while DIS, CMM and IMAI were applied only to Y_2.

Starting with Y_3, if we consider individual variables and the first group of indicators related to the pure error localisation capability, for *alpha* values the best performance is shown by CANCEIS-SCIA and SOM; for *beta* values the best are CANCEIS-SCIA and SVM, while for the overall *delta* the best is always CANCEIS-SCIA.

If we consider the other indicators for the only continuous variable (age), MLP is the best for the influential error detection (root average error, RAE), while CANCEIS-SCIA shows the best performance for estimation accuracy (m_1 and m_2). Instead, Support Vector Machine (SVM) is the best for the preservation of true values (R^2 and d_{L2}), followed again by MLP and CANCEIS-SCIA.

Considering now Y_2, for variable age and indicator R^2 the best method is SOM (with random draws from normal PDF), while for d_{L2} is SVM. Again, CANCEIS-SCIA shows the best performance for estimation accuracy, together with IMAI.

5.1.3 Evaluation results in the Danish Labour Forces Survey (DLFS)

The peculiarity of the Danish Labour Forces Survey (15,579 observations) is that only the variable "income" contains missing values. The distributions of all other variables, categorical, are complete. This reflects a real situation, in which 27% of interviewees refused to respond to this question. The corresponding true values of non respondents can be found in administrative registers, so this is the only non simulated situation, in which it is possible to evaluate the imputation performance in the presence of a real non-response mechanism. The following methods have been applied:

1. MLP;
2. CMM (with different imputation methods: nearest neighbour, random neighbour, mean, weighted mean and median);
3. SOM (with nearest neighbour or random neighbour);
4. SVM (greedy or stratified);
5. IMAI (Regression Based Nearest Neighbour linear or log-linear, with or without noise);
6. Linear Regression;
7. Random Hot Decking;
8. DIS.

As for the *predictive accuracy*, MLP (with 20 neurons) shows the best values for slope (together with CMM and SVM), R^2, d_{L1}, d_{L2} and the MSE, followed by the Linear Regression.

In the *distributional accuracy* group of indicators, MLP is still among the best for KS(2), but SOM is the absolute winner for KS, KS(1) and KS(2).

As for *aggregate preservation*, SOM reveals to be the best for the preservation of the first moment (indicator m1), followed by MLP, while IMAI (log-linear without noise) is the best for the preservation of the second moment (indicator m 2).

5.1.4 Evaluation results in Swiss Environmental Protection Expenditures Survey (SEPE)

EPE data contains 1,039 observations and 54 variables. As in the case of ABI, there is a three-level hierarchy, where at the top level we can find the 4 most important key variables, that are totals of 20 variables, some of which are in turn totals of other 30 variables. Evaluation was carried out concerning the four highest level variables. These methods were applied to the Y_3 version:

1. CHERRYPIE plus multivariate regression plus ratio hot deck method (CP-MRH);
2. DIS;
3. Epidemic Algorithm plus POEM (EA-POEM);
4. Transformed Rank Correlation plus POEM (TRC-POEM);
5. Univariate WAID plus node mean imputation (UWAID);
6. CMM,

and these others to Y_2:

1. Multivariate regression plus ratio hot deck method (MRH);
2. Censoring;
3. SOM plus deterministic imputation or mean or random draw from normal PDF;
4. DIS;
5. CMM.

Considering methods applied to Y_3, there is no evidence of a method clearly doing better than the others in error localisation. On the contrary, the CP-MRH method ranks first with respect to the majority of imputation indicators.

Considering the Y_2 version of dataset, the overall good performances belong to methods MHR and SOM.

5.1.5 Evaluation results in German Socio-Economic Household Panel (GSOEP)

The GSOEP is a panel survey with six different waves, from 1991 to 1996. The dataset contains 30 variables, of which two can present missing values. Both are related to income: personal and household income. Because of the waves, we have up to 12 different variables to be imputed, six for personal income (from 91 to 96) and six for household income (again from 91 to 96). Imputation has been carried out by means of the following methods: SOM (with random draw from normal PDF), CMM (with 5 different imputation options: 2 real donor and 3 model donor), DIS and IMAI (using RBNN imputation method with a log-linear regression model without noise term).

For all quality indicators, IMAI always results to outperform the other methods. To explain this, it is worth while to remark that IMAI is the only method that made use of the panel characteristics of the survey. In fact, while all the other methods modelled auxiliary information on a cross-section basis, wave by wave, IMAI did so only for the first wave (1991): for next waves, information on previous values of income (individual and household), actual and imputed, was considered as auxiliary information, and added to the set of explicative variables in the models. In any case, even if we consider only the first wave, where this advantage for IMAI is not present, values of indicators still are in favour of the method, though less markedly. CMM is the second best, at least for personal income, while SOM is better for household income.

5.1.6 Evaluation results in Financial Time Series

Two datasets have been considered: one containing information concerning *shares and bonds* (daily prices for 51 time series from 1995 to 1999), and the other one related to *options* (36 time series of daily prices over the same period). These are the methods used for imputation:

1. Last Value Carried Forward (LVCF);
2. Multivariate regression imputation (R1) using stock market indicators and exchange rats as covariates;
3. Non-parametric multivariate regression imputation using a moving window of length 100 (NP100), with the same covariates than R1;
4. Multivariate autoregression imputation of lag1 (MARX1), with the same covariates than R1;
5. Univariate autoregression imputation of lag1-lag5 (ARX5), with the same covariates than R1;
6. Univariate multi-layer perceptron (MLP) imputation, with the same input considered in R1;
7. Black-Scholes pricing with cross sectional average imputation of missing volatilities (BSBASE);
8. Black-Scholes pricing with LVCF imputation of missing volatilities (BSLVCF);
9. Black-Scholes pricing with EM imputation of missing volatilities (BSEM);
10. Black-Scholes pricing with MLP imputation of missing volatilities (BSMLP).

The first six were applied to bonds and shares dataset, while the last four were experimented on options dataset.

LVCF is a somehow naïve method consisting in replying for a missing value in the series the more recent value observed for the same unit.

Methods (2) and (6) are not peculiar of time series context. Methods (3) to (5), on the contrary, are based on time relationships among observations.

Black-Scholes is a pricing formula, well known and widely used in financial institutions. The price of an asset at time t is dependent on a set of entities: all of them are usually available, with the exception of the so called *volatility*. When a price is missing in a time series, also volatility is: so, to be able to use the Black- Scholes formula, it is necessary first to estimate volatility. This can be done by using a variety of imputation methods: cross-sectional averaging, last value carried forward, EM algorithm imputation, univariate MLP imputation.

For each dataset, also in this case two versions were considered: one with only missing, and one with missing and errors.

As for shares and bonds, considering the dataset with only missing, LVCF is the worst method (essentially in terms of predictive accuracy), while NP100 is slightly better than the others. But if we consider the dataset version with also errors, we have exactly the opposite situation: LVCF becomes the best method (followed by ARX5), while NP100 results to be the worst.

Considering the options dataset, BSLVCF and BSMLP are best methods for imputing missing data. This is true for both versions of this dataset.

As a general conclusion, it can be said that methods that work on lagged variables are better than those exploiting cross-sectional information.

5.2 Best methods

On the basis of previous analyses, we tried to individuate best methods inside those selected to be investigated in the EUREDIT project. It is important to underline the fact that the concept of *best* is sometimes very relative, as performance for a given method may vary accordingly to the considered (groups of) indicators and subsets of variables. Very seldom a method outperforms all others in all possible situations.

Among *standard methods*, we can say that CANCEIS-SCIA revealed the best performance for categorical data, both for error localisation and imputation. CHERRY PIE was the best for error localisation in continuous data; for imputation, multivariate regression plus hot deck method showed the best results, followed by IMAI predictive mean matching method.

Among *neural network based methods*, MLP applications always stand in the first positions, both for error localisation and imputation, followed by TS-SOM.

In the class of *robust methods*, univariate forward search (BACON) for outlier detection outstands as the best. In association with imputation methods as reverse calibration and nearest neighbour, this method is the best also for imputation of both missing data and errors in continuous data. As second best, univariate WAID obtains comparable results in this class.

6. Conclusions

Experience made in the EUREDIT project led us to say that there is no "best" method, in the sense that no method works best in all situations. In addition, for a given situation, i.e. for a given typology of survey, the procedure for error localisation and imputation can hardly be constructed by utilising a simple method: very often, it will be a *complex* procedure, composed by different steps, possibly involving various methods, accordingly to the various nature of errors to be dealt with, and the different non-response mechanism.

Therefore, the value added of the EUREDIT project is not only (and not prevailingly) in the final indications concerning best methods to be used for different typologies of data (the *winners*). It is rather in the methodological path that was followed in its activities, that can be replied by anyone in order to *continuously* improve editing and imputation procedure. This path can be summarised as follows:

i. for any typology of data of interest, individuate candidate methods for error localisation and imputation;
ii. define a set of indicators useful to evaluate the performance of selected methods;
iii. adopt a simulation approach, by introducing missing values and errors in data in a controlled way so to replicate real situations;
iv. develop procedures containing selected methods and apply to data;
v. evaluate and compare results in order to choose best methods.

Another lesson learnt is in the fact that the more *information* related to (i) data structure, (ii) error nature and (iii) missing data patterns you can introduce in the procedure for error localisation and imputation, the more you can obtain in terms of accuracy of the results. This means that a lot of analysis of these three elements is needed. This job can be done only by expert statisticians, and cannot be delegated to naïve users: it is not just a matter of applying software to data.

Nevertheless, the availability of software is a crucial aspect: some of the investigated methods are so complex that a corresponding software is very costly to develop. So, a value added is also in the software that will be made available to EUREDIT partners and to external users: a software incorporating all robust methods and some of the neural network methods; and also a software useful for the evaluation process, to simulate missing and errors in data, and to produce evaluation indicators. Other software, especially rule-based standard software developed by national statistical institutes, is already available on demand.

The activity of the EUREDIT project will be hopefully continued in the VI Framework European Research Programme. One of the first objective of future work will be the creation of a *knowledge base* containing all the information related to the different methods and tools: methodological and operational aspects, suitable typologies of data, performance.

References

Bankier, M., M. Lachance, and P. Poirier. 2000. *2001 Canadian Census Minimum Change Donor lmputation Methodology*. Work Session on Statistical Data Editing, UNECE Cardiff UK.

Barnett, V., and T. Lewis. 1994. *Outliers in Statistical Data*. New York: John Wiley.

Beugin, C., and B. Hulliger. 2001. *Detection of Multivariate Outliers by a Simulated Epidemic*. Proceedings of ETK/NTTS 2001 Conference, EUROSTAT, pp. 667-676.

Billor, N., A.S. Hadis, and P.F. Velleman. 2000. *BACON: Blocked Adaptive Computationally Efficient Outlier Nominators*. Computational Statistics and Data Analysis.

Bishop, M.C. 1995. *Neural Network for Pattern Recognition*. Oxford: Oxford Clarendon Press.

Breiman, L., J.H. Friedman, R.A. Olshen, and C.J. Stone. 1984. *Classification and Regression Trees*. Belmont, CA, U.S.: Wadsworth International Group.

Chambers, R. 1986. Outlier Robust Finite Population Estimation. *Journal of the American Statistical Association*, N. 81, pp. 1063-1069.

Chambers, R., J. Hoogland, S. Laaksonen, D.M. Mesa, J. Pannekoek, P. Piela, P. Tsai, and T. De Waal. 2001. *The AUTlMP Project: Evaluation of lmputation Software*. Research Paper 0122, Statistics Netherlands.

Cristianini, N., and J. Shawe-Taylor. 2000. *An lntroduction to Support Vector Machines*. Cambridge: Cambridge University Press.

De Waal, T. 2002. *An Algorithm for Consistent lmputation in Mixed Data*. EUREDIT Deliverable 5.1.1. Statistics Netherlands.

Dempster, A.P., N.M. Laird, and D.B. Rubin. 1977. Maximum Likelihood from lncomplete Data via the EM Algorithm. *Journal of the Royal Statistical Society,* B 39, pp. 1-38.

Deville, J.C., and C.E. Sarndal. 1992. Calibration estimators in survey sampling. *Journal of the American Statistical Association,* N. 87, pp. 376-382.

Di Zio, M., U. Guarnera, and O. Luzi. 2002. *GElS application on ABl data - Description of the applied editing methods*. EUREDIT Deliverable 5.1.1, Istat.

Di Zio, M., U. Guarnera, and O. Luzi. 2002. *GElS application on ABl data - Description of the applied editing methods*. EUREDIT Deliverable 4.1.1, Istat.

Gnanadesikan, R., and J.R. Kettenring. 1972. Robust Estimates, Residuals and Outlier Detection with Multiresponse Data. *Biometrics*, N. 28, pp. 81-124.

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

Hadi, A.S., and J.F. Simonoff. 1993. Procedure for the ldentification of Multiple Outliers in Linear Models. *Journal of the Royal Statistical Society*, B 56, pp. 393- 396.

Hidiroglou, M.A., and J.M. Berthelot. 1986. Statistical Editing and lmputation for Periodic Business Surveys. *Survey Methodology*, vol.12, N. 1, pp. 73-83.

Kohonen, T. 1997. *Self-Organising Maps*. Heidelberg: Springer-Verlag.

Koikkalainen, P. 1999. *Tree-Structured Self-Organising Maps. In Kohonen Maps* pp. 121-130. Amsterdam: Elsevier Science.

Kovar, J.G., J.H. MacMillian, and P. Whitridge. 1988. *Overview and Strategy for the Generalised Edit and lmputation System*. Report, Statistics Canada, Methodology Branch (updated February 1991).

Laaksonen, S. 2000. Regression-based Nearest Neighbour Hot Decking. *Computational Statistics*, N. 15, pp. 165-171.

Manzari, A. 2002. *Application of CANCElS and SClA to the UK SARs data. Description of the application.* EUREDIT Deliverables 4.1.1-5.1.1, Istat.

Nordbotten, S. 1995. Editing Statistical Records by Neural Networks. *Journal of Official Statistics*, Vol. 11, N. 4, pp. 391-411.

Nordbotten, S. 1995. Editing and lmputation by means of Neural Networks. *Statistical Journal of UNECE*, Vol. l3, N. 2, pp. 119-129.

Pannekoek, J. 2002. *(Multivariate) Regression and Hot-deck lmputation Methods*. EUREDIT Deliverable 5.1.1. Statistics Netherlands.

Riani, M., and A.C. Atkinson. 2000. Robust Diagnostic Data Analysis: Transformations in Regressions. *Technometrics*, N. 42, pp. 384-398.

Riccini, E. 2002. *CONCORD User Guide*. Istat Internal Document.

Ripley, B.D. 1996. *Pattern Recognition and Neural Networks*. Cambridge: Cambridge University Press.

Scavalli, E. 2002. *Edit and lmputation Using MLP Neural Networks in SARs data*. EUREDIT Deliverable 4.3-5.3, Istat.

Vapnik, V. N. 1995. *The Nature of Statistical Learning Theory*. New York: Springer.

Yar, M. 1988. *The Development of the Donor lmputation System*. Technical report. Office for National Statistics, U.K.