

# **New developments in edit and imputation practices – needs and research**

John Charlton

*Office for National Statistics, Methodology Group,*

*1 Drummond Gate,*

*London SW1V 2QQ, UK*

*John.Charlton@ons.gov.uk*

Ray Chambers

*University of Southampton, Social Statistics Department,*

*Highfield, Southampton, UK*

*rc@alcd.soton.ac.uk*

Svein Nordbotten

*University of Bergen, Information Science,*

*N-5020 Bergen, Norway*

*Svein.Nordbotten@ifi.uib.no*

## **1. Research needs**

Editing and imputation are undertaken to improve quality in data and statistics in virtually all National Statistics Offices. Currently 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 localization, 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.

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)

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

## **2. Opportunities offered by recent developments**

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<sup>1,2</sup>; non-parametric regression<sup>3</sup>; neural networks, including multi layer perceptron<sup>4</sup>, correlation matrix memory<sup>5,6</sup>, self-organizing maps<sup>7</sup>, and support vector machines<sup>8</sup>. These new approaches will be discussed briefly in the presentation. The above methods are all currently being evaluated in a project called EUREDIT.

## **3. Current research being undertaken in the EUREDIT project**

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:

1. To establish a standard collection of data sets for evaluation purposes
2. To develop a methodological evaluation framework and develop evaluation criteria
3. To establish a baseline by evaluating currently used methods.
4. To develop and evaluate a selected range of new techniques.
5. To evaluate different methods and establish best methods for different data types.
6. To disseminate the best methods via a software CD and publications.

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). Work on the statistical evaluation criteria is now complete (see Euredit web site), and the evaluation criteria, which include operational characteristics of the methods, are being finalised. The methods being evaluated include traditional ones such as those incorporated in NIM, AGGIES, SOLAS, DEIS and logistic regression, and newer methods: multivariate robust methods; multi-layer perceptron (MLP); correlation matrix memories (CMM); self organising maps (SOM); support vector machines (SVM); and new methods for panel data and time series. Results of experiments using these methods should be available by March 2002. The final evaluation comparing all methods in different situations should be available by March 2003. There will also be a CD Rom available, containing prototype software and documentation. A conference is planned for May 2002 to disseminate and discuss the findings.

#### 4. Determining the best E&I methods for different situations

We briefly describe here the approach adopted in evaluating the statistical properties of each method. The operational characteristics of each method will also be evaluated.

**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). There are two performance requirements for editing: *efficient error detection* (detect as many errors as is feasible) and *influential error detection* (detect the errors that would lead to significant errors in the analysis unless detected). Error detection can be evaluated in terms of both the number of errors correctly identified and the number of incorrect detections it makes. Statistical outlier detection can be considered as a form of editing. Formulae for categorical and continuous data will be presented. It is also important to see how these measures vary across identifiable subgroups in the data, e.g. in a business survey performance of editing procedures may be different for different industry groups

**Imputation** is the process by which missing or suspicious values are replaced. Here we shall only concern ourselves with assessing the imputation of identifiable missing values (which could be missing because they were detected as incorrect by an edit process). 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 measures are defined below (not necessarily ordered by desirability):

- Predictive accuracy (the process should preserve true values – imputed values are as “close” as possible to true values)
- Ranking accuracy (the process should preserve the ranks of true values)
- Distributional accuracy (the process should preserve the distribution of true values)
- Estimation accuracy (the process should provide unbiased and efficient estimators for

parameters of the distribution of true values – given that these are unavailable)

- Imputation plausibility (the values imputed by the procedure should be plausible – e.g. pass all edit tests)

Formulae to operationalise each of these concepts will be described. Note that not all the above properties are meant to apply to every variable that is imputed – in particular the second property requires that the variable be at least ordinal, while the third and fourth properties are only distinguishable when the imputed variable is scalar.

## REFERENCES

1. Chambers RL, 1986. Outlier robust finite population estimation. *JASA* **81**:1063-1069.
2. Chambers RL, Kokic PN, 1993. Outlier robust sample survey inference, Invited paper, *Proc. 49<sup>th</sup> ISI Session, Firenze, August 1993*.
3. Breckling J, Sassin O, 1995. A non-parametric approach to time-series forecasting, In *ZEW-Wirtschaftsanalysen Band 5: Quantitative Verfahren im Finanzmarktbereich*, ed Schroeder Nomos Verlagsgesellschaft, Baden-Baden, 1995.
4. Nordbotten S, 1995. Editing statistical records by neural networks. *J.O.S.* **11** no 4:391-411.
5. Austin J, 1996. Distributed associative memories for high-speed symbolic reasoning. *Fuzzy Sets and Systems* 82: 223-233.
6. Austin J, Lees K, 1999. A novel search engine based on correlation matrix memories. *Neurocomputing* – special issue, Elsevier Science.
7. Kohonen T, 1989. *Self-Organisation and Associative Memory*, (Third Edition), Springer Series in Information Sciences, Springer-Verlag.
8. Vapnik VN, 1996. Structure of statistical learning theory, In: *Computational Learning and Probabilistic Reasoning*, Ed A. Gammerman, Wiley.

## RESUME

This paper first discusses the research needs for edit and imputation, including 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. We present a range of statistical criteria for comparing different edit and imputation methods that are appropriate for different types of analyses and data. Finally we describe a large ongoing research project to develop and evaluate new methods alongside existing ones.

Le document fait tout d'abord état des recherches nécessaires à mener en ce qui concerne la vérification et l'imputation, ceci inclut le développement de nouvelles méthodes plus rapides, plus efficaces, plus flexibles, ainsi qu'une méthodologie pour comparer les méthodes de sorte que des choix bien informés puissent être faits afin de définir les méthodes les plus appropriées pour une situation particulière. Nous présentons une gamme de critères statistiques pour comparer les différentes méthodes de vérification et d'imputation appropriées pour différents types d'analyses et de données. Finalement, nous décrivons un vaste projet en cours pour développer et évaluer de nouvelles méthodes en parallèle avec d'autres déjà existantes.