Publications of the Laboratory of Data Analysis
Data-analyysin laboratorion julkaisuja No. 1

DataClean 2002 - Abstracts

A conference for dealing with erroneous and missing data
29th - 31st May 2002, Jyväskylä, Finland

Edited by Pasi Koikkalainen

Center for Computational and Mathematical Modelling
University of Jyväskylä
2002

Publications of the Laboratory of Data Analysis
Data-analyysin laboratorion julkaisuja

Distribution:
Laboratory of Data Analysis
Center for Computational and Mathematical Modelling
University of Jyväskylä
P.O.Box 20
FIN-53851 Jyväskylä
Finland

Electronic publications: http://erin.mit.jyu.fi/datalab/publications

Publications of the Laboratory of Data Analysis
Data-analyysin laboratorion julkaisuja
Jyväskylä 2002 No. 1

DataClean 2002 - Abstracts

A conference for dealing with erroneous and missing data
29th - 31st May 2002, Jyväskylä, Finland

Pasi Koikkalainen (editor)

Center for Computational and Mathematical Modelling
University of Jyväskylä
2002

Laboratory of Data Analysis
University of Jyväskylä
Jyväskylä 2002
ISBN 951-39-1238-8
(digital version)
ISSN 1458-7254

DataClean 2002

29th - 31st May 2002, Jyväskylä, Finland

This conference is devoted to techniques for dealing with erroneous and missing data in large scale statistical data processing. Such data represent a fundamental problem for the data systems of official statistical agencies as well as private enterprises. In particular, the conference focus is on the identification and correction of errors and outliers in data and on imputation for missing data values. Although this topic is not a new one, the focus will be on recent developments in the application of computer intensive methods to these problems, particularly those based on the application of neural net and related methods, and their comparison with more established methods.

Local Organization Committee

Pasi Koikkalainen (University of Jyväskylä, Organization Chair)

Sara Ahola (University of Jyväskylä)
Ismo Horppu (University of Jyväskylä)
Salme Kärkkäinen (University of Jyväskylä)
Seppo Laaksonen (R&D of Department of Statistics Finland)
Anssi Lensu (University of Jyväskylä)
Markku Mielityinen (University of Jyväskylä)
Antti Penttinen (University of Jyväskylä)
Jouni Raitamäki (University of Jyväskylä)

Scientific Programme Committee

John Charlton (Office for National Statistics, Conference Chair)

Jim Austin (Univ. of York, UK)
Giulio Barcaroli (ISTAT, Italian Statistical Institute)
Raymond Chambers (Univ. of Southampton, UK)
John Charlton (Office for National Statistics, UK)
Alex Gammerman (Royal Holloway University, UK)
Beat Hulliger (Swiss Federal Statistical Office)
Pasi Koikkalainen (University of Jyväskylä)
Phil Kokic (Insiders, Germany)
Seppo Laaksonen (R&D of Department of Statistics Finland)
Birger Madsen (Novo Nordisk, Denmark)
Pascal Riviere (INSEE, France)
Ton de Waal (Statistics Netherlands)

INVITED PRESENTATIONS

Thursday, May 30. 9:45:
Evaluation of Editing and Imputation Methodology
John Charlton

Thursday, May 30. 19:30:
Multiple imputation: a general approach to many problems in statistics
Donald Rubin

Friday, May 30. 9:30:
Robust and Nonparametric Multivariate Methods
Hannu Oja

MULTIPLE IMPUTATION (Thursday 30th of May)

11:00: **The effect of item-nonresponse and multiple imputation of missing data on the estimation of productivity using establishment panel data**
Susanne Raessler

11:30: **Examples of multiple imputation in large-scale surveys**
N. T. Longford, De Montfort University, Leicester, UK

12:00: **Handling missing data by multiple imputations in the analysis of Women Prevalence of Cancer**
Ula A. Nur

EDITING AND IMPUTATION SYSTEMS (Thursday 30th of May)

13:30: **CANadian Census Edit and Imputation System**
Michael Bankier and Paul Poirier

14:00: **A high performance scalable imputation system**
M. Weeks, K. Lees, S. O'Keefe, and J. Austin

14:30: **Presentation of the INSPECTOR project**
Gregory Farmakis and Photis Stavropoulos

15:30: **Unified environment for data production and data analysis**
Pasi Koikkalainen, Ismo Horppu

16:00: **Algorithms for Automatic Error Localisation and Modification**
Ton de Waal

16:30: **Combining Editing and Imputation methods in Household surveys: an experimental application on Census data.**
Antonia Manzari

17:00: **How to deal with ineffective edits in probabilistic editing algorithms: the EUREDIT experience on bussiness data**
M.Di Zio, O.Luzi, U.Guarnera

IMPUTATION (Thursday 30th of May)

13:30: **Bayesian Networks for Imputation in Official Statistics: A case study**
Lucia Coppola, Marco Di Zio, Orietta Luzi, Alessandra Ponti, Mauro Scanu

14:00: **Coupling neural networks and predictive matching for flexible imputation**
Fernando TUSELL

14:30: **Imputation Methods for Estimating Pay Distributions from Household Survey Data**
Gabriele Beissel and Chris Skinner

15:30: **Classical and Neural Approaches for imputation**
Pasi Piela and Seppo Laaksonen

16:00: **The development of a donor imputation system**
Heather Wagstaff and Nargis Rahman

16:30: **Tree-based Classifiers for Conditional Missing Data Incremental Imputation**
Roberta Siciliano

OUTLIER DETECTION (Friday 31st of May)

11:00: **Using robust tree-based methods for outlier and error detection**
Ray Chambers, Xinqiang Zhao, and Adao Hentges

11:30: **Detecting Multivariate Outliers in Incomplete Survey Data with the Epidemic Algorithm**
Beat Hulliger and Cédric Béguin

12:00: **Detecting Multivariate Outliers in Incomplete Survey Data with the BACON-EM algorithm**
Cédric Béguin and Beat Hulliger

NEURAL NETWORKS (Friday 31st of May)

11:00: **Edit and Imputation using a binary neural network**
K. Lees, S. O'Keefe, and J. Austin

11:30: **Kernel Methods for the Missing Data Problem**
Hugh Mallinson, Alex Gammerman

12:00: **Neural networks for Editing and Imputation**
Pasi Koikkalainen

SELECTIVE EDITING (Friday 31st of May)

13:30: **Development of a Graphically Oriented Macro-Editing Analysis System for NASS Surveys**
Dale Atkinson

14:00: **Developing selective editing methodology for surveys with varying characteristics**
Pam Tate, Office for National Statistics

14:30: **A Technical Framework for Input Significance Editing / The Application of Output Statistical Editing**
Keith Farwell

15:30: **Selective editing by means of a plausiblity indicator**
Jeffrey Hoogland

16:00: **Demonstration of The Graphical Editing and Analysis Query System**
Paula Weir, U.S. Department of Energy, EIA

16:30: **Stopping criterion: a way of optimising data editing and assessing its minimal cost**
Pascal Rivière, INSEE

INVITED PRESENTATIONS

MULTIPLE IMPUTATION: A GENERAL APPROACH TO MANY PROBLEMS IN STATISTICS

Donald Rubin

Harward University
Faculty of Arts and Sciences 1 Oxford Street, MA 02138
USA

Multiple imputation (MI, Rubin, 1987) was originally proposed as a method to handle missing data due to nonrepsonse in surveys, especially surveys destined to support the production of large public-use data sets. There have now been many very successful applications of MI to such data sets in the US (e.g., SCF,NHANES III, FARS, NMES), as well as to other data sets with missing data (e.g., randomized pharmaceutical trials presented to the US FDA). Recent applications also include the handling of "matrix sampling" in educational settings (e.g., NAEP) and in marketing contexts for business surveys. Even more novel applications involve the use of MI to address noncompliance in human randomized trials of anthrax vaccines and to try to build a bridge between these studies and randomized trials of macaques. In the macaque studies, true survival outcomes are measured, as well as biomarkers (e.g., blood antibody levels), whereas in the human trials, only the biomarkers are available. This presentation will be a free flowing exposition of some of these applications, and the crucial role, both conceptually and computationally, that MI makes to valid statistical analysis.

EUREDIT - EVALUATION OF EDITING AND IMPUTATION METHODOLOGY

John Charlton

Office for National Statistics
1 Drummond Gate
London SW1V2QQ
UK

Imputation-based methods for dealing with incomplete or inconsistent data are used in virtually all National Statistics Institutes (NSIs), and in academic and business research. Currently, these methods are typically based on simple statistical ideas (e.g. nearest neighbours). Also, little is known about the comparative performance of each method, across the wide variety of data sources being used.

Recent, advances in computing capabilities have made possible the application of the more complex statistical modeling techniques. The EUREDIT project 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.

In EUREDIT the fundamental approach adopted involves identifying sound scientific and technical, user-oriented criteria to enable a meaningful comparison of current and new promising methods for data editing and imputation.

ROBUST AND NONPARAMETRIC MULTIVARIATE METHODS

Hannu Oja [1]

Department of Mathematic and Statistics
University of Jyväskylä
P.O.Box 35 (MaD)
FIN-40351 Jyväskylä
Finland

1 Introduction

Classical multivariate methods (principal component analysis, multivariate regression, canonical correlation, discriminant analysis, Mahalanobis distance, Mahalanobis angle, etc.) are based on the sample mean vector and sample covariance matrix. Mean vector and covariance matrix are optimal if the data come from a multivariate normal distribution but they are very sensitive to outlying observations and loose in efficiency in the case of heavy tailed distrutions. In this talk, robust and nonparametric competitors of the mean vector and covariance matrix and their use in multivariate inference are considered.

2 Location vector, scatter matrix, shape matrix

2.1 Definitions

We assume that $X = \{x_1, \ldots, x_n\}$ is a random sample from a k-variate elliptically symmetric distribution with cumulative distribution function (cdf) F, symmetry center μ and covariance matrix Σ (if they exist). The aim is to consider and compare the location vector, scatter matrix and shape matrix functionals. The location, scatter and shape functionals are then denoted by $T(F)$, $C(F)$ and $V(F)$, or alternatively by $T(x)$, $C(x)$ and $V(x)$ if x is a random vector with cdf F. To be specific, a k-vector valued functional $T = T(F)$ is a **location vector** if it is affine equivariant, that is, if $T(Az+b) = AT(z)+b$ for any $k \times k$ nonsingular matrix A and k-vector b. A matrix valued functional $C = C(F)$ is a **scatter matrix** if it is $PDS(k)$ (a positive definite symmetric $k \times k$ matrix) and affine equivariant, which in this case means that $C(Az + b) = AC(z)A^T$. Finally, functional $V = V(F)$ is a **shape matrix** if it is $PDS(k)$, $Tr(V) = k$ and it is affine equivariant in the sense that $V(Az + b) = [k/Tr(AV(z)A^T)]AV(z)A^T$. Note that if $C(F)$ is a scatter

[1]Research supported by a grant from the Academy of Finland.

matrix then the related shape matrix is given by $V = [k/Tr(C)] \cdot C$. The affine equivariance property implies that, if the distribution of z is a spherically symmetric distribution with cdf F_0, mean vector 0 and covariance matrix $\Sigma = I_k$, then, for all location, scatter and shape functionals T, C and V,

$$T(Az+b) = b, \quad C(Az+b) = c_0 AA^T \quad \text{and} \quad V(Az+b) = \frac{1}{Tr(AA^T)} AA^T$$

where constant c_0 depends of both functional C and distribution F_0. Note that, for elliptic models, location vectors and shape matrices are directly comparable without any modifications.

2.2 Influence functions and efficiency

The influence function is a tool to describe the robustness properties of an estimator; it also often serves a way to consider the asymptotic properties. The influence function (IF) of a functional T at F measures the effect of an infinitesimal contamination located at a single point x as follows. We consider the contaminated distribution

$$F_\varepsilon = (1-\varepsilon)F + \varepsilon\delta_x$$

where δ_x is the cumulative distribution function of a distribution with probability mass one at x. The influence function is defined as

$$IF(x, T, F) = \lim_{\varepsilon \downarrow 0} \frac{T(F_\varepsilon) - T(F)}{\varepsilon}.$$

The influence functions of location scatter and shape functionals $T(F)$, $C(F)$ and $V(F)$ at spherical F_0 are then given by

$$IF(x; T, F_0) = \gamma_T(r)u,$$

$$IF(x; C, F_0) = \alpha_C(r)uu^T - \beta_C(r)I_k,$$

and

$$IF(x; V, F_0) = \alpha_V(r)\left[uu^T - \frac{1}{k}I_k\right],$$

for a contamination point x, $r = ||x||$ and $u = ||x||^{-1}x$. See Croux and Haesbroeck (1999). If $V = (k/Tr(C))C$ then $\alpha_V(r) = k\alpha_C(r)/Tr(C(F_0))$. Note that the regular estimates (mean vector, covariance matrix) use weight functions $\gamma_T(r) = r$, $\alpha_C(r) = r^2$ and $\beta_C(r) = E(r^2)/k$. For robust functionals, the influence functions are continuous and bounded.

The constants

$$E(\gamma_T^2(r), \quad E(\alpha_C^2(r)), \quad E(\beta_C^2(r)) \quad \text{and} \quad E(\alpha_V^2(r))$$

are then used in efficiency comparisons. It is easy to see for example that, under general assumptions in the spherical case F_0,

$$\sqrt{n}T_n \to_d N_k\left(0, \frac{E(\gamma_T^2(r))}{k} I_k\right),$$

and

$$\sqrt{n}(V_n - I_k) \to_d N_{k\times k}\left(0, E(\alpha_V^2(r))U_k\right)$$

where

$$U_k = E[(uu^T) \otimes (uu^T)] - \frac{1}{k} I_{k^2}.$$

3 Robust and nonparametric alternatives

In this talk we consider and compare multivariate location, scatter and shape estimates of three kinds, namely M-estimates, S-estimates and R-estimates.

Let again $X = \{x_1, \ldots, x_n\}$ is a random sample from a k-variate elliptically symmetric distribution with cumulative distribution function (cdf) F. The location and scatter estimates are then constructed as follows. Let C_n be a $PDF(k)$ matrix and T_n a k-vector. Consider transformed observations $z_i = C_n^{-1/2}(x_i - T_n)$, $i = 1, ..., n$, with symmetric $C_n^{-1/2}$. Write $r_i = ||z_i||$ and $u_i = ||z_i||z_i$, $i = 1, ..., n$. The multivariate **location and scatter M-estimates** are the choices T_n and C_n for which

$$\text{ave}_i\{w_1(r_i)u_i\} \;=\; 0 \;\text{ and }\; \text{ave}_i\left\{w_2(r_i)u_iu_i^T\right\} \;=\; \text{ave}_i\{w_3(r_i)\} \cdot I_k.$$

for some weight functions $w_1(r)$, $w_2(r)$ and $w_3(r)$. See Maronna (1976) and Huber(1981). Next we define S-estimates. The multivariate **location and scatter S-estimates** are the choices T_n and C_n which minimize $det(C_n)$ subject to $\text{ave}_i\{\rho(r_i)\} \le 1$ for some function $\rho(r)$. See Rousseeuw and Leroy(1987) and Davies(1987). For the relation between M- and S-estimates, see Lopuhaä (1989). Finally, Ollila, Hettmansperger and Oja (2002) introduced estimates based on multivariate sign vectors. In their approach **location and scatter estimates based on signs** are the choices T_n and C_n for which

$$\text{ave}_i\{S(z_i)\} \;=\; 0 \;\text{ and }\; \text{ave}_i\left\{S(z_i)S^T(z_i)\right\} \;=\; \frac{\text{ave}\left\{S^T(z_i)S(z_i)\right\}}{k} \cdot I_k$$

where $S(z)$ is a multivariate sign function. Multivariate rank vectors may be used similarly as well and the resulting family of estimates can be called multivariate **location and scatter R-estimates**.

4 Applications

4.1 Subspace estimation

In our first example we consider the problem of subspace estimation. Let $X = \{x_1, \ldots, x_n\}$ be a random sample from a k-variate elliptically symmetric distribution with covariance matrix $\Sigma = P\ \Lambda\ P^T$ where P is an orthogonal matrix with the eigenvectors of Σ in its columns and Λ the diagonal matrix with the corresponding distinct eigenvalues $\lambda_1 > .... > \lambda_k > 0$ as diagonal entries. Write $P = (P_1 P_2)$ where the r columns on P_1 and s columns of P_2, $r + s = k$, are supposed to span the signal and noise subspaces, respectively.

Any shape matrix estimate V_n may now be used to estimate the signal space. If $P_n = (P_{n1}, P_{n2})$ is the estimate of $P = (P_1, P_2)$ obtained from V_n then

$$D^2(P_{1n}, P_1) = ||P_{2n}^T P_1||_F^2,$$

where $||A||_F = Tr(A^T A)$ is the so called Frobenius matrix norm, measures the distance between the estimated and true signal subspace. See Crone and Crosby (1995).

If we then compare the accuracy of the estimates based on V_n and V_n^*, a natural measure is

$$\frac{E_F[D^2(P_{1n}, P_1)]}{E_F[D^2(P_{1n}^*, P_1)]} \to \frac{E_{F_0}[\alpha_V^2(r)]}{E_{F_0}[\alpha_{V^*}^2(r)]}$$

as $n \to \infty$.

4.2 Mahalanobis distance, Mahalanobis angle

Let again $X = \{x_1, \ldots, x_n\}$ be a random sample from a k-variate elliptically symmetric distribution and let T_n and C_n be location and scatter estimates. Mahalanobis distance is sometimes used to measure a distance of an observation from the center of the data

$$D_i = (x_i - T_n)^T C_n^{-1} (x_i - T_n)$$

The so called Mahalanobis angles

$$D_{ij} = (x_i - T_n)^T C_n^{-1} (x_j - T_n)$$

measure angular distances between vectors $x_i - T_n$ and $x_j - T_n$. Finally, the Mahalanobis distance between two observations x_i and x_j is given by

$$(x_i - x_j)^T C_n^{-1} (x_i - x_j).$$

All the measures are naturally affine invariant. Again, mean vector and regular covariance matrix give measures which are sensitive to outlying observations; we end this talk with a discussion on the robustified versions of these measures.

References

[1] Crone, L.J., Crosby, D.S. (1995): Statistical applications of a metric on subspaces to satellite meteorology. *Technometrics*, **37**, 324-328.

[2] Croux, C., Haesbroeck, G. (1999): Principal component analysis based on robust estimators of the covariance and correlation matrix: Influence function and efficiencies. *Biometrika*, **87**, 603-618.

[3] Davies, P.L. (1987): Asymptotic behavior of S-estimates of multivariate location parameters and dispersion matrices. *Ann. Statist.*, **15**, 1269-1292.

[4] Hampel, F.R., Ronchetti, E.M., Rousseeuw, P.J., Stahel, W.A. (1986): *Robust Statistics: The Approach Based on Influence Functions*. Wiley, New York.

[5] Huber, P.J. (1981): *Robust Statistics*. Wiley, New York.

[6] Lopuhaä, H.P. (1989): On the realtion between S-estimators and M-estimators of multivariate location and scatter. *Ann. Statist.*, **17**, 1662-1683.

[7] Maronna, R.A. (1976): Robust M-estimators of multivariate location and scatter. *Ann. Statist.*, **4**, 51-76.

[8] Ollila, E., Hettmansperger, T.P., Oja, H. (2002): Affine equivariant multivariate sign methods, Submitted.

MULTIPLE IMPUTATION (Thursday 30th of May)

THE EFFECT OF ITEM-NONRESPONSE AND MULTIPLE IMPUTATION OF MISSING DATA ON THE ESTIMATION OF PRODUCTIVITY USING ESTABLISHMENT PANEL DATA

Susanne Raessler

Institute of Statistics and Econometrics
Faculty of Business Administration, Economics and Social Sciences
Friedrich-Alexander-University Erlangen-Nuremberg
Lange Gasse 20
D-90403 Nuremberg

This paper illustrates the effects of missing data in panel surveys on the results of multivariate statistical analysis. Large data sets of the German IAB Establishment Panel are used which typically contain more than 10000 cases in each wave. Due to item-nonresponse continuous as well as categorical variables of interest are affected by missing values. Using only the available cases for the estimation task reduces the data set considerably. Thus, we multiply impute the missing data applying a Bayesian data augmentation algorithm. The imputer's model is based on a multivariate normal distribution for the data and some noninformative prior distributions for the parameters. The handling of so-called semi-continuous variables is explained to impute the incomplete mixed data suitably.

The analyst's model is based on a translog production function with labour and capital as input factors. Also the influence of industries and the use of modern technologies are considered. Furthermore, we include interaction variables that indicate deviations in the parameters concerning the two parts of Germany. Besides differences between industries, we find a higher productivity, if modern technologies are installed. The results for Eastern and Western Germany only differ for some industries and the constant. Using only the available cases valuable information seems to be discarded. Calculating the regression coefficients using the imputed data sets and combining the results according to the multiple imputation principle reduce these differences up to 11%-points. Additionally, the differences of the industrial branches between Eastern and Western Germany decrease when inference is based on multiple imputation.

EXAMPLES OF MULTIPLE IMPUTATION IN LARGE-SCALE SURVEYS

N. T. Longford

De Montfort University, Leicester
James Went Building 2-8, De Montfort University, The Gateway,
Leicester LE1 9BH, UK.
Email: NTL@dmu.ac.uk

Missing data are a ubiquitous problem in large-scale surveys. Such incompleteness is usually dealt with either by restricting the analysis to the cases with complete records or by imputing for each missing item an efficiently estimated value. The deficiencies of these approaches will be discussed, especially in the context of estimating a large number of quantities. The main part of the paper will describe two examples of analysses using multiple imputation.

In the first, the ILO employment status is imputed in the British Labour Force Survey by a Bayesian bootstrap method. It is an adaptation of the hot deck method which seeks to fully exploit the auxiliary information. Important auxiliary information is given by the previous ILO status, when available, and the standard demographic variables.

Missing data can be interpreted more generally, as in the framework of the EM algorithm. The second example is from the Scottish House Condition Survey, and its focus is on the inconsistency of the surveyors. The surveyors assess the sampled dwelling units on a large number of elements, or features of the dwelling, such as internal walls, roof, and plumbing, which are scored and converted to a summarising 'comprehensive repair cost'. The level of inconsistency is estimated from the discrepancies between the pairs of assessments of doubly surveyed dwellings. The principal research questions are: how much information is lost due to the inconsistency and whether the naive estimators that ignore the inconsistency are unbiased. The problem is solved by multiple imputation, generating plausible scores for all the dwellings in the survey.

HANDLING MISSING DATA BY MULTIPLE IMPUTATIONS IN THE ANALYSIS OF WOMEN PREVALENCE OF CANCER

Ula A. Nur

University of Leeds
Nuffield Institute for Health (Block D)
71-75 Clarendon road
Leeds LS2 9PL
email: hssuamn@leeds.ac.uk

The UK Women's Cohort study aims to explore the relationship between diet and cancer incidence and mortality in a group of middle-aged vegetarian women in the UK. Standard statistical methods employed in epidemiological studies are valid only when applied to a representative sample of the population of interest. Even when the sample has been designed to be representative at the outset, the validity of the methods will be eroded if some of the subjects are subsequently lost to the survey, or failed to compete all items of the questionnaire. Missing data can rarely be avoided in large-scale studies in which subjects are requested to complete questionnaires with many items. Methods for handling missing data have been developed for a variety of contexts, With Multiple Imputations the uncertainty about the missing values is represented by the differences among the sets of imputed (plausible) values. Once the plausible values are generated, the remainder of the analysis is as complex as the planned (complete-data) analysis, except that it is conducted on the datasets completed by each set of the plausible values. The aim of this research is to compare multiple imputations to other methods of handling missing data in the statistical analysis-investigating link between prevalence of cancer and a number of life style and socio-economic factors.

EDITING AND IMPUTATION SYSTEMS (Thursday 30th of May)

CANADIAN CENSUS EDIT AND IMPUTATION SYSTEM

Michael Bankier and Paul Poirier

Census Research and Development Section,
Social Survey Methods Division,
Statistics Canada,
15th Floor, Coats Building,
Ottawa, Ontario K1A 0T6
Canada
E-Mail: bankier@statcan.ca

CANCEIS (CANadian Census Edit and Imputation System) is a generalized Edit and Imputation (E&I) system that was used early in 2002 to process the demographic variables from the 2001 Canadian Census on PCs. Later in 2002, it will perform E&I on the labour, mobility, place of work and mode of transport variables as well. It performs minimum change nearest neighbour imputation on a mixture of qualitative and quantitative variables. It is written in the C programming language and is highly efficient computationally. CANCEIS (or an earlier version of the software) will be used to process some of the variables in the 2001 Ukraine Census, the 2000 Brazilian Census and the 2001 Swiss Census. In addition, the 2001 Italian Census, having studied CANCEIS, will use a similar approach in their imputation methodology.

A HIGH PERFORMANCE SCALABLE IMPUTATION SYSTEM

M. Weeks, K. Lees, S. O'Keefe, and J. Austin

Advanced Computer Architectures Group
Department of Computer Science
University of York
Heslington
YORK
YO10 5DD
UK

This paper describes the implementation of a highly scalable method for imputation. Specialised hardware in a distributed environment provides high performance and scalability. Imputation is the process by which missing fields in a data set can be generated from known acceptable data. As part of the Euredit project for the development and evaluation of new methods for editing and imputation, we use the k-nearest-neighbour (kNN) approach to determine the imputed data. However, kNN processing can be slow as the vector distance must be calculated between the query point and all points in the data set. For very large data sets performance can be restrictively slow. To solve this problem we apply AURA (Advanced Uncertain Reasoning Architecture), which is a generic family of neural network based techniques and implementations intended for high speed approximate search and match operations on large data sets. AURA is based upon a binary neural network called a Correlation Matrix Memory (CMM). Hardware PRESENCE (PaRallEl StructurEd Neural Computing Engine) cards have been developed to accelerate core CMM functionality. Mapping a CMM to multiple PRESENCE cards provides data and performance scalability. The Cortex-1 distributed neural processor is a high performance environment for AURA development. It is a seven node PC cluster containing 28 PRESENCE cards, providing 3.5 gigabytes of CMM storage. Allowing AURA to perform a high speed sift of a large data set, we can create a smaller subset of the data. Traditional kNN methods can then be applied to this subset, in order to determine the imputed values. This paper describes how the AURA imputation technique maps onto Cortex-1, and determines how its performance scales with increasing data set size.

PRESENTATION OF THE INSPECTOR PROJECT

Gregory Farmakis and Photis Stavropoulos

Liaison Systems SA
77 Akadimias Str,
106 78 Athens
Greece
email: photis@liaison.gr

Quality in the Statistical Information Life-Cycle (INSPECTOR) is a research and development project partially funded by the IST Program , CPA4/2000: Statistical tools, methods & applications for the Information Society.

The main objective of the project is the design and development of a generic, distributed and flexible data validation system, which will be able to be seamlessly integrated in the current (or future) processes of statistical data collection, in order to ensure and monitor data quality. The system will be accessible in a distributed way in order to validate large statistical data sets before their transmission or even throughout their production, ensuring homogeneity and consistency which are critical quality parameters of the validation process and the data quality in general. Among the critical project objectives are, the development of a formal framework for the classification and semantic notation of validation rules, the design, development and population of the rules repository, the development of the validation engine and a, Java application implemented, Validation Client.

The main concept underlying our approach is the treatment of validation rules as metadata, as opposed to the usual approach of implementing case specific validation programs. In this talk we will present this novel representation of validation rules, its importance for the design of the validation software, and the project's progress that far.

UNIFIED ENVIRONMENT FOR DATA PRODUCTION AND DATA ANALYSIS

Ismo Horppu and Pasi Koikkalainen

Laboratory of Data Analysis University of Jyväskylä
P.O.Box 20, FIN-35851 Jyväskylä
Finland

Currently there is only limited software support for statistical editing and imputation. Most software systems are experimental and not designed for generic data production.

In this presentation we demonstrate a software that has been build on the top of our NDA (Neural Data Analysis) software platform. New methodology for data editing and imputation has been implemented in the software kernel, and a new user interface has been build to support the tasks of data editing and imputation.

The software is an attempt to implement a typical *data production process* (DPP) as done in official statistics and industrial data management. We consider that this is defined by the following requirements.

a) Software should support data manipulation, reorganization and visualization. These are common tasks in any type of data analysis.

b) Use of external knowledge, such as edit rules, must be supported. We have done this with a simple rule converter that translates edit rules to NDA type of expressions.

c) Variable selections and case spesific edit/imputation operations should allowed. The user should be able do them with minimal effort.

d) Several methodologies for editing and imputation must be supported.

e) Experimenting and playing with data should be easy.

f) There should several tools to evaluate the results of editing and imputation.

ALGORITHMS FOR AUTOMATIC ERROR LOCALISATION AND MODIFICATION

Ton de Waal

Statistics Netherlands,
PO Box 4000, 2270 JM Voorburg, Netherlands,
e-mail address: twal@cbs.nl

Automatic edit and imputation can be subdivided into three steps. The first step is error localisation during which the erroneous fields are identified. The second step is imputation. In this step the erroneous fields and missing data are imputed for. The final step is modification during which any edits that may still be violated after imputation are made satisfied by slightly modifying the imputed values. Statistics Netherlands has developed algorithms for error localisation and modification in a mix of continuous and categorical data. The developed algorithm for error localisation is based on the (generalised) Fellegi-Holt paradigm, which says that data should be made to satisfy all edits by changing the fewest (weighted) number of variables. The algorithm determines all optimal solutions to the error localisation problem, given a user-specified upper bound on the maximum number of errors. In case there are more errors in a particular record than the specified upper bound, this record is not corrected automatically. In this paper we describe the algorithm in some detail. The developed algorithm for modification is based on the paradigm that, after imputation, the imputed values should modified as little as possible to satisfy the edits. To measure the distance between an imputed record and the final, modified record, a distance function consisting of a sum of a part involving only the categorical variables and a part involving only the continuous ones is used. The categorical part of the distance function consists of a sum of positive weights, where each weight indicates the costs of changing the imputed value into a certain other value. The numerical part of the distance function consists of a weighted sum of absolute differences between the imputed values and the final values. In this paper we describe a heuristic that has been developed to minimise this distance function subject to the restriction that all edits become satisfied. Finally, in the paper we also describe modifications of the algorithm for error localisation problem so it can be applied to solve related problems.

COMBINING EDITING AND IMPUTATION METHODS IN HOUSEHOLD SURVEYS: AN EXPERIMENTAL APPLICATION ON CENSUS DATA.

Antonia Manzari

ISTAT
c/o Servizio MPS
Via Cesare Balbo,16 - Roma
Italy
manzari@istat.it

Data from Household surveys are generally characterised by a hierarchical structure: data are collected at the household level with information for each person within the household. Some collected variables are related to the household features, while the remaining ones concern the person. Some person variables are of demographic type, other ones are non-demographic. The majority of variables are of qualitative or categorical type (though integer coded data), but some variables can be of quantitative or numeric type. These features makes the Editing and Imputation (E&I) phase a complex matter: the relationships among the values of demographic variables referred to different persons within the household oblige the user of the E&I system to take into account the between persons constraints together with the constraints among the values of variables referred to a given person (within person constraints). Moreover, joint E&I of both qualitative and quantitative variables is required, but while constraints involving qualitative data are definable by logical edit rules, the relationships among numeric variables are generally expressed by arithmetic edit rules (generally linear inequalities). Therefore E&I system treating invalid or inconsistent responses for qualitative and numeric variables simultaneously are needed.

How E&I phase can be performed in so complex a situation by using automatic generalised system for micro-editing? Complex E&I task can be tackled dividing the E&I phase into simpler sub-problems and finding the most appropriate solution for each of them. The E&I strategy as a combination of several methodologies can be an useful way to clean data as the data quality is maintained because every peculiar problem is faced by a suitable tool.

In this paper the E&I strategy used in cleaning a perturbed Sample of Anonymised Records for individuals from UK Census 1991 (SARs data) is presented. The strategy has been developed in EUREDIT project by using currently in-use (or "standard") methods for data E&I, in order to obtain a performance benchmark for advanced techniques.

The E&I phase has been divided into two macro sub-phases where different automatic systems (CANCEIS and SCIA) were used. Inside the second macro sub-phase (SCIA) several applications were performed varying the variables to handle and/or the specified edit rules.

IMPUTATION
(Thursday 30th of May)

BAYESIAN NETWORKS FOR IMPUTATION IN OFFICIAL STATISTICS: A CASE STUDY

Lucia Coppola, Marco Di Zio, Orietta Luzi, Alessandra Ponti, Mauro Scanu

ISTAT,
via Cesare Balbo 14, 00184 Roma
Italy
e-mail: luzi@istat.it

Bayesian Networks are particularly useful for dealing with high dimensional statistical problems (Jensen, 1996). They allow reducing the complexity of the inspected phenomenon by representing joint relationships among a set of variables, through conditional relationships among subsets of these variables. Roughly, it is equivalent to split the overall problem in many sub-problems, but assuring that the combination of all the single sub-solutions (corresponding to the single sub-problems) will give the optimal global solution. Official Statistics is a natural application field for this technique because of the complexity of statistical surveys carried on in this context. In particular, following Thibaudeau and Winkler (2001), we used the Bayesian Networks for imputing missing values. We performed a first experiment on a subset of anonymous individual records and variables surveyed in 1991 U.K Population Census (SARS).

References

Jensen F. V. (1996) An introduction to Bayesian Networks, Springer Verlag, New York.
Thibaudeau Y., Winkler W.E. (2001) Bayesian networks representations, generalized imputation, and synthetic micro-data satisfying analytic constraints, unpublished manuscript.

COUPLING NEURAL NETWORKS AND PREDICTIVE MATCHING FOR FLEXIBLE IMPUTATION

Fernando TUSELL

Facultad de CC.EE. y Empresariales
Avenida del Lehendakari Aguirre, 83
E-48015 BILBAO
email: etptupaf@bs.ehu.es

We address the problem of imputation of vectors, such as is required, for instance, when a large survey is supplemented with another one in which only a subset of all questions is asked, and imputation on the non-asked ones is needed.

In Bárcena and Tusell(2000) we proposed a tree-based method that afforded easy, nonparametric imputation of multivariate responses, such as is required e.g. when linking two surveys. While its performance is quite competitive with existing methods (and best in some circumstances), the method suffers from the discreteness intrinsic to the approximation that trees can make of a continuous function.

The basic ideas present in our previous work –predictive matching and flexible, nonparametric approximation– are carried now one step further. We show how neural networks can be put to use so as to provide an approximation of a predictive distribution. Thus, a simple, nonparametric, distribution-free analogue of multiple imputation is obtained.

We will show the performance of the method in both real and simulated date.

references

Bárcena, M.J. and Tusell, F. (2000) "Tree based algorithms for missing data imputation", COMPSTAT'2000: Proceedings in Computational Statistics, Physica-Verlag, Heidelberg, 2000.

IMPUTATION METHODS FOR ESTIMATING PAY DISTRIBUTIONS FROM HOUSEHOLD SURVEY DATA

Gabriele Beissel[1] *and Chris Skinner*[2]

Department of Social Statistics,
University of Southampton,
Southampton SO17 1BJ, UK,
[1] *gbeissel@socsci.soton.ac.uk,* [2] *cjs@socsci.soton.ac.uk*

Distributions of hourly pay are important for a wide range of social and economic policy issues. However, it is difficult to obtain reliable data on both earnings and hours due to measurement error.

We use data from the U.K. Labour Force Survey, a large survey of households, which includes information on hours worked and earnings of employees. However, these variables appear to be subject to a considerable amount of measurement error, which is thought to lead to substantial upward bias of estimates of the lower end of the pay distribution. An alternative variable on hourly earnings is obtained directly and appears to give very accurate information but is subject to a high amount of missing data, because many individuals are not able to report their hourly pay. The aim is to impute the missing values taking into account information on the erroneous variable and other covariates, such that the imputation method effectively corrects for the measurement error in the pay variable. Under the assumption of ignorable nonresponse, an imputation method using a random hot deck procedure within imputation classes based on a regression model, is carried out and compared to more established methods such as predictive mean matching, as investigated in Skinner and Beissel (2001). The imputation is applied multiple times. A formula for variance estimation under this imputation method taking into account imputation, response and sampling variability and the complex weighting scheme of the survey, using a design-based approach (Rao and Sitter, 1995), is derived. A computer intensive simulation study is carried out showing good results for point and variance estimators.

In addition, we consider variance estimation using Rubin's multiple imputation formula. This formula is designed for proper multiple imputation, however, and we find that this approach underestimates the variance for the improper imputation procedure.

References

[1] Rao, J.N.K. and Sitter, R.R. (1995): Variance Estimation under Two-Phase Sampling with Applications to Imputation for Missing Data, Biometrika, 82, 2, pp. 453-460.
[2] Skinner, C.J. and Beissel, G. (2001): Estimating the Distribution of Hourly Pay from Survey Data, paper presented at the CHINTEX Workshop, The Future of Social Surveys in Europe, Helsinki 29, 30 Nov. 2001.

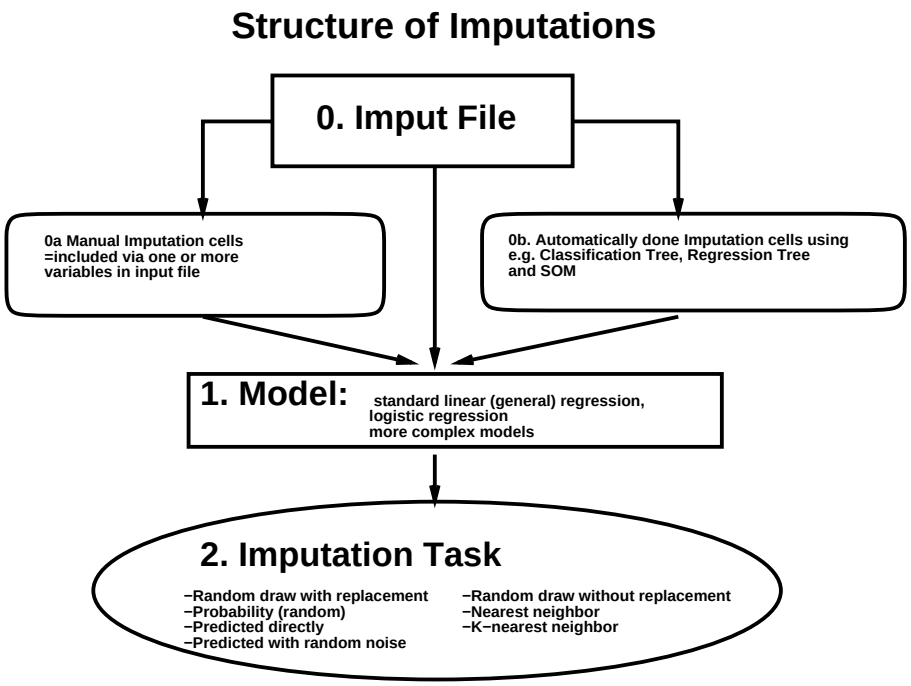
NEW AND TRADITIONAL TECHNIQUES FOR IMPUTATION

Seppo Laaksonen and Pasi Piela

Statistics Finland
FIN-00022 Tilastokeskus
Finland
Email: firstname.secondname@stat.fi

The purpose of the paper is to give an overview of new and traditional methods for imputation, and specify some of these methods with empirical exercises. It is nice that we may approach both to these so-called new techniques and traditional techniques in the same way as described in the attached figure. The difference is derived mainly from a model used prior to imputation. As far as newer techniques are concerned, the imputation model is more exploratory and non-parametric whereas in case of traditional techniques it is parametric and linear. Another difference is concerned the level of automation, traditional techniques being less automated.

Self-organizing maps (SOM) is an iterative method for classification and can thus also be used in finding the imputation classes. Imputations are made within clusters, located by corresponding neurons, in several ways that can be based on both traditional and neural methods as MLP models. Naturally, there are several modifications developed for the SOM, Tree-structured SOM as one of them.



THE DEVELOPMENT OF A DONOR IMPUTATION SYSTEM

Heather Wagstaff and Nargis Rahman

Office for National Statistics, Room D212, Drummond Gate,
London, SW1V 2QQ.
Email1: heather.wagstaff@ons.gov.uk
Email2: nargis.rahman@ons.gov.uk

As part of the EUREDIT Project, the Office for National Statistics (ONS) has developed a prototype donor imputation system (DIS) for the imputation of item non-response. The DIS implements the joint imputation method proposed by Fellegi & Holt 1976. The basic principle underlying this method is that all missing items within a record are imputed using a single clean record as a donor. Donor imputation methods select values from a wholly valid record, the 'donor', and copy the values to fill the missing items of another record, the 'recipient'. The ONS DIS supports a search algorithm which hunts for candidate donor records from the whole data matrix using a set of primary matching variables. Once a pool of candidate donor records has been found the nearest neighbour, based on statistical closeness, is selected to provide all missing items to the recipient. Thus the current DIS searches for a single donor for all missing items in a record but the current functionality also allows an option to select a different donor for each missing item should the user so choose. There is multiple choice of distance functions for both categorical and continuous matching variables. Current evaluation has produced evidence that the DIS achieves good results when a suitable set of matching variables are selected. However, there is also evidence that a comprehensive statistical analysis, together with sound knowledge, of the data set are necessary to obtain a good set of predictors.

TREE BASED CLASSIFIERS FOR CONDITIONAL INCREMENTAL MISSING DATA IMPUTATION

Claudio Conversano, Roberta Siciliano

Department of Matematics and Statistics
University of Naples Federico II
Via Cintia, Monte SantŠAngelo, I-80126
Napoli, Italy
conversa@unina.it, roberta@unina.it

1. Introduction

Missing or incomplete data are a serious problem in many fields of research because they can lead to bias and ineściency in estimating the quantities of interest. The relevance of this problem is strictly proportional to the dimensionality of the collected data. Particularly, in data mining applications a substantial proportion of the data may be missing and predictions might be made for instances with missing inputs.

In recent years, several approaches for missing data imputation have been presented in the literature. Main reference in the field is the Little and Rubin book on statistical analysis with missing data [L&R-87]. In most situations, a common way for dealing with missing data is to discard records with missing values and restrict the attention to the completely observed records. This approach relies on the restrictive assumption that missing data are Missing Completely At Random (MCAR), i.e., that the missingness mechanism does not depend on the value of observed or missing attributes. This assumption rarely holds and, when the emphasis is on prediction rather than on estimation, discarding the records with missing data is not an option [S-98]. An alternative and weaker version of the MCAR assumption is the Missing at Random (MAR) condition. Under a MAR process the fact that data are missing depends on observed data but not on missing data themselves. While the MCAR condition means that the distributions of observed and missing data are indistinguishable, the MAR condition states that the distributions diőer but missing data points can be explained (and therefore predicted) by other observed variables. In principle, a way to meet the MAR condition is to include completely observed covariates that are highly related to the missing ones. Actually, most of the existing methods for missing data imputation discussed in Shafer [Sc-97] just assume the MAR condition and, in these settings, a common way to deal with missing data are conditional mean methods or model imputation, i.e., to fit a model on the complete cases for a given variable treating it as the outcome and then, for cases where this variable is missing, plug-in the available data in the model to get an imputed value. The most popular conditional mean method employs least squares regression but it can be often unsatisfactory for nonlinear data and biased if model misspecification occurs.

In this work, in order to overcome the shortcomings of the conditional mean imputation method, d'The work was supported by EC Research Project IST-2000-26347 (INSPECTOR) Funds. we propose the iterative use of tree based models for missing data imputa-

tion in large data bases. The proposed procedure uses lexicographic order to rank missing values that occur in diőerent variables and deals with these incrementally, i.e, augmenting the data by the previously filled in records according to the defined order.

2. Overview of Tree Based Classifiers

Tree Based Classifiers consists of a recursive binary partition of a set of objects described by some explanatory variables (either numerical or and categorical) and a response variable [B-84]. Data are partitioned by choosing at each step a variable and a cut point along it, generating the most homogeneous subgroups respect to the response variable according to a goodness of split measure. The procedure results in a powerful graphical representation known as decision tree, which express the sequential grouping process. Once the tree is built, a response value or a class label is assigned to each terminal node. According to the nature of the response variable, they usually distinguish between Classification Tree (for the categorical response case) and Regression Tree (for the numerical response case). In the classification tree case, when the response variable takes value in a set of previously defined classes, the node is assigned to the class which presents the highest proportion of observations. Whereas, in the regression tree case, the value assigned to cases in a given terminal node is the mean of the response variable values associated with the observations belonging to the given node. Note that, in both cases, this assignment is probabilistic, in the sense that a measure of error is associated with it. Main advantages of these methods are:

a) their non parametric nature, since they do not assume any underlying family of probability distributions;
b) their flexibility, because they handle simultaneously categorical and numerical predictors and interactions among them;
c) their simplicity, from a conceptual viewpoint. In the following, we describe main steps of the proposed nonparametric approach, based on an iterative use of tree based classifiers for missing data imputation.

3. Proposed Approach

The basic idea is as follows: given a variable for which data are missing, and set of other $d(d < p)$ observed variables, the method works by using the former as response variable y and the latter as covariates $x_1, x_2, \ldots, x_d$. The resulting tree explains the distribution of the response variable in terms of the values of the covariates. Since the terminal nodes of the tree are homogeneous with respect to the response, they provide candidate imputation values for this variable. To deal with the data presenting missing values in many covariates, we introduce an incremental approach based on a suitably defined data preprocessing schema. The underlying assumption is the MAR process.

3.1 Missing Data Ranking

Let $\mathbf{X}$ be the $n \times p$ original data matrix, with d completely observed variables. For any record, if q is the number of covariates with missing data, which maximum value

is $Q = (p - d + 1)$, there might be up to $\sum_k \binom{Q}{q}$ type of missingness, ranging from the simplest case where only one covariate is missing to the hardest condition to deal with, where all the Q covariates are missing. We perform a two-way rearrangement of X, one with respect to the columns $(X_1, X_2, \ldots, X_d, \ldots, X_p)$ and one with respect to the rows $(1, 2, \ldots, m, \ldots, n)$. We propose to use a lexicographic ordering [C&T-91] [K&U-01] that matches the ordering by value, corresponding to the number of missing values occurring in each record. Practically, we form a string vector of length n that indicates the occurrence and the number of missing values for each row of $\mathbf{X}$. This allows to order $\mathbf{X}$ in a way that the first incomplete column $X_d (d \ll p)$ presents the lowest number of missing values and it follows the complete observed ones. Furthermore, columns also are ordered in the way that the first m rows (m £ n) contain instances with no missing values and the remaining $(n - m)$ rows present missing values. As a result, $\mathbf{X}$ is partitioned into four disjoint matrices asfollows:

$$\mathbf{X}_{n,p} = \begin{bmatrix} \mathbf{A}_{m,d} & \mathbf{C}_{m,p-d} \\ \mathbf{B}_{n-m,d} & \mathbf{D}_{n-m,p-d} \end{bmatrix}$$

Note that, as a consequence of the ordering schema, only $\mathbf{D}$ contains missing values while the other three blocks are completely observed with respect to their rows and columns.

3.2 Incremental Imputation

Once that the different types of missingness have been ranked and coded, the missing data imputation is iteratively made using tree based models. With respect to the records presenting only one missing attribute, a simple tree is used. Here, the variable with missing values is the response and the other observed variables are the covariates. The tree is built on the current complete data cases in $\mathbf{A}$ and the its results are used to impute the cases in $\mathbf{D}$. In fact, terminal nodes of the tree represent candidate Şimputed valuesŤ. Actual imputed values are get by dropping down the tree the cases of $\mathbf{B}$ corresponding to the missing values in $\mathbf{D}$ (for the variable under imputation), till a terminal node is reached. The conjunction of the filled-in cells of $\mathbf{D}$ with the corresponding observed rows in $\mathbf{B}$ generates new records which are appended to $\mathbf{A}$, that gains the rows whose missing values have been just imputed and a ŞnewŤ column corresponding to the variable under imputation.

For records presenting multiple missing values, trees are used iteratively. In this case, according to the previously defined lexicographic ordering, the tree is first used to fill in the missing values of the covariate presenting the smallest number of incomplete records. The procedure is then repeated for the remaining covariates under imputation. In this way, we form as many trees as the number of covariates with missing values in the given missingness. In the end, the imputation of joint missingness derives from subsequent trees. A graphical representation of both the data preprocessing and the imputation of the first missing value is given in figure 1.

A formal description of the proposed imputation process can be given as follows. Let $y_{r,s}$ be the cell presenting a missing input in the r-th row and the s-th column of the matrix

X. The imputation of this input derives from the tree grown from the learning sample $L_{r,s} = \{x_i^T, y_i; i = 1, \ldots, r-1\}$ where $x_i^T = (x_{i,1}, \ldots, x_{i,j}, \ldots, x_{i,s-1})^T$. Obviously due to the lexicographical order the imputation is made separately for each set of rows presenting missing data in the same cells. As a result, this imputation process is incremental, because as it goes on more and more information is added to the data matrix, both respect the rows and the columns. In other words, **A** is updated in each iteration, and the additional information is used for imputing the remaining missing inputs. Equivalently, the matrix **D** containing missing values shrinks after each set of records with missing inputs has been filled-in. Furthermore, the imputation is also conditional because, in the joint imputation of multiple inputs, the subsequent imputations are conditioned on previously filled-in inputs.

4. Concluding Remarks

Tree based classifiers appear particularly promising because when dealing with missing data because they enjoys two important properties:
a) they do not require the specification of a model structure;
b) they can deal with diőerent type of predictors (numerical and categorical). In this framework, we propose an automatic procedure that takes advantage of modern computing and allows to handle nonresponses in an easy and fast way, by defining a lexicographic order of the data. The results concerning the application of the proposed methodology on both artificial and real data set, not showed here due to the lack of space, confirm its eőectiveness, in most cases when data are nonlinear and heterosckedastic.

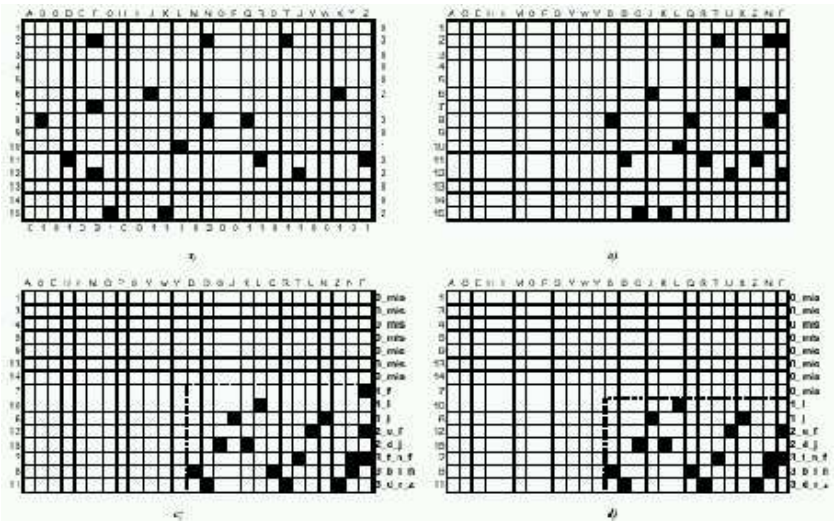


Figure 1. An illustration of the incremental imputation process: a) original data matrix; b) data rearrangement based on the number of missingness in each column; c) data rearrangement based on the number of missingness in each row and definition of the lexicographical order; d) the rearranged data matrix after the first iteration of the imputation process.

References

[**B-84**] Breiman, L., Friedman, J.H., Olshen, R.A., and Stone, C.J. (1984):*Classification and Regression Trees*. Monterey (CA): Wadsworth & Brooks.

[**C&T-91**] Cover, T.M., Thomas, J.A. (1991): *Elements of Information Theory*. New York: John Wiley and Sons.

[**K&U-01**] Keller, A.M., Ullman, J.D. (2001): *A Version Numbering Scheme with a Useful Lexicographical Order*: Technical Report, Stanford University.

[**L&R-87**] Little, J.R.A, Rubin, D.B. (1987):*Statistical Analysis with Missing Data*. John Wiley and Sons, New York.

[**S-98**] Sarle, W.S. (1998): *Prediction with Missing Imputs: Technical Report*, SAS Institute.

[**Sc-97**] Schafer, J. L. (1997): *Analysis of Incomplete Multivariate Data*. Chapman & Hall, London.

OUTLIER DETECTION (Friday 31st of May)

USING ROBUST TREE-BASED METHODS FOR OUTLIER AND ERROR DETECTION

Ray Chambers, Xinqiang Zhao, and Adao Hentges

Department of Social Statistics
University of Southampton
Highfield, Southampton, SO17 1BJ, U.K.

Editing in business surveys is often complicated by the fact that outliers due to errors in the data are mixed in with correct, but extreme, data values. In this paper we focus on a technique for error identification in such long tailed data distributions based on fitting outlier robust tree-based models to the outlier an error contaminated data. An application to a trial data set created as part of the EUREDIT project that contains a mix of extreme errors and "real" values will be demonstrated. The tree-based approach can be carried out on a variable by variable basis or on a multivariate basis. Intial results from both these approaches will be contrasted using this data set. Issues associated with "correcting" identified outliers in these data will also be explored.

DETECTING MULTIVARIATE OUTLIERS IN INCOMPLETE SURVEY DATA WITH THE EPIDEMIC ALGORITHM

Beat Hulliger and Cédric Béguin

Beat Hulliger
Swiss Federal Statistical Office
CH-2010 Neuchâtel
Switzerland
E-mail: Beat.Hulliger@bfs.admin.ch

The Epidemic Algorithm is a non-parametric method to detect multivariate outliers which is completely based on interpoint distances. It simulates an epidemic in a point cloud in p-dimensional space. The epidemic starts on the sample spatial median and then spreads through the point cloud with probabilities that decrease with the distance between points. Outliers typically are infected late. Thus the Epidemic Algorithm maps the multivariate space into time and outlying infection times are used to detect outliers. The cut-off for infection times to be considered outlying is arbitrary in a non-parametric context but it can be discussed under multivariate normality or other stochastic models.

Contrary to most of the established methods the complexity of the Epidemic Algorithm only depends linearly on the dimension p. However it depends quadratically on the number of points involved.

The adaption of the algorithm to missing values is straightforward by leaving out missing values from the distance calculations. The adaption to sampling is somewhat more involved since the infection probability in the population must be estimated from the sample.

The Epidemic Algorithm is applied to data sets of the EUREDIT project which contain missing values, sampling weights and various types of outliers. The choice of the infection probability function and its constants is discussed. The Epidemic Algorithm is an alternative to outlier detection methods which use the Mahalanobis distance and therefore assume an elliptical distribution of the bulk of the data. It can be seen as a data depth method.

DETECTING MULTIVARIATE OUTLIERS IN INCOMPLETE SURVEY DATA WITH THE BACON-EM ALGORITHM

Cédric Béguin and Beat Hulliger

Cédric Béguin
Swiss Federal Statistical Office
CH-2010 Neuchâtel
Switzerland
E-mail: cedric.beguin@bfs.admin.ch

The BACON (Blocked Adaptative Computationally-efficient Outlier Nominator) algorithm, one of the many forward search methods, is a very efficient outlier detection method in multivariate data with elliptical distribution. Starting from a small subset of good points BACON iteratively grows this good subset using Mahalanobis distances based only on the good observations. The largest Mahalanobis distances indicate the outliers when the growth of the good subset stops. The adaptation of BACON to complete survey data is straightforward by defining weighted estimates of the mean and the covariance matrix. Missing values are more problematic to deal with. The EM algorithm for multivariate normal data is used to evaluate the mean and the covariance matrix at each step of the BACON algorithm. The adaptation of EM to survey data is presented. The merging of both algorithms through the splitting of EM to use the advantage of the growing structure of BACON is discussed as well as the number of iterations of EM. The hypothesis on the missingness mechanism is the usual EM assumption, namely MAR (missing at random) data. Examples on well known datasets with challenging outliers are shown with up to 30% MCAR (missing completely at random) data. The BACON-EM algorithm is also applied to datasets of the EUREDIT project.

NEURAL NETWORKS
(Friday 31st of May)

EDIT AND IMPUTATION USING A BINARY NEURAL NETWORK

K. Lees, S. O'Keefe, and J. Austin

Advanced Computer Architectures Group
Department of Computer Science
University of York
Heslington
YORK YO10 5DD
UK

This paper describes a novel application of a binary neural network technology to the important practical task of statistical data editing and imputation. Editing and Imputation is used to improve data quality by most National Statistical Institutes and some commercial organisations. The paper describes how the AURA (Advanced Uncertain Reasoning Architecture) high-speed pattern matching system can be used to find a subset of data records similar to a given record. This can accelerate the processing of records with missing values and errors, allowing slower, conventional Euclidean distance based techniques to be used in the post-processing stage. A central part of the AURA system is the CMM (Correlation Matrix Memory) neural network. A binary version of the CMM is described, which has been studied at University of York for over 15 years, and some preliminary edit and imputation results are presented. The work at York is being carried out as part of the Euredit project. Euredit is supported under the European 5th Framework Programme, has 12 partners from 7 European states, and is investigating and comparing new and existing methods for edit and imputation, including neural network methods such as MLP, SVM, SOM, and CMM. The project is evaluating the performance of each method against a range of common datasets.

KERNEL METHODS FOR THE MISSING DATA PROBLEM

Hugh Mallinson and Alex Gammerman

Royal Holloway College
University of London
www.clrc.rhbnc.ac.uk

An imputation problem requires the completion of a dataset that is missing some values on some or all variables. A successful imputation action preserves the joint probability distribution of the dataset. We compare four imputation algorithms, a linear regressor, a group-mean imputation, a neural network and a Support Vector Machine (SVM). Our chief aim being to evaluate the SVM's performance. We artificially induce missing data patterns in three data sets; Boston Housing (BH), Danish Labour Force Survey(DLFS) and the Sample of Anonymised Records (SARS). Our performance measures include root-mean-square error and absolute error. We also compare the *full set* of imputations with the set of true values. This comparison (eg. using the Kolmogorov Smirnoff distance), measures how well the set of imputations preserves the marginal distribution as observed in the missing true values.

The Support Vector Machine is a new non-parametric prediction technique that has shown state-of-the art performance in some high-dimensional classificaton and regression problems, for example digit recognition and text retrieval. These algorithms exploit new regularisation concepts, e.g. the VC-Dimension, which control the capacity of models that are highly non-linear. SVMs require the solution of a convex quadratic program.

The imputation problem is in one sense harder than the standard prediction scenario as we must often restore values on more than one variable. Moreover during modelling or prediction of the j^{th} variable, one may have to use units that are lacking values on other variables. We propose a method that trains a model for each variable missing values and offer two approaches to selecting training sets for each of the models.

The experiments undertaken show SVMs to be a useful tool. On BH data the SVM rmse is best by a 5% margin. On DLFS the SVM has 5% lower root-mean-square-error. SARS results show the SVM to perform best relative to the others on scalar variables.

NEURAL NETWORKS FOR EDITING AND IMPUTATION

Pasi Koikkalainen

Laboratory of Data Analysis
University of Jyväskylä
P.O.Box 35, FIN-40351 Jyväskylä
Finland

The concern of this presentation is how neural networks can be used for editing and imputation.

In imputation tasks the promise is that neural networks can overcome the problem of "curse of dimensionality":
`Dense samples as needed to learn` **pdf** `well, but dense samples are hard to get in high dimensions.`

When using neural networks the dimensionality of data is not the problem, rather it is the COMPLEXITY of data. The self-organizing map, for example, combines dimension reduction and data modelling under a single learning algorithm. This allows us to model multivariate distributions with relatively effectively. The imputation model is then obtained by conditionlizing the modelled distribution by observed values.

In editing neural networks can be used for both strong and weak type of error localization. Strong knowledge assumes that errors can be modelled, while weak knowledge expects that we are able to discriminate between acceptable and erroneous observations. The use of weak knowledge is more common in neural systems. The objective is to build a model that explains well all clean observations, but which gives low matching probabilities for erroneous ones. This can be done in two ways:

i) Clean data is used for model building. As most models are based on mean values, also a measure of accepted spread around the model is needed.

ii) When no clean training data is available, robust methods must be used during training. Then, according some criteria, samples that are suspicious are given less weight, or totally ignored, from the model. As well as in case i) a measure of accepted spread must be computed before actual error (outlier) detection can be done.

SELECTIVE EDITING
(Friday 31st of May)

DEVELOPMENT OF A GRAPHICALLY ORIENTED MACRO-EDITING ANALYSIS SYSTEM FOR NASS SURVEYS

Dale Atkinson

USDA/NASS
3251 Old Lee Highway
Fairfax, VA 22030-1504
email: datkinson@nass.usda.

In 1997 the responsibility for the quinquennial census of agriculture was transferred from the U.S. Bureau of the Census (BOC) to the National Agricultural Statistics Service (NASS) in the U.S. Department of Agriculture. This fulfilled a goal of NASS to become the national source of all essential statistics related to U.S. agriculture, and provided an opportunity for the Agency to improve both the census and its ongoing survey and estimation program through effective integration of the two. However, the timing of the transfer severely limited the changes NASS could make to the procedures and processing tools for the 1997 Census of Agriculture (COA). Following the 1997 census, however, a team was assembled to reengineer its entire survey program, more effectively integrating the census. One of the immediate needs in effecting this integration was the development of a new processing system for NASS surveys. Considerable progress has been made on the specification and develo! pment of the new processing system, planned for initial use with the 2002 census.

This paper focuses on one of the key components of the new system the analysis module, which features a graphically oriented state-of-the-art macro-editing approach to identify problematic data values and correct erroneous data. The system contains considerable capability in identifying problems through statistical graphics and data roll-ups, with comparisons to administrative and historical data. All analysis screens provide drill-down capability to individual records, with linkages to the Agency's data warehouse for historical data and to its sampling frame database for name and address and sampling information. The module also provides tools for managing the analysis process, so that the analyst can efficiently perform all necessary review without omission or duplication.

DEVELOPING SELECTIVE EDITING METHODOLOGY FOR SURVEYS WITH VARYING CHARACTERISTICS

Pam Tate

Office for National Statistics, Room D140,
Government Buildings, Newport, NP10 8XG, UK
E-mail: pam.tate@ons.gov.uk

Recent research in ONS on data editing processes for business surveys has focused on developing new methods that would improve efficiency, without impacting adversely on data quality. A major element in this has been the development of suitable methodology for selective (or significance) editing - an approach that aims to reduce the amount of editing by concentrating the checking of suspicious data on cases where it is thought likely to have a material effect on the survey estimates.

Such a methodology was developed for the Monthly Inquiry for the Distribution and Services Sector (MIDSS), and successfully piloted, achieving a reduction of some 40% in validation follow-up with negligible effect on the survey outputs. The approach has accordingly been implemented on MIDSS, and is now being extended and adapted to other surveys.

This paper briefly describes the selective editing methodology for MIDSS and its effects. It then discusses the factors that need to be considered when extending and adapting the method to other surveys. Some of the key issues are illustrated by reference to the process of development of suitable methods for the Monthly Production Inquiry, Annual Business Inquiry and Monthly Wages and Salaries Survey. Lastly, further issues for investigation are described.

A Technical Framework for Input Significance Editing

Keith Farwell, Robert Poole, Stephen Carlton

Australian Bureau of Statistics
GPO Box 66A
Hobart 7001
Australia
Email: keith.farwell@abs.gov.au

The term "significance editing" refers to a general editing approach which incorporates survey weights and estimation methodology into edits and maintains a link between individual responses and output estimates. The Australian Bureau of Statistics (ABS) has been using significance editing in varying degrees over the last decade. When significance editing is applied at the input stage of a collection it is called 'input significance editing' within the ABS. Rather than using edits that result in messages such as "we should check that value", input significance editing uses a score to prioritise editing effort. The larger the score, the more significant the respondent reporting error is to possible bias in estimates. Respondents can be ordered or ranked in order of decreasing score giving a prioritised list of units to edit. This method was first applied within the ABS by superimposing it on the edit failures of an existing input editing system. Scores were generated for the edit failures and used to target editing effort towards a subset consisting of those most likely to provide the largest reduction of reporting bias on estimates. This approach has resulted in a noticeable improvement in editing efficiency.

This paper will outline further progress on input significance editing. It will outline progress on a technical framework for input significance editing which uses a model to describe the effect of editing unit record data. It looks at the joint problems of minimising reporting bias for a fixed cost and minimising cost to achieve a fixed level of reporting bias and shows that, in the case of uniform cost per unit, both problems have analytical solutions.

This paper will also outline the results of an empirical study set up to verify the applicability of the above framework to the editing of actual data from the Australian Survey of Employment and Earnings. The work investigates some of the practical aspects of estimating the model.

SELECTIVE EDITING BY MEANS OF A PLAUSIBLITY INDICATOR

Jeffrey Hoogland

Statistics Netherlands
Prinses Beatrixlaan 428
2273 XZ Voorburg
The Netherlands
jhgd@cbs.nl

In 2001 a new uniform system for the annual structural business statistics was implemented. An important part of the new system is the editing process, which makes use of selective editing. Four score functions are used to select questionnaires that contribute most to publication aggregates or that have an unexpected small influence on publication totals. Three other score functions are used to select questionnaires that are likely to contain large errors. The philosophy behind these score functions will be explained in detail.

The seven score functions are combined to one plausibility indicator (PI). When a form has a low score on this indicator the form is likely to contain influential errors. This form is therefore checked by a person and, if necessary, corrected. Forms with a large value for the PI are edited automatically with the software package SLICE, which has been developed at the Methods and Informatics department of Statistics Netherlands. If SLICE signals a violation of edit rules then regression imputation is used to alter values of variables.

At this moment the records for the annual structural business statistics of 2000 are edited. Information from clean data of 1999, like medians and publication aggregates are used to evaluate the plausibility of raw records of 2000. Also, clean data from 2000 from other sources are used, like short-term statistics and tax information. For a number of publication cells all records that have been edited automatically are also edited by a statistical analyst to evaluate the performance of the score functions and SLICE.

DEMONSTRATION OF THE GRAPHICAL EDITING AND ANALYSIS QUERY SYSTEM

Paula Weir

Department of Energy, EIA
Energy Information Administration, EI-42
1000 Independence Ave., S.W.
Washington, D.C. 20585
USA
email: paula.weir@eia.doe.gov

The Graphical Editing and Analysis Query System (GEAQS) is software developed by EIA as an in-house tool for editing and validating survey data. The graphical approach used in the system is based on best practices of four other systems examined at the beginning of the software development. GEAQS uses a top-down approach to examining data at the macro or aggregate level, highlighting questionable aggregates, and drilling down through lower level aggregates to identify the potential micro or reported data outlier. The graphical views include anomaly maps according to the various dimensions of the data, scatter plots, box-whisker plots, and time series graphs. In the recent version of the system, all of these graphical displays are available for both macro and micro views, with complementary metadata provided through accompanying spreadsheets with point-and-click mapping to the graphs. Outliers are identified by their position relative to the other respondents' values and,! in the case of scatter graphs, by their position relative to the fit line. The user can select the scale of the data, linear or log, to facilitate unclustering of data if necessary. The graphs display reported and imputed data with the data points colored according to user selected edit scores or measures of influence, allowing users to also evaluate the other edit or imputation rules through the graphic presentation. The original GEAQS PowerBuilder code has recently been rewritten to capture the capabilities of the newer PowerBuilder version, and to run on an Oracle database for the efficiency required for larger datasets, as necessitated by larger surveys and longer time series.

STOPPING CRITERION: A WAY OF OPTIMISING DATA EDITING AND ASSESSING ITS MINIMAL COST

Pascal Rivière

INSEE
18 Boulevard Adolphe, Pinard
75675 Pari Cedex 14
France

There is generally no scientific criterion to stop the manual checking of a survey: the editing process is stopped because everything has been checked, or because there is no time left for verification, or for other practical reasons. In this paper, we deliberately consider a simplistic goal for the editing process: ensuring, with a certain level of confidence, that the error rate falls below a certain threshold. For that purpose, we calculate approximate confidence intervals for the proportion of errors, and approximate predictive intervals for the number of remaining errors after checking and fixing part of the returns. Using the upper bound of the one-sided predictive interval, we can then easily define a stopping criterion: whenever the upper bound of the rate of remaining errors is greater that the target error rate, data editing continues; as soon as it falls below the threshold, we can stop editing. Such an approach allows to reduce the cost of data editing by avoiding unnecessary manual checks. The main results of this paper are in section 5, in which we define a relationship between the cost of editing and four main parameters: target error rate, number of target domains, number of returns, and level of confidence. In the last section, we examine the issues raised by the principle of stopping criterion, in order to generalise the criterion that we suggested.

ADDITIONAL PAPERS

THE APPLICATION OF OUTPUT STATISTICAL EDITING

Keith Farwell

Australian Bureau of Statistics
GPO Box 66A
Hobart 7001
Australia
Email: keith.farwell@abs.gov.au

The term "significance editing" refers to a general editing approach which incorporates survey weights and estimation methodology into edits and maintains a link between individual responses and output estimates. The Australian Bureau of Statistics (ABS) has been using significance editing in varying degrees over the last decade. When significance editing is applied at the output stage of a collection it is called 'output statistical editing' within the ABS.

Output statistical editing can be used to direct resources to those areas where editing effort is expected to have greatest benefit. Output editing involves a combination of detecting outliers, detecting any remaining significant reporting errors, and analysing the trends in estimates (such as movements for continuing surveys). Within a significance editing framework, we need to focus our attention on the actual weighted contributions of respondents to estimates.

Three separate initial output scores are created for a specific item based on contributions to the estimate, the movement, and the standard error. These are combined into a single 'item' score which can be used to order or rank respondent data in order of importance. Item scores can be further combined to produce a 'provider score' which can be used to derive an ordering of respondents for a variety of estimates. An output statistical editing strategy will allow collection areas to predetermine the amount of editing work they will do and to have a better understanding of that works impact.

This paper will outline results of recent investigations into the application of output statistical editing using data from ABS Agricultural collections. It will outline the development and refinement of the editing strategy and describe how it could be applied to other collections.

TREE BASED MODELS AND CONDITIONAL PROBABILITY FOR AUTOMATIC DERIVATION OF VALIDATION RULES

Photis Stavropoulos

Liaison Systems SA
77 Akadimias Str,
106 78 Athens
Greece
email: photis@liaison.gr

One of the most important steps in any survey that collects large amounts of dat a is (automatic) editing. The starting point in any editing application is a set of edits defined according to some check plan, i.e. a list of a "error sources" , by a group of subject matter specialists. The present paper is concerned with a recently proposed approach for the automatic derivation of edits from clean da tasets. We deal with validation rules (i.e. conditions that data must satisfy), which the approach views as conditional probability statements. In other words, a rule involving certain variables is seen as a statement of what are the most p robable values of some of them given the others. The approach proceeds by specif ying the domains of the variables involved in a given rule and then estimating t he conditional probabilities on this probability space. In this way, a generic v alidation treatment is created which is free from formally defining rules.

As a tool to practically estimate the probabilities we propose the use of segmen tation via tree based models. Suppose a dataset contains N cases described by K explanatory variables (numerical and/or categorical) and a response variable. Th e data are partitioned in an optimal way according to the values of the explanat ory variables and the result is a tree. Each node corresponds to certain values of the explanatory variables and contains cases with a certain distribution of t he response variable. This conditional distribution of each node is a validation rule.

In the paper we investigate ways of obtaining as many rules as possible and also ways of overcoming the theoretical and computational problems of the approach. The work presented is carried out under the INSPECTOR IST project on automatic d ata validation.