STATISTICAL COMMISSION and
ECONOMIC COMMISSION FOR EUROPE

Working paper No. 35
English only

CONFERENCE OF EUROPEAN STATISTICIANS

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

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

FIRST RESULTS FROM THE EUREDIT PROJECT – EVALUATING METHODS FOR DATA EDITING AND IMPUTATION

Contributed paper

Submitted by the Office for National Statistics (ONS), United Kingdom[1]

Abstract: This paper discusses the EUREDIT project, a large research project that involves twelve participating organisations in seven countries over a period of three years. Its aims include developing new methods that are faster, more efficient, and more flexible, and also a methodology for comparing methods so that informed choices can be made on the most appropriate methods to be used in a particular situation. A range of statistical criteria have been developed for comparing different edit and imputation methods that are appropriate for different types of analyses and data. The project and its methods are described, and progress is reported. The project has now entered its final year and some 32 papers on interim results were presented at the project meeting in March 2002. Experimentation on individual methods will continue until July, and subsequently methods will be compared according to the established criteria, with conclusions published in Spring 2003.

I. INTRODUCTION

1. Current imputation methods are typically based on simple statistical ideas such as nearest neighbour, but little is known about the comparative performance of each method across the wide variety of data sources used. For the purposes of this paper we define "edit" as error localisation, i.e. identifying doubtful or erroneous data values. Once incorrect (or missing) values have been identified they will need to be corrected. Imputing new values is often preferred over the alternative of re-weighting because of its simplicity and because imputation provides a dataset that can easily be used by many different future users.

2. Research and development is needed to identify edit and imputation (E&I) methods:

- That improve the efficiency of the edit/ imputation process compared with current methods
- That improve the quality of the edit/ imputation process compared with current methods
- Are faster than current methods
- Can cope with complex data structures that are difficult to specify in terms of simple edit rules
- Can deal with mixtures of discrete and continuous data
- That ensure consistency with all edit rules specified
- That limit edits selectively to those that really affect data quality (e.g. automatic macro-editing algorithms/ significance editing)
- Incorporate auxiliary data in edit and imputation (e.g. previous survey or administrative data)

[1] Prepared by John Charlton (john.charlton@ons.gov.uk).

3. Research is also required to:

- Investigate, develop and evaluate new methodologies that have a bearing on E&I
- Assess robustness of different imputation methods and their effects on outcomes including multivariate analyses of data with imputations
- Identify the merits of different E&I methodologies for different data/ analysis requirements
- Establish when to automatically impute/ contact data supplier when edit failures occur

4. Advances in methods and computing capabilities have made possible the application of more complex statistical modelling techniques. There have been developments in statistics including: outlier-robust statistical methods (Chambers 1986; Chambers and Kokic 1993); non-parametric regression (Breckling and Sassin 1995); neural networks, including multi layer perceptron (Nordbotten 1995), correlation matrix memory (Austin 1996; Austin and Lees 1999), self-organizing maps (Kohonen 1989), and support vector machines (Vapnik 1996). These new approaches will be discussed briefly in the presentation. The above methods are all currently being evaluated in a project called EUREDIT.

II. THE EUREDIT PROJECT

A. Project objectives

5. We describe here the research undertaken in a large multi-national collaboration (see http://www.cs.york.ac.uk/euredit), involving twelve partners from seven countries, largely funded by the European Commission and aimed at meeting many of the needs mentioned above. EUREDIT will combine recent developments in statistical and computer science to develop and evaluate novel edit and imputation methodologies, focusing on the use of new statistical, neural network and related methods for edit and imputation in large-scale statistical data-sets. The project should establish a general framework in which new E&I methods, both within the project and beyond, can be evaluated in comparative terms, so that the choice of appropriate methods, depending on data type, error types, and intended application, should be easier for users in the future. The study will be based on real data and real problems encountered in official statistical data. The project has the following objectives:

i) To establish a standard collection of data sets for evaluation purposes
ii) To develop a methodological evaluation framework and develop evaluation criteria
iii) To establish a baseline by evaluating currently used methods.
iv) To develop and evaluate a selected range of new techniques.
v) To evaluate different methods and establish best methods for different data types.
vi) To disseminate the best methods via a software CD and publications.

6. EUREDIT is organized in 9 work-packages (WPs). These have the following main tasks:

i) Project management
ii) Selection and compilation of datasets for evaluating methods
iii) Determining objective quality criteria for evaluating methods
iv) Develop and test selected new methods for error location
v) Develop and test selected new methods for imputation
vi) Evaluation and validation of results from WP 4 and WP 5.
vii) Integration of methods into a package for wider dissemination
viii) Dissemination and exploitation
ix) Project evaluation (internal)

WP 4 and 5 are subdivided into 5 and 7 sub-work-packages, in which different new methodological approaches are studied by teams of interested partners.

B. Participants

7. Partners in the EUREDIT project include national statistics institutes (NSIs), universities and private organizations. The partners should represent a comprehensive knowledge of statistical production, research in statistics and areas of emerging technologies, computer implementation of methods evaluated and recommended, and end users. The NSO's partners are the Office for National Statistics (ONS) UK, Statistics Netherlands (CBS), Statistics Finland, Swiss Federal Statistical Office (SFSO), ISTAT, Italy, and Statistics Denmark (DSt). The universities represented are: University of Jyvaeskylae – Finland; Royal Holloway and Bedford New College – University of London, University of Southhampton, UK, and University of York, UK. The Numerical Algorithms Group (NAG), UK, and Qantaris GmbH are the partners representing computer implementation and end users.

C. Investigation of currently used methods

8. In both WP 4 and WP5 it is considered important to establish benchmarks to which the new methods can be compared and evaluated. An effort is made to obtain access to the E&I methods currently used and considered successful. Even though these methods are in current use, they have to be adapted for use with the standards and formats required in EUREDIT.

9. Among the methods/packages considered are:

- Canadian Edit system (NIM/ CANEDIT, GEIS) from Statistics Canada.
- Agricultural Generalized Imputation and Edit System (AGGIES) from US NASS.
- Cherry-Pie from CBS in the Netherlands.
- Donor Imputation System (DIS) from the UK ONS.
- Automatic Control and Imputation System (SCIA) and CONCORD from Italian ISTAT.
- Multivariate regression/ classification trees and MCMC for imputation
- SOLAS, SAS, NORM, SPSS MVA, including multiple imputation.

Some methods may work well in special fields, other will work best in other. It will therefore be necessary to have several benchmark methods for the evaluation of the new methods.

D. Development of new methods

10. The new methods, which will be implemented and tested, include:
(A) Multivariate robust methods;
(B) Multi-layer perceptron (MLP);
(C) Correlation matrix memories (CMM);
(D) Self-organising maps (SOM);
(E) Support vector machines (SVM); and
(F) New methods for panel data and time series.
In EUREDIT, all new and standard methods are tested on standard data sets so results can be compared.

(A) Multivariate robust methods

11. A very important aspect of statistical data editing is outlier detection. Besides graphical tools, which are of limited use in high dimensions, robust mathematical algorithms can be used to detect outliers. Imputation of continuous variables in the presence of outliers needs robust methods to give plausible results. The methods will address the following problems:

- Treatment of high dimensional datasets with continuous and categorical variables;

- Treatment of missing item values
- Distinction between non-representative and representative outliers in data (Chambers 1986; Chambers and Kokic 1993).
- Adaptation to the concept of sample surveys, taking account of survey weights
- Choice of level of aggregation that should be used for outlier detection - outliers may appear and disappear according to which reference population is used
- Error localisation within continuous variables and between categorical and continuous variables

12. New methods for outlier detection, robust imputation and outlier robust estimation have been developed and implemented and are now being evaluated. These include methods using a Mahalanobis distance with a **robust covariance estimator**. Simple covariance estimator methods should be directly computable without iteration or searches. They should be particularly useful as basic steps for tree-based methods. The newly developed SMP method takes bivariate Spearman correlations as starting point (cf. Gnanadesikan and Kettenring 1972) and ensures positive definiteness of the resulting covariance matrix by back-transforming the diagonal matrix of robust scales in the direction of the principal components. Complex covariance estimator methods are based on some iterative or other high- cost robust computation of a covariance matrix and include:

- Minimum covariance determinant
- Modified Stahel-Donoho estimator (Stahel 1982, Donoho 1982, Franklin et al. 2000) - an iterative method based on projection techniques.

13. Another approach is based on **growing of a good subset** of the data using iterative computations of the empirical covariance matrix on a subset of non-outlying observations. These methods end with the partition of the dataset in two parts, good points and outliers:

- Kosinski algorithm (Kosinski, 1999 and De Boer and Feltkamp, 2000). Here several small subsets are selected as starting points of a two levels iterating algorithm..
- BACON, Blocked Adaptive Computationally efficient Outlier Nominators (Billor, Hadi and Velleman 2000). Here one initial subset is selected and an the algorithm adds and excludes points according to an outlier criterion.
- Atkinson forward search starts with a good subset and adds one point at a time.

14. **Data depth methods** are based on some notion of data depth (and have high computation costs). They include:

- Methods using simplicial depth (Liu 1990, 1999) or a similar depth with high computation cost.
- Multivariate M-quantiles (Breckling and Chambers 1988).
- A stochastic type of data depth is realised by the Epidemic Algorithm (EA). The EA simulates an epidemic starting from a robust center. Outliers are detected by their late infection (Hulliger and Béguin 2001).

15. **Tree-based methods** are based on classification and regression trees that incorporate robust measures and weighting. They include:

- Outlier detection using WAID's (weighted automatic interaction detection) classification/regression tree methodology to build a robust "tree" that explains the sample distribution in terms of categories defined by some covariate known for the whole population. Outliers will be detected as early formed and small terminal nodes in the tree. The splitting criterion for the forming of child nodes is robust. For more detail see Chambers 2000.

16. The search for an optimal partition intends to determine the best reference population in the sense that an overall discrepancy measure is minimised over possible partitions based on a (categorical) covariate X. The partitioning is done recursively, e.g. by WAID.
Error localisation in a mixed set of variables (continuous Y and categorical X) tries to compare the reduction in outlyingness of the continuous variables Y of an observation when we change the observation to a

different reference population than the one it belongs to according to its covariates X. If the reduction in outlyingness is large compared with the change in the covariates we may decide that the covariates X are in error instead of the variables Y.

Imputation methods

17. The **Winsorization** approach assumes that there is a region A of acceptable values for Y. An outlier is then replaced by the value of A which is closest to it. A variant is that we do not have a region but a set of observations that have acceptable values of Y. Then we choose the observation closest to the considered outlier for imputation. This is a **Nearest Neighbour** method where donors belong to the non-outlying observations. We may relax the definition of closeness to include a set of acceptable observations as candidates (donors) for imputation to a certain outlier. Then we may impute from this set randomly, eventually with varying probabilities. The case where the variables X are considered in error can be subsumed here. Then the X-values of observations of the reference population will be imputed instead of Y values. Again the distance to the original X may play a role in the definition of the set of donors or in the probability to choose a donor. This is a random nearest neighbour imputation with.

18. **Robust estimation and reverse calibration** assume that there is a particular estimation problem, maybe multivariate, and that we have a robust estimate. Note that a robust estimate solves the problem of outlier detection and (implicite) imputation at the same time. To turn the estimate back into an imputation method we stick to the robust estimate and look for values of Y which, together with the weights, yield the robust estimator as a weighted mean (See Chambers 2000). As estimation methods robust calibration (Duchesne 1999), M-estimators (Chambers 1986), adaptive censoring (cf. Searls 1966) and adaptive winsorisation (Chambers, Kokic *et al* 2000) are considered.

(B) Multi-layer perceptrons (MLP).

19. The multi-layer perceptrons (MLP) or feed-forward neural networks (NN) can be considered as very primitive models of biological neural networks. From being a tool mainly of interest for a few groups in biology and physics, mathematicians and statisticians have more recently become aware of MLP. In statistics, a number of articles and books have been devoted to the relationship for instance between MPL and multivariate theory (Bishop 1995). It has so far been tested out as an E&I method on several data sets (Nordbotten 1995, 1996, 1999). An MLP application is based on a network trained with a small representative sample of statistical records that have been edited as well as possible by human experts.

20. Assuming that if a large number of human experts were available at no cost, human editing would give the best results and be preferred. As less expensive substitutes, traditional, automatic editing methods are based on retrieval and formalisation of the knowledge of experts. Experience indicates that this kind of 'knowledge engineering' is difficult because people have problems in expressing their expertise.

21. Application of MLP networks takes another approach. The experts are asked to use their expertise on a small sample of records, and the networks are trained to imitate the experts. The trained networks can later be applied for editing as well as imputation of all records. In addition to providing training material, the small edited sample can also be sub-divided into a training sub-sample used for training and a separate test sub-sample used to estimate imputation errors. Both samples will finally be an important source for knowledge about where and how errors are generated.

(C) Correlation matrix memories (CMM)

22. AURA models are used to here to implement a *k*-nearest neighbour (*k-NN*) approach to imputation. The implementation of the *k-NN* using AURA technology has been described in (Zhou et al 1999) and (Zhou and Austin 1998). It has been demonstrated that:

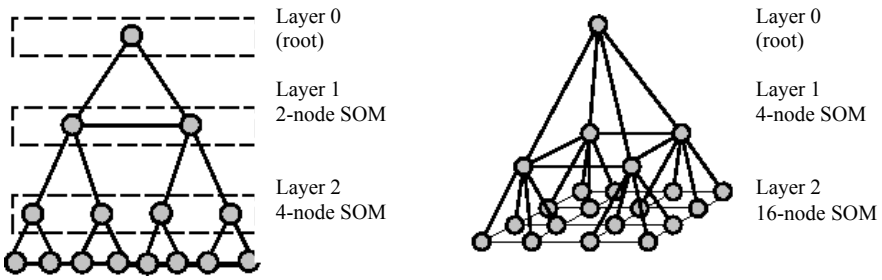
- The results obtained by a statistical *k-NN* technique can be reproduced using a neural network implementation of the technique.
- The speed can be improved significantly

23. The significant difference between the MLP and the AURA techniques lies in data space reduction. MLP systems attempt to reduce the dimensionality of the data space and represent it by a number of simpler functions. This model is then used to reproduce the output values. AURA models, however, attempt to learn all of the training data, albeit in a compressed form. The difference is obvious when the output function is highly non-regular – the MLP will over-smooth the output space, and produce estimates based only on the gross structure of the output function. The AURA technique retains the detailed structure of the output function and will return the closest points in the known, detailed output space to any new input. The corollary of this is that it may be poor at generalising outside the space spanned by the training data. These results – the closest neighbours in the output space – still need to be combined in some way to give the required result. The parameter *k* tells the system how large a neighbourhood around the test point should be combined to give the output. The appropriate value of *k* can be obtained by testing with a range of values. The sensitivity of the performance to the value of *k* is a measure of the regularity or complexity of the output space.

(D) Self-organizing maps (SOM)

24. Early versions of the self-organising map (SOM) were developed by Kohonen (1982). It is closely related to principle curves (Hastie 1989) and surfaces (Tibshirani 1992). The princpal approach adopted by EUREDIT is a tree-structured variant of SOM (TS-SOM), using software developed by the University of Jyvaskyla, Finland, called Neural Data Analysis (NDA) (Häkkinen 2001; Kohonen1997; Koikkalainen, P. and Oja, 1990). The basic SOM defines a mapping from the input data space $\boldsymbol{R}^n$ onto a latent space consisted typically of a two-dimensional array of nodes or neurons. The Tree-Structured Self-Organizing Map is made of several SOMs arranged to a tree structure (see Figure 1). The topmost layer ($L = 0$) has one neuron. Layer 1 has four neurons in two-dimensional and two neurons in one-dimensional case. Thus, each neuron has its own associated subgroup of data, four subgroups on layer 1 but one group, the data set itself on layer 0. The complexity of the SOM model is controlled by the degree of smoothing – without smoothing every data point in the training set will be included in the map, and with infinite smoothing SOM approximates the largest principal component subspace of the data. In TS-SOM the number of neurons allowed determines the degree of smoothing.

Figure 1. Illustrations of one and two-dimensional TS-SOM structures.



25. Editing (localization and correction of errors) can be logical (check conflicts with logical edit rules) or model-based (find observations that cannot be explained by the model). Models can be built from clean data or using robust training algorithms. The clean part of the data is assumed simpler than the erroneous

part. After a certain level of complexity the model is expected to separate clean data from errors. Error detection of a single observation is done such that the model classifies the observation space into possible and not possible observations. There is also distribution detection where the projection of the data is compared with the distribution for clean data – significant differences indicate errors.

26. A natural approach to imputing the missing values is to impute values within the clusters located by associated neurons. A simple starting point is to derive analogous processes from the classical imputation methods. Nearest neighbour imputation can be made by filling missing components of the data vector from the nearest data vector within the same cluster. Group means imputation (replacing the missing value by the average value of the observations belonging to the same class/subgroup) can be made by taking the centroid of the cluster, $\boldsymbol{m}_b$, and replacing the missing component j of the data vector $\boldsymbol{x}_i$ by corresponding $\boldsymbol{m}_b(j)$, which is actually the fastest way to impute. More complex, regression based, imputation modelling can also take advantage of TS-SOM mapping. (See paper 29 for early results)

(E) Support vector machines (SVM)

27. The support vector machines (Vapnik 1996) is a new tool for prediction and function estimation. Given a training set of input-output pairs $(\mathbf{x_1},y_1)$, $(\mathbf{x_2},y_2)$ … $(\mathbf{x_n},y_n) \in \boldsymbol{R}^n \times \{\pm 1\}$, the SVM algorithm estimates a function f such that, for $(\mathbf{x},y)$ drawn according to the same distribution, $P(\mathbf{X},Y)$ as the training set, $f(\mathbf{x}) = y$. The SVM can be adapted to perform multi-class classification ($y \in \{1,2,..N\}$) and regression ($y \in \boldsymbol{R}$). The SVM presented here is an extension of the perceptron algorithm. The perceptron learns a linear discriminant function, $f(\mathbf{x}) = \text{sign}(+\mathbf{w} \bullet \mathbf{x_i}, + \mathbf{b})$. The SVM extend this algorithm in two respects. It introduces non-linear decision surfaces, and a means of avoiding overfitting, The latter will be explained below. The former is achieved through a non-linear projection of the data into a higher dimensional feature space prior to estimation of the linear discriminant. The linear model learned in this space is equivalent to a non-linear model in the input space. For example, a point X in $\mathbf{R^3}$ with position, $(x_1,x_2,x_3)^t$ could be projected to the new space $\mathbf{R^5}$ with coordinates $\mathbf{X'} = (x_1,x_2,x_3,c_1x_1^2,c_2x_2^2)^t$. The SVM algorithm finding a linear discriminant function in this feature space is equivalent to the estimation of a polynomial discriminant in the original space $\mathbf{R^3}$. The second extenstion of the perceptron algorithm concerns capacity control or regularisation. The SVM achieves good generalisation by choosing a discriminant function that maximally separates the two classes in the feature space. The euclidean distance between the closest point and the decision surface is known as the *margin.* Maximising the margin acts as a form of regularisation. This is due to constants c_i that are associated with the added dimensions. (c_1 and c_2 in the example above). Their effect is to penalise discriminants that exploit the new features. This SVM algorithm can be formulated in such a way that it only requires the calculation of the dot product $+ \bullet ,,$ between training points. Moreover in test or prediction phase, test points also only occur as dot products: $f(\mathbf{x}) = \text{sign}(+\mathbf{w} \bullet \mathbf{x_i}, + \mathbf{b})$. An algorithm with this feature is known as having a 'dual form'. The SVM algorithm exploits the dual form by finding functions that perform the non linear projection described above, and the dot product in one step. Thse 'kernel functions' K(x,y) equate $+N(\mathbf{x_i}) \bullet N(\mathbf{x_j}),.$ The positions of the points in the feature space are in fact never calculated. There are many choices of kernel function, some of which have implicit feature spaces of infinite dimension. Such feature spaces provide a large number of models. Maximising the margin however is able to effectively choose the model with lowest capacity.

28. Data sets where missing variables are correlated with observed ones may be handled effectively by this tool, as long as enough training data (fully observed pairs $(\mathbf{x},y)$) can be extracted. However, if no such relationship exists it is likely that prediction tools such as the SVM, or the multilayer perceptron (MLP) will approximate a mean imputation.

29. Variance in the variable will then be underestimated, and the marginal distribution will not be preserved. Simpler methods, such as donor imputation, would be better suited to such data structures.

This extended perceptron algorithm can be reformulated as a convex quadratic programme. Such QPs are well understood and efficient methods exist for solving them.

(F) New methods for panel data and time series

30. Multivariate M-quantiles for outlier identification are being investigated for time series data. The objective in this part of the EUREDIT project is to develop a method for non-parametric multivariate outlier detection that has a probabilistic interpretation under an appropriate multivariate distribution model, but which is robust against reasonable depatures from this model. The method under consideration is based on multivariate *M*-quantiles, (Breckling and Chambers1988; Breckling, Kokic, and Lübke (2001). Each *M*-quantile, θ_{rp}, is defined in terms of a unit direction vector, r, and a value $0 \leq p \leq ½$, where values of p close to 0 lie in the extremities of the data, and those with p close to ½ lie in the centre. By inverting this procedure a p-value can be determined for each point in a data set: - hence those points with small p-values can be consider as outlying. However, due to the highly nonparametric nature of the multivariate *M*-quantiles all points on the convex hull of the data set will be automatically identified as outlying ($p = 0$), which is an undesirable result. To solve this problem we begin at the other extreme by making a specific distributional assumption about the underlying data. We then introduce a parameter that determines the degree to which this distributional assumption enters the final estimating equations or alternatively, the degree to which departures from the distribution assumption can be tolerated. In the process a mapping is obtain from p to a probability-based measure of "outlyingness" under the assumed distribution. A procedure, somewhat similar to ridge parameter selection in ridge regression, is proposed for the selection of the mixing parameter. This approach permits the greatest departures from the base distribution feasible while overcoming the problem described above.

31. A number of methods are being considered for imputation for panel time series data. The objective in this case is to investigate the performance of a range of increasingly sophisticated imputation methods for panel time series data. They are being tested against a financial data set, which consists of a mixture of time series for US and UK shares, bonds and simple share options. Special techniques for financial time series are also being examined in this work. Amongst the basic approaches considered, essentially for the purposes of benchmark comparisons with more sophisticated approaches, are the methods of last-value carried forward, linear interpolation, Black-Scholes pricing, and standard term structure pricing of bonds. The new methods being investigated include univariate and vector ARMA, linear and non-parametric regression and multilayer perceptron models for imputation. Since most of these methods utilise other time series as covariates, which themselves contain missing observations, the EM algorithm (Dempster, Laird and Rubin, 1977) is an appropriate tool. This algorithm considerably simplifies the process of estimating the parameters in the underlying model through maximum likelihood techniques (in the M-step). Nevertheless it is still necessary to develop approaches that fully utilise all observed information (in the E step) since, unlike the forecasting problem in time series, data is usually available both before and after the missing observations. Thus the development of appropriate methods and algorithms for each model is a major undertaking in this part of the EUREDIT project.

E. Experimentation based on standard datasets

32. The standard evaluation datasets comprise: the UK Census (1 percent sample of anonymised household records); stock price time series; Danish registry data linked to their labour force survey; the UK Annual Business Enquiry; a Swiss Environmental Protection Expenditure survey; and the German Socio-economic Panel Survey (GSOEP).

F. Evaluation - Determining the best E&I methods for different situations

33. A key aspect of the EUREDIT project will be its focus on evaluation of the methodologies for edit and imputation that will be investigated. Work on the statistical evaluation criteria is now complete, and was presented at the 2000 UN/ECE meeting (paper 5 - see EUREDIT web site for fuller version). Here we briefly describe the evaluation criteria adopted by EUREDIT in order to achieve this aim. In doing so, we need to distinguish between criteria aimed at assessing editing performance and those developed to assess imputation performance.

34. **Editing** can be of two different types, *logical* (pre-defined rules must be obeyed) and *statistical* (a value is unlikely – it might be wrong). Here we shall be concerned with evaluation of overall editing performance (i.e. detection of data fields with errors). This is not the same as *error localisation*, which corresponds to deciding which of the fields in a particular record that "fail" the edit process should be modified. The key aspect of localisation performance is finding the "smallest" set of fields in a record such that at least one of these fields is in error. The localisation performance will be measured using a Gini-type index provided the editing procedure produces a "score" corresponding to the probability that a field is in error. The main focus of EUREDIT, however, will be on error detection. In this context, two performance criteria will be investigated: *efficient error detection* (ability to detect as many errors as is feasible) and *influential error detection* (ability to detect the errors that would lead to significant errors in the analysis unless detected). Efficient error detection can be evaluated in terms of both the number of errors correctly identified and the number of incorrect detections made. In this context, let Y_{ij} denote the observed value for the j^{th} item and the i^{th} sample case. The corresponding "true" value is Y_{ij}^*. The editing process itself is characterised by a set of variables E_{ij} that take the value one if the measured value Y_{ij} passes the edits (Y_{ij} is acceptable) and the value zero otherwise (Y_{ij} is suspicious). For each variable j we can therefore construct the following cross-classification of the n cases in the dataset:

	$E_{ij} = 1$	$E_{ij} = 0$
$Y_{ij} = Y_{ij}^*$	n_{aj}	n_{bj}
$Y_{ij} \neq Y_{ij}^*$	n_{cj}	n_{dj}

Then $\hat{\alpha}_j = n_{cj}/(n_{cj}+n_{dj})$ is the proportion of cases where the value for variable j is incorrect, but is still accepted 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. Similarly $\hat{\beta}_j = n_{bj}/(n_{aj}+n_{bj})$ is the proportion of cases where a correct value for variable j is judged as suspicious by the editing process, and estimates the probability that a correct value is incorrectly identified as suspicious. Finally, $\hat{\delta}_j = (n_{bj} + n_{cj})/n$ is an estimate of the probability of an incorrect outcome from the editing process for variable j, and measures the inaccuracy of the editing procedure for this variable. A good editing procedure would be expected to achieve small values for $\hat{\alpha}_j$, $\hat{\beta}_j$ and $\hat{\delta}_j$ for all p variables in the data set.

35. Turning now to efficient influential error detection, the aim here is not so much to find as many errors as possible, but to find the errors that matter (i.e. the influential errors) and then to correct them. From this point of view the size of the error in the measured data (measured value - true value) is the important characteristic, and the aim of the editing process is to detect measured data values that have a high probability of being "far" from their associated true values. In this context, one can view the editing procedure as leading to a set of post-edit values defined by $\hat{Y}_{ij} = E_{ij}Y_{ij} + (1-E_{ij})Y_{ij}^*$. The key performance criterion in this situation is the "distance" between the distributions of the true values Y_{ij}^* and the post-edited values $\hat{Y}_{ij}$. The aim is to have an editing procedure where these two distributions are as close as possible, or

equivalently where the difference between the two distributions is as close to zero as possible. A number of measures of this difference can be calculated, based on the moments and distribution of the post-edited errors $D_{ij} = \hat{Y}_{ij} - Y^*_{ij} = E_{ij}(Y_{ij} - Y^*_{ij})$, see section 2.4.1 of Chambers (2001). When survey weights are available, these measures are also weighted.

36. Note that statistical outlier detection is also a form of editing. As with "standard" editing, the aim is to identify data values that are inconsistent with what is expected, or what the majority of the data values indicate should be the case. However, in this case there are no "true" values that can be ascertained. Instead, the aim is to remove these values from the data being analysed, in the hope that the outputs from this analysis will then be closer to the "truth" than an analysis that includes these values (i.e. with the detected outliers included)

37. Finally, it is important to note that in many cases the measures described above will vary across subgroups of the data. An important part of the evaluation of an editing procedure will therefore consist in showing how these measures vary across identifiable subgroups. For example, in a business survey application, the performance of an editing procedure may well vary across different industry groups.

38. **Imputation** is the process by which missing or suspicious values are replaced. Here we shall only concern ourselves with assessing the imputation of identifiable values, i.e. where we know the records in the data set that need to be imputed (e.g. because they were detected as incorrect by an edit process and set to "missing"). Ideally an imputation procedure should be capable of effectively reproducing the key outputs that would have been obtained from "complete data". Since this is impossible a number of alternative criteria for imputation performance have been identified. These are defined below in order of those that are hardest to achieve to those that are easiest, but not necessarily in order of desirability.

- Predictive Accuracy: The imputation procedure should maximise preservation of true values. That is, it should result in imputed values that are "close" as possible to the true values.
- Ranking Accuracy: The imputation procedure should maximise preservation of order in the imputed values. That is, it should result in ordering relationships between imputed values that are the same (or very similar) to those that hold in the true values.
- Distributional Accuracy: The imputation procedure should preserve the distribution of the true data values. That is, marginal and higher order distributions of the imputed data values should be essentially the same as the corresponding distributions of the true values.
- Estimation Accuracy: The imputation procedure should reproduce the lower order moments of the distributions of the true values. In particular, it should lead to unbiased and efficient inferences for parameters of the distribution of the true values (given that these true values are unavailable).
- Imputation Plausibility: The imputation procedure should lead to imputed values that are plausible. In particular, they should be acceptable values as far as the editing procedure is concerned.

Note that not all the above properties are relevant to every imputed variable. For example, ranking accuracy requires that the variable is at least ordinal, while distributional and estimation accuracy are identical when the imputed variable is not scalar.

39. Generally, imputation plausibility can be checked by treating the imputed values as measured values and assessing how well they perform relative to the statistical editing criteria described earlier. However, methods for evaluating imputation performance relative to the other criteria listed above depend on the scale of measurement of the variable being imputed. To illustrate, suppose the variable being imputed is categorical. The distributional accuracy of the imputation procedure can then be assessed using an extension (Stuart, 1955) of McNemar's statistic (without a continuity correction) for marginal homogeneity in a 2×2 table. This is

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

where **R** is the c-vector of imputed counts for the first c categories of the variable, **S** is the c-vector of actual counts for these categories and **T** is the square matrix of order c corresponding to the cross classification of actual vs. imputed counts for these categories. Under relatively weak assumptions about the imputation process (essentially providing only that it is stochastic, with imputed and true values independently distributed conditional on the observed data), it can be shown that large n distribution of W is chi-square with c degrees of freedom. Similarly, the predictive accuracy of an imputation procedure for a categorical variable can be assessed via the statistic

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

where $\hat{Y}_i$ denotes the imputed version of Y_i and Y_i^* is its true value. Provided we cannot reject the hypothesis that the imputation method preserves the marginal distribution of Y, we can estimate the variance of D by

$$\hat{V}(D) = n^{-1} - n^{-2}\mathbf{1}^t\{diag(\mathbf{R}+\mathbf{S}) - \mathbf{T} - diag(\mathbf{T})\}\mathbf{1} = n^{-1}(1\tilde{}\ D)$$

where **1** denotes a c-vector of ones. If the imputation method preserves individual values, D should be identically zero. For an ordinal scale variable the issue of ranking accuracy is important. In this case a weighted version of D

$$D = n^{-1}\sum_{i=1}^{n} d(\hat{Y}_i, Y_i^*)$$

can be used. Here $d(t_1, t_2)$ is the "distance" from category t_1 to category t_2. Large values of this statistic are indicative of weak ranking accuracy for an imputation procedure.

40. Finally, it can be argued that an editing and imputation system is essentially useless, no matter how excellent its statistical properties, unless it can be practically implemented. Consequently it is vital that any such system demonstrates its operational efficiency before it can be recommended by the EUREDIT project. In particular the resources needed to implement and maintain the system (both in terms of trained operatives and information flows) need to be spelt out. Comparison of different editing and imputation systems in this way is of necessity qualitative, but that does not diminish its importance.

41. The evaluation comparing all methods in different situations should be available by April 2003. There will also be a CD Rom available, containing prototype software and documentation.

III. RESULTS OF THE EUREDIT PROJECT

42. The first public result of the EUREDIT was the establishment of a website, http://www.cs.york.ac.uk/euredit . The partners will present their results in research papers and journal articles and there will be a final report comparing all methods according to the different types of data to which they were applied. Some 40 interim papers have already been prepared and are available for review by project partners. At this stage however not all methods have been applied to all evaluation datasets, and it is thus too early to draw valid comparisons. The Data-Clean2002 conference planned to immediately follow this workshop will disseminate and discuss preliminary findings – see http://erin.mit.jyu.fi/data-clean **.** Proceedings will be compiled from the presentations and discussions, and these will be made available for the statistical community. The EUREDIT project goals also include integration of E&I methods and implementation into a prototype software system. This will be available on CD-ROM with documentation. Some of the data used by the project will also be made available in anonymised form to permit external researchers to compare their methods with the EUREDIT methods. It is intended that the results will aid the identification of current best methods for different types of data.

IV. CONCLUSIONS

43. EUREDIT is an ambitious project involving over thirty researchers. By 2003 it should have evaluated the major new methods for statistical editing and imputation, established a methodology for future evaluations, and made a substantial contribution to knowledge of how different methods compare for different types of data.

References

Austin J, 1996. Distributed associative memories for high-speed symbolic reasoning. *Fuzzy Sets and Systems* 82: 223-233.

Austin J, Lees K, 1999. A novel search engine based on correlation matrix memories. *Neurocomputing* – special issue, Elsevier Science.

Bishop, M. C.(1995). Neural Networks for Pattern Recognition. Clarendon Press. Oxford. 1995.

Bailey S., Charlton J. (2000*), Draft procedures for the evaluation of edit and imputation methods*, EUREDIT.

Breckling J, Sassin O, 1995. A non-parametric approach to time-series forecasting, In ZEW-Wirtschaftsanalysen Band 5: *Quantative Verfahren im Finanzmarktbereich*, ed Schroeder Nomos Verlagsgesellschaft, Baden-Baden, 1995.

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

Breckling, J. and Chambers, R. (1988). M–quantiles, *Biometrika*, 75, 761–771.

Breckling, J., Kokic, P. and Lübke, O. (2001). A Note on Multivariate M-Quantiles. To appear in *Statistics and Probability Letters*.

Chambers, R.L., (2000), *Robust Editing and Imputation*, University of Southampton Research Plan for EUREDIT WP4.2/5.2.

Chambers R.L. (2001), *Evaluation Criteria for Statistical Editing and Imputation*, National Statistics Methodology Series No 28.

Chambers R.L., Kokic Ph., Smith P., Cruddas M. (2000*), Winsorization for Identifying and Treating Outliers in Business Surveys*, Proceedings of the ICESII conference, Buffalo.

Chambers RL, 1986. Outlier robust finite population estimation. *JASA* 81:1063-1069.

Chambers RL, Kokic PN, 1993. Outlier robust sample survey inference, Invited paper, *Proc. 49th ISI Session, Firenze, August 1993.*

De Boer P., Feltkamp V. (2000), *Robust Multivariate Outlier Detection. Report*, Statistics Netherlands, Voorburg.

Dempster, A.P., Laird, N.M. and Rubin, D.B. (1977). Maximum Likelihood from incomplete data via the EM algorithm (with discussion). *Journal of the Royal Statistical Society B*, 39, 1-38.

Depoutot, R. (1998*): Quality of International Statistics: Comparability and Coherence. Conference on Methodological Issues in Official statistics. Stockholm.*

Donoho D.L. (1982), *Breakdown Properties of Multivariate Location Estimators*, Ph.D. Qualifying Paper, Department of Statistics, Harvard University.

ECE/UN (1994*):Statistical Data Editing: Methods and Techniques. Volume No.1. United Nations. NY and Geneva.*

ECE/UN (1996*):Statistical Data Editing: Methods and Techniques. Volume No.2. United Nations. NY and Geneva.*

Franklin S., Thomas S., Brodeur M. (2000), *Robust Multivariate Outlier Detection Using Mahalnobis' Distance and a Modified Stahel-Donoho Estimator*, to appear in Proceedings of the ICESII conference, Buffalo.

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

Granquist, L. (1997*); The New View on Editing. International Statistical Review. Vol. 65. No.3. pp. 381-387.*

Häkkinen, E. (2001). *Design, Implementation and Evaluation of the Neural Data Analysis Environment*. PhD thesis, Diss. Jyväskylä Studies in Computing. University of Jyväskylä, Finland.

Hastie T ND Stuetzle W (1989). Principle curves. *JASA* 84: 502-516.

Kohonen, T. (1997). *Self-Organizing Maps*. Springer, Berlin, Heidelberg.

Hulliger B., Kassab, M. (1998), *Evaluation of Estimation Methods for the Survey on Envirionment Protection Expenditures of Swiss Communes*, Technical Report, Swiss Federal Statistical Office, Bern.

Hulliger B. (1999), *Simple and Robust Estimators for Sampling*, Proceedings of the Survey Research Methods Section of the American Statistical Association.

Hulliger, B. (2000), *A simple algorithm for a robust covariance matrix estimator*, Internal Note, Swiss Federal Statistical Office.

Hulliger B, Beguin C (2001). Detection of multivariate outliers by a simulated epidemic. In *Proceedings of the ETK/NTTS Conference*, pp667-676 Eurostat.

Kohonen T, 1989. *Self-Organisation and Associative Memory*, (Third Edition), Springer Series in Information Sciences, Springer-Verlag.

Kohonen, T. (1997). *Self-Organizing Maps*. Springer, Berlin, Heidelberg.

Koikkalainen, P. and Oja, E. (1990). Self-Organizing Hierarchical Feature Maps. In *Proc. IJCNN-90-Wash-DC, Int. Joint Conf. on Neural Networks*, volume II, pages 279-285, Piscataway, NJ., IEEE Service Center.

Kosinski A.S. (1999), *A Procedure for the Detection of Multivariate Outliers*, Computational Statistics & Data Analysis, 29, 145-161.

Liu R.Y. (1990), *On a notion of data depth based on random simplices*, Annals of Statistics, 18, 405-414.

Liu R.Y., Parelius J., Singh K. (1999*), Multivariate analysis by data depth: descriptive statistics, graphics and inference (with discussions)*, Annals of Statistics, 27.

Mesa D.M., Tsai P., Chambers R.L. (2000*), Using Tree-Based Models for Missing Data Imputation: an Evaluation Using UK Census Data*, Report, University of Southampton.

Nordbotten S, 1995. Editing statistical records by neural networks. *J.O.S*. 11 no 4:391-411.

Nordbotten, S. *(1996):"Neural Network Imputation Applied to the Norwegian 1990 Population Census Data". Journal of Official Statistics. Vol 12, No. 4. pp 385-401.*

Nordbotten, S. (1999): *Small Area Statistics from Survey and Imputed Data. Statistical Journal of the United Nations ECE. Vol 16. pp. 297-309.*

Platek, R. and Särndal, C-E. (2001*): Can a statistician deliver? Journal of Official Statistics*. Vol 17, No. 1, pp. 1-20.

Rousseeuw P.J., Leroy M.L. (1987), *Robust regression & outlier detection*, John Wiley and Sons, New York.

Rousseeuw P.J., Molenberghs G. (1993), *Transformation of Non Positive Semidefinite Correlation Matrices*, Communications in Statistics-Theory and Methods 22, 965-984.

Rousseeuw P.J. (1999), *Regression Depth (with discussion)*, Journal of the American Statistical Association 94, 388-445.

Ruiz-Gazen A. (1996), *A Very Simple Robust Estimator of a Dispersion Matrix*, Computational Statistics and Data Analysis 21, 149-162.

Stahel W.A. (1981), *Robuste Schätzungen: Infinitesimale Optimalität und Schätzungen von Kovarianzmatrizen*, Ph.D. Thesis Nr. 6881, Swiss Federal Institute of Technology, Zurich, Switzerland.

Tibshirami R (1992). Principal curves revisited. Statistics and Computing 2: 183-190.

Vapnik VN, 1996. Structure of statistical learning theory, In: Computational Learning and Probabilistic Reasoning, Ed A. Gammerman, Wiley.