UNECE WP25 2003

Evaluating New Methods for Data Editing and Imputation Results from EUREDIT

John Charlton

http://www.cs.york.ac.uk/euredit/

The Euredit Project

March 2000 – February 2003

To develop and evaluate new edit and imputation methodologies alongside existing methods

- Develop evaluation criteria / framework for comparisons
- Produce standard datasets for experiments
- Establish a baseline = current good methods
- Develop and evaluate new methods
- Compare all methods to establish "best methods" for different data types
- Disseminate methods via software components and publications

Essentially a simulation experiment

Evaluation datasets (1)

For edit and imputation:

- **SARS** –1% of all GB households in 1991, hierarchical census data, mainly categorical
- **ABI** – UK Annual Business Inquiry - typical business survey with mixed numeric and categorical data
- **EPE** – Swiss environmental expenditure business survey, more challenging than the ABI because of large numbers of zeros
- **Times series data** - financial instruments/ share prices

For imputation only

- **Danish Labour Force Survey** linked to population register – the data came from registers but the missingness was created by using real non-response for those individuals who had not responded to the equivalent survey (income data not missing at random)
- **European Household Panel Survey Data**

Evaluation datasets (2)

Standard versions of datasets:

- True dataset (retained by ONS)
- Dataset with missing values (Y_2)
- Dataset with missing values & errors (Y_3)
- Separate subset of true dataset (for training – e.g. of neural network-type methods)

Evaluation software to calculate some 30+ formulae

Overall evaluation across methods, using statistical and operational criteria (best practice guidelines)

Evaluating Editing

- Measures of editing efficiency – detect as many errors as possible and avoid classifying correct values as errors Alpha   Proportion of false negatives/ false positives etc..
- Measures of influential error detection performance - based on size of errors D_{ij} in post-edited data.
- Measures of outlier detection performance (absolute relative errors of k-mean of moments)

Evaluating Imputation

- **Missing or suspicious values are replaced**

  - 5 levels of assessment:

  - Predictive accuracy (preserve true values, i.e. imputed close to real values)
  - Ranking accuracy (maximise preservation of order in imputed values)
  - Distributional accuracy (preserve distribution of true data – preserve marginal and higher order distributions)
  - Estimation accuracy (reproduce lower order moments of distributions of true values)
  - Imputation plausibility (acceptable – all logical edit rules should be satisfied)

Note:

(1) Ranking accuracy requires ordinal data, distributional/ estimation accuracy require scalar data, etc.

(2) Measures depend on scale of measurement, e.g. scalar, categorical, etc.

Evaluating E&I - **Operational characteristics**

- General features of system
  - Accept/ export data/ documentation/ versatility with different sources & data types
- Resource requirements
  - Knowledge/ skill required, software/hardware requirements, time taken, human intervention required
- Features indirectly affecting accuracy
  - Judgement required (choice of explanatory variables, edit rule design etc), help provided by system, dependency of results on expertise of user, time required to set up, pre-processing required, tools for validating output, e.g. visualisation
- Final output
  - Audit trail, ability to interpret changes made etc.

EUREDIT experimenters recorded operational characteristics of methods, to assist potential users

Standard methods (baseline - in NSI use)

Combined E&I systems - data changed

- CANCEIS (NIM) – Statistics Canada, nearest neighbour
- SCIA – ISTAT, inter-individual edits for household
- GEIS – Statistics Canada, Felleghi-Holt (continuous data)
- Cherry Pie (was Cherry Pi) +EC system - CBS Netherlands
  - Software differs in terms of the sort of data designed to handle, e.g. continuous/ categorical/ mixed

For imputation only - all types of data

- DIS (Donor Imputation System) - ONS

New methods (1)- multivariate outlier detection and outlier-robust imputation

- Methods:
  - Mahalanobis distance - robust covariance estimator
  - Growing “good” subsets: Kosinski; BACON; and Epidemic algorithms
  - Data depth: simplicial depth - multivariate M-quantiles
  - Tree-based methods: WAID - optimal partition
  - Regression methods incl. Robust calibration
  - Winsorisation and nearest neighbour imputation
  - Robust estimation and reverse calibration – values of Y which yield robust estimator

New methods (2) - Neural networks

- Advantages: Easy to use, make few assumptions about data, are flexible and resilient to noise. Train network on a small representative "clean" dataset, network "learns" from what experts did.
- **Neural-type Methods in EUREDIT**
  - **Multi-layer Perceptron** (MLP)- Classic neural network, previously tested for E&I (Nordbotten).
  - **Self Organising Maps (SOM)** - a NN which defines a mapping from input data space $\boldsymbol{R}^n$ onto a latent space consisting typically of a 2-dimensional array of nodes or neurons, giving imputation classes
  - **Correlation Matrix Memory (CMM)** - based on simplification of MLP algorithm – uses binary weights instead of continuous ones, implement k-nearest neighbour approach to edit and imputation - very fast
  - **Support Vector Machines (SVM) -** learn complex dependencies

New methods (3) - Panel &times-series methods

- BASIC METHODS: last-value carried forward, linear interpolation, Black-Scholes pricing, and standard term structure pricing of bonds.
- NEW METHODS: univariate and vector ARMA, linear and non-parametric regression and multilayer perceptron models for imputation.
- Since most of these methods utilise other time series as covariates, which themselves contain missing observations, the EM algorithm (Dempster, Laird and Rubin, 1977) is an appropriate tool.

Conclusions (1)

- No one method works best in all situations
  - Depends on the dataset and variable (e.g. scale of measurement, dependencies between variables, type of missingness)
  - Best methods usually capitalised on structure of data

- Some winners by dataset:
  - SARS: CANCEIS/SCIA nearest neighbour, neural networks worked reasonably well (SVM,CMM,MLP) but not T-SOM
  - ABI: outlier-robust methods, also T-SOM
  - EPE: Classical hot deck, and regression methods, logical edits
  - DLFS: T-SOM, also MLP, CMM, SVM
  - GSOEP: IMAI (statistical modelling approach)
  - TIMESERIES: nearest neighbour, neural networks

- "Black boxes" worked less well generally

Conclusions (2)

- Usefulness of pre-specified edits depends on method
- Good training data important for calibration and developing strategy -- keep “before” data for future work.
- Data should always be analysed prior to E&I to learn about relationships etc. Naïve users will not get maximum benefit from complex systems and may do better with simpler less efficient systems in default mode
- An editing strategy is likely to be a mixture of methods tuned to each particular dataset

Conclusions (3)

- **Promising new methods** are:
  - WAID: robust regression-tree models for skewed business survey data
  - Robust multivariate outlier detection methods (BACON-EM, EPIDEMIC algorithm, Transformed Rank Correlation) for skewed business survey data
  - T-SOM for a wide variety of surveys
  - MLP for imputation
  - CMM for very large datasets with minimum user intervention
  - SVM for imputation of categorical data
  - POEM and reverse calibration for data with outliers

General Outcomes

- Euredit web site: **http://www.cs.york.ac.uk/euredit/**
  - Statistical evaluation criteria
  - Coming soon: results papers
- 30-31 May 2002 "Data-Clean 2002" conference in Finland, **http://erin.it.jyu.fi/dataclean/** – initial results, papers on website

Now completed:

- Final report on evaluation of all methods (D6.1)
- User guide for E&I based on Euredit Findings (D6.2)
- Software to implement new methods (D7.1)
- Software to perturb data and apply statistical evaluation criteria
- Some standard data for future evaluations

D6.1 Report on results of experiments

1. Standard Methods - the benchmark for new methods
2. Robust Methods
3. MLP Neural Network Approaches
4. SOM Neural Network Approaches
5. CMM Neural Network Approaches
6. Support Vector Machines
7. Methods for Panel Data and Time-Series
8. Evaluation criteria
9. Technical appendices on details of project, e.g. preparation of data for experiments

D6.2 User guide: Towards effective E&I

1. Editing and Imputation Issues – to describe types of data, problems, and principles of edit and imputation
2. The Euredit Project - describes the project, the datasets chosen, the rationale for chosing these datasets
3. An Overview of Each of the Methods Tested
4. Overall Evaluation of Approaches Tested in EUREDIT - gives the results from a user's perspective, dataset by dataset
5. Recommendations Towards an Edit/Imputation Strategy