WP. 25

ENGLISH ONLY

UNITED NATIONS STATISTICAL COMMISSION and ECONOMIC COMMISSION FOR EUROPE

CONFERENCE OF EUROPEAN STATISTICIANS

Work Session on Statistical Data Editing
(Madrid, Spain, 20-22 October 2003)

Topic (ii): Developments related to new methods and techniques

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

Supporting Paper

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

I. INTRODUCTION

1. Advances in statistical and computer science have created opportunities for the application of newer technologies such as neural networks and outlier-robust methods for edit and imputation. The EUREDIT project, now completed, set out to assist users to make informed choices regarding the methods they use for automatic edit and imputation, by evaluating alternatives systematically, using real data. This paper describes in overview the findings of the EUREDIT project, which was funded under EU Fifth Framework research programme - a large multi-national collaboration involving twelve partners from seven countries (see http://www.cs.york.ac.uk/euredit). The project was based on real data and real problems encountered in official statistical data, and had the following objectives:

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

The project involved 12 partners from 7 countries, and its methods has been described previously - see http://www.unece.org/stats/documents/2002/05/sde/35.e.pdf.

2. The first stage of the project involved developing a framework for statistical evaluation (Chambers, 2001) – see Tables 1,2. This framework was then used within the project to evaluate both existing methods and those developed within the life of the project. It is envisaged that NSIs could also use the framework outside the Euredit project to evaluate other edit and imputation methods, including new techniques developed in the future. Operational characteristics of the methods examined (Tables 3 – 5) were also recorded, such as general features, resource requirements, judgement/ experience required etc. In this paper an overview of the findings, with selected examples, will be presented. Here we use the term "edit" to describe the process of detecting values or records with errors, and "imputation" as the process of correcting these errors or filling in holes in the data. The edit and imputation methods

[1] Prepared by John Charlton John.Charlton@ons.gov.uk

examined and their characteristics are described in Tables 1 to 2 below. The web version of the Euredit publication will be available at http://www.cs.york.ac.uk/euredit/CDindex.html.

II. STATISTICAL PERFORMANCE MEASURES USED IN EUREDIT EVALUATIONS

Table 1

Measures for evaluating performance of edit methods

Measures of editing efficiency – smaller values denote better performance	
Alpha	Proportion of false negatives resulting from edit process for variable j (errors that are accepted as valid by the edit process). Estimates the probability that the editing process does not detect an incorrect value.
Beta	Proportion of false positives resulting from edit process for variable j (correct values that the edit process identifies as errors). Estimates the probability that a correct value is incorrectly identified as suspicious.
Delta	Proportion of editing errors overall. Provides a global measure of the inaccuracy of the editing process.
A	Proportion of cases that contain at least one incorrect value and that pass all edits (false negatives)
B	Proportion of cases containing no errors that fail at least one edit (false positives)
C	Proportion of incorrect case-level error detections
G	Gini coefficient for measuring error localisation performance. N.B. only applicable to edit processes that assign probabilities of being in error to variables
Measures of influential error detection performance - based on size of errors in post-edited data. Smaller values denote better edit performance	
RAE	Relative average error (scalar variables only), the ratio of the mean of post-edit errors to the mean of the true values
RRASE	Relative root average squared error (scalar variables only), the ratio of the square root of the mean of the squares of the post-edit errors to the mean of the true values
RER	Relative error range (scalar variables only), the ratio of the range of post-edit errors to their inter-quartile range
Dcat	Categorical or nominal data measure of relative error – weighted proportion of cases where post-edit and true values disagree
t_j	t-test for how effective editing process has been for error reduction for variable j – values >2 indicate significant failure of edit process (continuous and categorical versions available)
Measures of outlier detection performance, smaller values denote better edit performance	
AREm1	Absolute relative error of the k-mean for 1st moment
AREm2	Absolute relative error of the k-mean for 2nd moment

Table 2

Measures for evaluating performance of imputation methods

Performance measures for predictive accuracy of imputation	
Categorical data	
D	Proportion of imputed cases where true values differ from imputed values. The smaller the better - ideal is zero.
Eps	Test statistic for preservation of true values in imputation, based on D
Dgen	Generalised version of D that takes into account the distances between categories
Continuous data	
mse	Mean square error from regressing true values on imputed values (zero intercept) using weighted robust regression - the smaller the better
t-val	t-statistic for testing slope=1 in above (smaller is better)
slope	Slope of regression line - should be close to 1
R^2	R^2 for above regression - proportion of variance in Y* explained by $\hat{Y}$
DL1	Mean distance between true and imputed values (L1 norm)
DL2	Mean distance between squares of true and imputed values (L2 norm)
DLinf	Distance measure between true and imputed values (L infinity norm – maximum distance between imputed and true values)
Performance measures of distributional accuracy of imputation methods	
Categorical variables	
W	Wald statistic for testing preservation of marginal distributions of categorical variables - distribution is chi-square with degrees of freedom =c (number of categories) for large n for stochastic imputation methods. Compares marginal distributions of imputed and true values.
Continuous (scalar) variables	
K-S	Kalmogorov-Smirnov for testing preservation of distribution (compares distributions of imputed and true values)
K-S_1	Alternative Kalmogorov-Smirnov for testing preservation of distribution using L1 norm (compares distributions of imputed and true values)
K-S_2	Alternative Kalmogorov-Smirnov for testing preservation of distribution using L2 norm (compares distributions of imputed and true values)
Performance measures for estimation accuracy	
m_1	Absolute difference between 1st moments of true and imputed values
m_2	Absolute difference between 2nd moments of true and imputed values
MSE	Evaluation of outlier-robust imputation. Mean square error of imputed values compared with true values
R_k	For time-series data, a measure of the relative discrepancy between estimated lag k auto-correlates for true and imputed values

III. EDIT AND IMPUTATION METHODS INCLUDED IN EUREDIT EVALUATIONS

3. The methods have been described in http://www.unece.org/stats/documents/2002/05/sde/35.e.pdf and in more detail in the volume *'Towards effective statistical Editing and Imputation Strategies – Findings of the Euredit Project"* (Charlton (ed), 2003).

Table 3. Characteristics of Edit and combined edit/imputation methods evaluated in EUREDIT

METHOD	CANCEIS	SCIA	GEIS	MLP neural networks	SOM/ NDA	CMM	IMAI	Cherry Pie[22]
Is it:								
Based on Fellegi-Holt?	No	Yes	Yes	No	No	No	No	Yes
Does method cover:								
Logical edit rules?	Yes	Yes	Yes	No	See[11]	No	No	Yes
Logical imputation rules?	Yes	Yes	Yes	No	See[12]	No	No	No
Does method require:								
Pre-specified edits?	Yes	Yes	Yes	No	No	No	No	Yes
Pre-specified parameters?	No	No	Yes	Yes	No	No[18]	No	No
Which parameters?	N/A	N/A	See [5]	See [8]	See[13]	See[19]	N/A	N/A
Pre-specified imputation rules	Yes	Yes	Yes	No	No	No	No	N/A
Other pre-specified imputation parameters	Yes	Yes[3]	No	No	No	See[20]	No	N/A
Training sample with raw values	No	No	No	Yes	No	No	No	No
Training sample with target values	No	No	No	Yes	No	No	No	No
Pre-process scaling of data?	Yes[1]	Yes[4]	No	Yes	Yes	No	Yes	No
Other pre-process transformation of data?	Yes[2]	No	Yes[6]	Yes[10]	See [14]	See[21]	Yes	No
Post process rescaling of data?	No	No	Yes[7]	Yes	See [15]	No	No	No
Post process other transformation of results?	No	No	No	No	No	No	No	No
Methodological experts?	Yes	Yes	Yes	Yes	See [16]	No	Yes	No
IT experts?	No	No	Yes	No	See [17]	No	No	No
Does it operate:								
Sequentially for each variable?	No	No	No	Yes	No	Yes	Yes	No
Simultaneously for set of variables?	Yes	Yes	Yes	Yes[9]	Yes	Yes	No	Yes
Types of variables dealt with:								
Categorical, nominal variables?	Yes	Yes	No	Yes	Yes	Yes	Yes	Yes
Categorical ordinal variables?	Yes	Yes	No	Yes	Yes	Yes	Yes	Yes
Continuous variables?	Yes	No	Yes	Yes	Yes	Yes	Yes	Yes

Table 4. Characteristics of imputation methods evaluated in EUREDIT

METHOD	DIS	Multi-variate regression	Hot-deck ratio	Hot-deck donor	EC System	Censor-ing	EM	Time Series methods[1]	SVM
Does method cover:									
Logical imputation rules?	No	No	Yes	No	Yes	No	No	No	No
Does method require:									
Pre-specified edits?	No	No	No	No	Yes	No	No	No	No
Pre-specified parameters?	No	No	No	No	No	No	Yes	Yes	No
Which parameters?	N/A	N/A	N/A	N/A	N/A	N/A	See [8]	See [2]	N/A
Pre-specified imputation rules	No	No	No	No	No	No	No	No	No
Other pre-specified imputation parameters	No	No	No	No	No	No	No	Yes [see 2]	No
Training sample with raw values	No	No	No	No	No	No	No	No	Yes
Training sample with target values	No	No	No	No	No	No	No	No	Yes
Pre-process scaling of data?	No	No	No	No	No	No	Yes	No	Yes
Other pre-process transformation of data?	No	No	No	No	No	No	No	Yes[3]	Yes[7]
Post process rescaling of data?	No	No	No	No	No	No	Yes	No	Yes
Post process other transformation of results?	No	No	No	No	No	No	No	Yes[4]	Yes
Methodological experts?	Yes	Yes	Yes	Yes	No	Yes	Yes	Yes[5]	No
IT experts?	No	No	No	No	No	No	No	No	No
Does it operate:									
Sequentially for each variable?	No	No	No	No	No	Yes	No	No	Yes
Simultaneously for set of variables?	Yes	Yes	Yes	Yes	Yes	No	Yes	Yes[6]	No
Types of variables dealt with:									
Categorical, nominal variables?	Yes	No	No	Yes	Yes	No	Yes	No	Yes
Categorical ordinal variables?	Yes	No	No	Yes	Yes	No	Yes	No	Yes
Continuous variables?	Yes	Yes	Yes	No	Yes	Yes	Yes	Yes	Yes

Table 5. Characteristics of outlier robust edit/imputation methods evaluated in EUREDIT

METHOD	TRC/ POEM	BEM/ POEM	EA/ POEM	Univariate WAID	Mutivariate WAID	Forward Search/ Regression Imputation
Is it:						
An edit method?	Yes	Yes	Yes	Yes	Yes	Yes
An imputation method?	Yes	Yes	Yes	Yes	Yes	Yes
Based on Fellegi-Holt?	No	No	No	No	No	No
Does method cover:						
Logical edit rules?[2]	No	No	No	No	No	No
Logical imputation rules?	No	No	No	No	No	No
Does method require:						
Pre-specified edits?	No	No	No	No	No	No
Pre-specified parameters?	Yes	Yes	Yes	Yes	Yes	Yes
Which parameters?	Tuning	Tuning	Tuning	Tuning	Tuning	Tuning
Pre-specified imputation rules	No	No	No	No	No	No
Other pre-specified imputation parameters	No	No	No	No	No	Yes[3]
Training sample with raw values	No	No	No	No	No	No
Training sample with target values	No	No	No	No[4]	No[5]	No
Pre-process scaling of data?[6]	Yes	Yes	Yes	Yes	Yes	Yes
Other pre-process transformation of data?	No	No	No	No	No	No
Post process rescaling of data?	No	No	No	No	No	No
Post process other transformation of results?	No	No	No	No	No	No
Methodological experts?	Yes	Yes	Yes	Yes	Yes	Yes
IT experts?	No	No	No	No	No	No
Does it operate:						
Sequentially for each variable?	No	No	No	Yes	No	Yes
Simultaneously for set of variables?	Yes	Yes	Yes	No	Yes	No
Types of variables dealt with:						
Categorical, nominal variables?	No	No	No	No	No	No
Categorical ordinal variables?	No	No	No	No	No	No
Continuous variables?	Yes	Yes	Yes	Yes	Yes	Yes

IV. DATASETS USED FOR EUREDIT SIMULATION EXPERIMENTS

4. A range of *standard datasets* was selected (Table 6), representative of the different types of data encountered by National Statistics Offices and other potential users of edit and imputation methods. The datasets included in EUREDIT experiments needed to be suitable for the evaluation of a wide variety of edit and imputation techniques and cover a range of data sources, such as social surveys, business surveys, time series, censuses and registers. Within each dataset, a range of error types and missingness was required, allowing the data to exhibit inconsistencies, non-response (item and unit), outliers and missingness.

5. The specific reasons for including particular datasets were:

***The Danish Labour Force Survey*:**
A combination of information sampled from the register from a population register combined with a true non-response pattern for income from a social survey (Labour Force Survey). The Income variable (known from the register) needed to be imputed for non-respondents to the survey. This represents a real pattern of non-response, and known missing values.

U.K. Annual Business Inquiry
A business survey (self-completion questionnaire) containing commonly measured continuous variables such as Turnover and Wages. It is currently edited through re-contact of cases that fail logical edits checks. Information was available regarding the types of errors and pattern of missingness.

Sample of Anonymised Records from the 1991 U.K. Census
A random 1% sample of household records from a census. This was the largest dataset in EUREDIT, containing information on people within households – a hierarchical structure. From Census

documentation, patterns of errors and missingness in the pre-edited data were recreated in the data distributed to participants.

Swiss Environment Protection Expenditures

A Business Survey containing some categorical variables plus mainly continuous data (expenditures), including a large number of true zero responses (i.e. where there was no expenditure), and outliers. The originators of the data themselves recreated the pattern of errors and missingness – missing in cases where data suppliers had to guess expenditure, and with errors as found in data as originally received.

German Socio-Economic Panel Survey

A social survey dataset, with a longitudinal aspect, consisting of information from a panel of people interviewed over a number of years. There is also an element of hierarchical data with information on people within households. Complete records were selected, and missing values for income were created according to the pattern of missingness in the full dataset.

Time Series Data for Financial Instruments

Financial time series, consisting of daily closing prices of over 100 stocks covering a time period of up to 5 years. This was the only dataset to contain time series information. The suppliers were also able to provide a simulated dataset for use in developing methods.

Table 6 Description of datasets used for EUREDIT evaluations

Dataset name	Type of dataset	Type of variables	Number of variables	Number of records
Danish Labour Force Survey (used for imputation only)	Administrative records with pattern of missingness from social survey.	Continuous variable for imputation (income), Ordinal, Nominal.	14	15,579
UK Annual Business Inquiry (ABI)	Business Inquiry Questionnaire	Mostly continuous (£000 sterling), 1 nominal (industry)	35	9,580
Sample of Anonymised Records from U.K.1991 Census (SARs)	Population Census	Categorical, Ordinal.	35	494,024
Swiss Environment Protection Expenditures (EPE)	Environmental Questionnaire	Continuous (SF 000), Binary, Categorical.	70	1,239
German Socio-Economic Panel Survey (GSOEP)	Panel Survey	Nominal, Ordinal, Continuous (income)	169	5,383
Times Series: Financial Instruments	Time Series	Continuous	124 time series	522 obs. per series

6. Treatment of datasets for evaluation purposes

Table 7 shows the notation used to describe the different versions of any single dataset. In the context of EUREDIT, a missing value is not an error, and is thus ignored in the evaluation of error detection - they are easily identified in the data and are the targets for imputation.

Table 7 – Notation to describe versions of datasets

Errors?	Missing?	
	Yes	No
Yes	Y_3	Y_1
No	Y_2	Y^*

7. The $\mathbf{Y}^*$ version of the dataset is assumed to be complete and without errors. For the purposes of the EUREDIT evaluations, 'true data' means data that the NSI provider considered to be satisfactorily cleaned by their edit and imputation procedures. One could also consider this as 'target data'. Version Y_2 (with missing values but no errors), and Y_3 (with missing values and errors) were distributed to partners for use in their experiments. No Y_1 (errors but no missing values) dataset was provided since it did not seem to represent a realistic situation.

8. The Danish Population Register/Labour Force Survey and GSOEP datasets each have two versions, Y*, Y_2, as they are to be used solely for imputation. The other four datasets have three versions: Y*, Y_2, Y_3, where Y_2 and Y_3 have different observation numbers for individual records to prevent potential disclosure of errors. For each dataset the Y* data were retained by the co-ordinator (ONS), and the perturbed data, Y_2, Y_3, were distributed to partners for edit and imputation

9. Developmental datasets
Some methods, particularly neural networks, need to estimate parameters from clean data. In real life situations such networks would learn from data that had been meticulously manually edited – usually a previous survey of the same type or a sample of the actual data. In order to develop and test prototype systems, six development datasets based on a small subset of each original dataset were provided for use with these methods. Each of these were available in the three versions:

- True data (Y*)
- Data with missing values but no errors (Y_2)
- Data with both errors and missing values (Y_3)

V. RESULTS

10. Altogether 191 experiments were run on the six datasets using the different methods, and each experiment involved a number of variables (see Table 6) and for each variable there might be around 30 evaluation formulae (Tables 1-2), for Y_2 and Y_3 versions of perturbed data. Thus there was a huge amount of information to synthesise in comparing methods. The full set of results will be published in the Project volume: *'Towards effective statistical Editing and Imputation Strategies – Findings of the Euredit Project"* (Charlton (ed), 2003). In this short paper it is necessary to be selective in presenting results, and we will present some results for just two of the six datasets.

Danish Labour Force Survey
11. The Danish Labour Force Survey (DLFS) contains one variable with missing values, which is income. All other variables contain full information. The missing values in the variable income are due to the fact that the respondent refused to participate in the survey or could not be reached at home. The income comes from Statistics Denmark's income register. All other variables are also found in registers. As income is known for all persons in the dataset it is possible to analyse the efficiency of the imputation concerning the bias of non-response. The following variables were available to assist imputation: Telephone/postal interview or neither; No. times interviewed in the panel; Telephone contact made; Postal follow up; Male/female; Age of respondent; Marital Status; Duration of education; Last employment; Employed/unemployed; Any children at home; Living with another adult; Area of residence. In all subgroups, except for persons aged 15-25, the persons who did not participate had lower incomes. This is attributable to characteristics associated with the non-response rate, e.g. persons with the lowest level of education and highest unemployed account for the highest non-response.
12. In order to describe the effectiveness of the imputation the linear regression model can, e.g. be estimated: $Y^* = b\hat{Y} + e$, here $\hat{Y}$ denotes the imputed value of Y and Y_i^* is the true value. It appears from Table 8 that for all 24 various imputations that have been tested on DLFS, a slope b, which is less than 1, are estimated. Thus, the imputation is not able to remove entirely the non-response bias.
The neural networks 'MLP quick 20' and SOM yielded the best estimation of the mean income ($\left|m(Y^*) - m(\hat{Y})\right|$), with total numerical errors of less than 1,000. But the CMM median, NN and SVM also yielded good results. However, the standard method linear regression (REG) is a comparatively good method, and is better than many of the other methods. Random errors are also imputed in cases where donor imputation is applied, reflecting the variation in the data material, which might suggest that the results achieved by e.g. RBNN and SOM with respect to the bias are not so good. This is especially true when looking at the slope and R^2. In the light of this, the results achieved by SOM are particularly good.

Another essential problem is the efficiency of the imputation used to describe the distribution and the variation in the material. The Kolmogorov-Smirnov distance (K-S) measures the largest difference between the two numerical distribution functions. The area between the two empirical distributions functions in first and second moment is also calculated. The area between the two distribution functions is lowest for SOM and NN, but also for "RBNN loglinear without noise" and "DIS without area". In general, SOM is superior with respect to the Kolmogorov-Smirnov distances. SOM is in general also among the best to minimize the mean error (m 1), but pays a price with respect to other measures of the quality of the imputation.

13. In summary, there is no clear winner, but "MLP quick 20", SOM and linear regression are the three best methods overall. In evaluating which imputation method is best suited, depends on the purpose of the imputation. It is therefore difficult to give a simple prescription, but it will be fair to say that SOM showed great strength. SOM was able to achieve total numerical errors of less than 1,000, which must be regarded as an impressive reduction of the bias in connection with non-response. Also, SOM was able to obtain good results with respect to imputing the variation in the data. In Table 9 the 5 best results on each measure have been underlined – some experiments have been omitted to save space.

Table 8 Results for various imputation methods on the DLFS data (some experiments omitted to save space)

EXPERI-MENT	METHOD	ESTI-MATION /DONOR	SLOPE	T VAL	MSE	R*R	DL1	DL2	DLINF	K S	K-S 1	K-S 2	M 1	M 2	MSE
DL21120	MLP quick 20	E	0.939	-14.1	6241867909	0.457	46538	78639	868315	0.0711	0.0128	0.00035	604	5466857151	1020321
DL21130	MLP dynamic	E	0.892	-27.6	6400848668	0.445	47599	80042	829097	0.2072	0.0232	0.00165	9551	3162220193	7528796
DL21140	MLP multiple	E	0.913	-21.0	6235500054	0.458	47018	78701	848750	0.0922	0.0182	0.00072	5288	4197224414	2995240
DL21150	MLP prune	E	0.893	-28.1	6292790808	0.454	46804	79329	822795	0.1126	0.0221	0.00116	8771	3138454147	6509871
YL20001	CMM n. neighbour	D	0.832	-35.2	9569009086	0.262	62391	102680	850380	0.0599	0.0109	0.00036	7290	888155314	4904610
YL20003	CMM mean	E	0.886	-28.8	6648029533	0.424	48754	81714	834028	0.1492	0.0261	0.00190	11499	3050703655	10460707
YL20005	CMM median	D	0.946	-13.9	6679189338	0.423	45132	81157	834249	0.1349	0.0218	0.00111	1132	6904997159	1047917
FL20001	N.neighbour	D	0.842	-28.1	10100000000	0.232	66379	103504	870720	0.0498	0.0082	0.00018	1322	1811656379	1197455
FL20005	RBNN log-linear w/o noise	D	0.835	-30.7	9882576471	0.250	64607	103779	880970	0.0517	0.0079	0.00018	4344	3745782	2446269
RL2002	SVM greedy bottom up	E	0.941	-14.7	6473310831	0.439	45113	79937	848858	0.0989	0.0180	0.00075	1400	6242058236	1109103
RL2003	SVM stratified	E	0.941	-14.1	6477665209	0.438	45695	79962	847997	0.0946	0.0185	0.00076	1446	6259512366	1117663
OL20001	DIS with all variables	E	0.807	-40.9	10200000000	0.224	64602	107385	870720	0.0822	0.0141	0.00067	11252	2286197959	10181466
OL20002	DIS without area vbl	E	0.834	-32.8	9658856881	0.243	63225	102042	869105	0.0580	0.0116	0.00037	6315	543320167	3927580
JL20003	SOM random donor	D	0.850	-27.3	10400000000	0.196	64992	104274	955189	0.0359	0.0080	0.00012	402	3611230060	1055018
JL20005	SOM n. neighbour	D	0.862	-28.6	9350027835	0.265	60267	98870	955189	0.0434	0.0076	0.00014	947	2533995253	1121543
Cl20001	REG Linear regression	E	0.922	-18.7	6352749474	0.449	46960	79278	836901	0.0771	0.0183	0.00063	3181	4974625603	1710695
DL21600	SOLAS Hot Deck	D	0.743	-44.7	13100000000	0.105	78753	124914	965724	0.0798	0.0137	0.00061	11553	3621201643	10704072

UK Census (SARS)

14. Editing

We present results for the census variables age, sex, relat (relationship to head of household), and mstatus (marital status). Tables 9 and 10 show the values for α and β respectively for each experiment where editing was carried out. For a good editing procedure both α and β should be small. Here we can see that the methods CANCEIS/SCIA and MLP achieved consistently low values for both α and β. The CANCEIS/SCIA, and MLP editing procedures show particularly good performance for the variable sex. Higher α values can be seen for the continuous variable age, but it should be noted that the perturbations for the SARS dataset included a large number of minor perturbations, for example, age may have been perturbed from 33 to 34, and most editing systems will ignore such minor perturbations as they are not

considered important. Relationship to household head did not have a strong relationship with other variables so performance as measured by α was less good here. Overall the probability of identifying a correct value as suspicious (β) is small for the CANCEIS/SCIA method. Table 11 shows the statistic δ (the probability of an incorrect outcome from the editing process) for the variables age, sex, relat and mstatus. For all variables and all editing methods δ is small, with the CANCEIS/SCIA and MLP methods achieving smaller values than SOM.

Table 9: Alpha values (probability of accepting errors as valid) for four SARs variables where editing has been applied.

Experiment	Method	Age	Sex	Relat	Mstatus
IS30001b	CANCEIS/SCIA	0.593281	0.078518	0.435005	0.243563
IS30003	MLP	0.630947	0.105027	0.312877	0.302392
JS30001	SOM	0.800808	0.114187	0.198952	0.446133
JS30002	SOM	0.582831	0.113751	0.184758	0.446133

Table10: Beta values (probability of identifying a valid value as an error) for four variables where editing has been applied.

Experiment	Method	Age	Sex	Relat	Mstatus
IS30001b	CANCEIS/SCIA	**0.004183**	**0.000275**	**0.000821**	**0.000255**
IS30003	MLP	**0.00751**	**0.000373**	**0.011732**	**0.001585**
JS30001	SOM	0.008246	0.000862	0.047294	0.000746
JS30002	SOM	0.057456	0.002457	0.054983	0.000746

Table 11: Delta values for four variables where editing has been applied.

Experiment	Method	Age	Sex	Relat	Mstatus
IS30001b (A)	CANCEIS/SCIA	0.045277	0.005354	0.028296	0.011676
IS30003 (A)	MLP	0.050999	0.007166	0.030788	0.015705
JS30001 (C)	SOM	0.063533	0.008218	0.056891	0.021653
JS30002 (C)	SOM	0.094104	0.009682	0.063195	NA

15. Imputation

We now assess the performance of the imputation processes. Imputation was carried out on both the Y2 and Y3 datasets. The Y2 dataset did not have errors in the data whereas Y3 was used to assess the ability to make imputations in the presence of errors. For categorical variables we can assess the *predictive accuracy* of an imputation procedure using the measure D. This measure gives, for each variable, the proportion of cases where the imputed value does not equal the true value. An imputation process with good predictive accuracy would achieve small values for D, ideally zero. The variable "Ltill" is whether or not the person has a limiting long-term illness.

Table 12: Measure of predictive accuracy, D, for four variables (Y3 data, with errors).

		D			
Experiment	Method	Sex	Relat	Mstatus	Ltill
IS30001a (B)	CANCEIS/SCIA	**0.24**	**0.11**	**0.18**	0.14
IS30002 (B)	MLP	**0.24**	0.17	**0.20**	**0.11**
RS3001 (B)	SVM	0.27	**0.09**	0.21	**0.12**
RS3005 (B)	SVM	0.27	0.11	0.21	**0.12**
RS3006 (B)	SVM	0.27	0.09	0.21	**0.12**
JS30001 (C)	SOM	0.45	0.68	0.48	**0.12**
JS30002 (C)	SOM	0.45	0.70	0.48	**0.12**
JS30004 (C)	SOM	0.45	0.70	0.48	**0.12**

16. For the Y3 data CANCEIS/SCIA, MLP, SVM do reasonably well in accurately predicting values but performance varies according to the variable imputed. SOM is less good, only slightly better than naive baseline methods (not shown here). As expected, for the Y2 data that are not contaminated with errors (Table 13) the performance is improved. CANCEIS/SCIA is overall best across the different variables. All methods do extremely well for the variable 'bath'. SVM also has good results for the

variables sex, mstatus and relat. However, SOM and IMAI/SOM have not performed particularly well for these variables.

Table 13: Measure of predictive accuracy, D, for four variables (Y2 data – no errors).

		D			
Experiment	Method	Sex	Relat	Mstatus	Bath
IS20001	CANCEIS/SCIA	0.23	0.05	0.16	0.0006
IS20002	MLP	0.23	0.15	0.17	0.0005
OS20001	DIS	0.33	0.35	0.32	0.008
RS2001	SVM	0.25	0.06	0.19	NA
RS2002	SVM	0.28	0.07	0.21	NA
RS2006	SVM	0.27	0.07	0.19	NA
YS20001	CMM	0.26	0.28	0.29	0.0005
JS20001	SOM	0.28	0.29	0.23	0.004
JS20002	SOM	0.34	0.24	0.36	0.014
JS20003	SOM	0.28	0.30	0.22	0.0009
FS20001	SOM/donor	0.30	0.25	0.34	0.0007
FS20002	IMAI/SOM	0.29	0.12	NA	NA

17. For the continuous variable Age we use the R^2, dL_2, m_1, m_2 and Kolmogorov-Smirnov statistics to assess imputation performance. Table 14 gives the results for R^2, dL_2, m_1 and m_2 for the Y3 data. The statistic R^2 should be close to one for a good imputation procedure. It can be seen that the methods CANCEIS/SCIA, MLP and SVM have values of $R^2 > 0.8$. For the method SOM however $R^2 <$ 0.5. We assess preservation of true values using dL_2. This is a distance measure, so smaller values indicate a better imputation performance. We can see that the methods CANCEIS/SCIA, MLP and SVM achieve the smallest values, in fact SVM performs very well, especially in terms of preserving first and second moments m_1 and m_2. SOM has the highest dL_2 values. For all methods m_1 indicates that the mean of the empirical distribution for age has been reasonably well preserved by the imputation procedures, apart from SOM. We can see how well the variance of the empirical distribution is preserved by using the statistic m_2. The methods CANCEIS/SCIA, MLP and SVM perform better than SOM. For the methods CANCEIS/SCIA, SVM and SOM the Kolmogorov-Smirnov statistic has values ranging from 0.01 to 0.07 confirming that these imputation methods have preserved the distribution for the variable age, but results for MLP and SOM are not as good.

Table 14: Values for selected imputation criteria for variable age (Y3 data - with errors).

Experiment	Method	R^2	dL_2	m_1	m_2	KS
IS30001a (B)	CANCEIS/SCIA	**0.850**	**9.04**	**0.24**	**21.94**	**0.0075**
IS30002 (B)	MLP	**0.853**	8086	0.83	60.69	0.1092
RS30001 (B)	SVM	0.75	11.51	1.15	**2.65**	0.0692
RS30005 (B)	SVM	**0.92**	**6.67**	**0.40**	65.35	**0.0280**
RS30006 (B)	SVM	**0.92**	**6.62**	**0.43**	66.40	**0.0292**
JS30001 (C)	SOM	0.51	16.63	3.74	544.86	0.1764
JS30002 (C)	SOM	0.37	17.64	3.08	514.98	0.1591

18. We now assess the imputation performance on the Y2 data that was error-free. Table 15 gives results for R^2, dL_2, m_1 and m_2 for the continuous variable age. Again CANCEIS/SCIA, SVM MLP and CMM perform well. The methods IMAI/SOM and SOM with regression also perform very well. SVM and CANCEIS/SCIA achieved the smallest dL_2 values and values of $R^2 > 0.9$. CANCEIS/SCIA, SVM, CMM, and SOM with regression are best in preserving the mean (m1) while CANCEIS/SCIA, SOM/donor and IMAI/SOM achieved the best results for preserving the raw second moment of the empirical distribution. The Kolmogorov-Smirnov statistic is in the range 0.01 to 0.09 for the methods CANCEIS/SCIA, SVM, CMM and IMAI/SOM confirming that these methods have preserved the distribution for age. MLP, and DIS also have low values (<0.13). In summary the methods

CANCEIS/SCIA, SVM, CMM, IMAI/SOM and SOM with regression show good performance for the imputation of the variable age, while the performance of the methods SOM and DIS was not as good.

Table 15: Values for selected imputation criteria for variable age (Y2 data - without errors).

Experiment	Method	R^2	dL_2	m_1	m_2	KS
IS20001	CANCEIS/SCIA	**0.926073**	**6.249487**	**0.171317**	**17.29991**	**0.00607**
IS20002	MLP	0.863691	8.482209	0.524163	148.7667	0.103244
OS20001	DIS	0.591329	17.45294	6.019617	593.8579	0.131801
RS2001	SVM	**0.946346**	**5.315706**	0.613262	79.23829	**0.036347**
RS2002	SVM	**0.937552**	**5.702528**	0.237599	58.01568	**0.024444**
RS2006	SVM	**0.942952**	**5.457148**	0.261328	36.08478	**0.022656**
YS20001	CMM	0.820548	9.721793	**0.025185**	46.32909	0.052261
JS20001	SOM	0.197077	24.05214	4.69175	335.1553	0.09484
JS20002	SOM	0.572164	14.92676	**0.181788**	234.4577	0.203346
JS20003	SOM	0.008383	29.13619	9.20166	574.4546	0.23152
FS20001	SOM/donor	0.865529	8.49288	**0.140562**	**6.962375**	**0.00613**
FS20002	IMAI/SOM	0.891531	7.638982	**0.152414**	**4.859234**	0.011724

19. In summary, for the imputation of the SARS dataset the donor method implemented in CANCEIS/SCIA performed better than the neural network methods. CANCEIS/SCIA was the best performer across all measures for the imputation of the continuous variable Age. SVM, CMM, IMAI/SOM and SOM with regression achieved good results on several of the measures. CANCEIS/SCIA and SVM may be better suited for imputation of datasets where most variables are continuous. The CANCEIS/SCIA and MLP editing procedures show promising results. It was apparent from the experiments that thorough exploratory analysis of the data is crucial to achieving a highly successful edited/imputed dataset and the extent to which this is done well will affect the results. The selection of appropriate matching variables and other tuning parameters may require many hours of analysis. In addition to this most systems require lengthy set up times and run times, but more time and expertise invested in preparation should result in higher quality imputed datasets.

VI. CONCLUSIONS

20. Conclusions

In an ideal world the series of experiments conducted in the course of the project would enable the identification of general procedures for editing and imputation of statistical data that would be "best" across a wide variety of data types, including census data, business survey data, household survey data and time series data. Not unexpectedly, the conflicting requirements and data types implicit in these different data scenarios meant that it was impossible to find a "one size fits all" solution to the many different editing and imputation problems posed within them. In real life situations it is likely that a mixture of solutions will be needed, tailored to characteristics of the dataset being processed. Overall the Euredit project was very productive, achieved most of its objectives, and many important lessons were learnt in the process of carrying out the research, developing new methods, and in the evaluation stages. Due to lack of space it has only been possible to present results for two of the six datasets included in the experiments, but full details are available in the project's publications.

VII REFERENCES

Charlton JRH, 2002. First Results from the EUREDIT Project. UNECE Work Session on Statistical Data Editing: http://www.unece.org/stats/documents/2002/05/sde/35.e.pdf

Chambers R, 2001. Evaluation criteria for statistical editing and imputation. NS Methodology Series No 28. See http://www.statistics.gov.uk/StatBase/Product.asp?vlnk=9227

Charlton JRH (Ed), 2003 (forthcoming volume). *Towards effective statistical editing and imputation strategies – findings of the EUREDIT project.* (see http://www.cs.york.ac.uk/euredit)

Footnotes to Table 3.
1. Missing value cannot be represented by a blank;
2. Data must be split into strata and imputation groups; the household head variables must be located in first position;
3. Yes optionally: Key variables, auxiliary matching variable, degree of fixity of the variables, marginal variables, max no of times that each donor can be used, max size of donor record;
4. The system requires positive integer coded data;
5. Variable weights, max cardinality of solutions, data groups/edit groups, matching variable, max no of times that each donor can be used, max allowed time to find solutions, Hiroglou-Berthelot algorithm parameters;
6. Data translation in order to avoid negative values;
7. If data have been translated in order to avoid negative values, back transformation of data is required;
8. Network topologies (no of hidden layers, no of neurons per layer, error function, activation function, training rate, stopping criteria etc.;
9. Yes but it is time and resources consuming;
10. Possible preparation of error indicators for the training phase;
11. SOM does not need, but it can be used with edit rules;
12 SOM does not need but it can be used with imputation rules;
13. Always: number of neurons, selection of variables, Imputation: method and related parameters (if any), Editing: sigma1 and sigma2 (for robustness);
14., 15. Depends on the data set;
16. Some understanding is recommended;
17. Depends on the software implementation used;
18. Parameters determined automatically by system, but can be overridden by user.;
19. K (the number of neighbours for K-NN processing), also the number of quantisation bins. Parameters determined automatically by system. User can override these;
20. Five "modes" for imputation are available. Default mode selected by system generally gives good results. ;
21. Data is represented in a CMM binary neural network to allow fast identification of similar matching records.;
22. Cherry Pie is the only method described in this table that is not designed to perform imputations;
23. AGGIES was also part of the software investigated, but we could not get it to work properly and NAS and others were not able to solve the problems.

Footnotes to Table 4.
1. LVCF, R1, NP100,MARX1, AR5X, MLP, BSBASE, BSLVCF, BSEM, BSMLP;
2. Choice of covariates, choice of dependent variables, and choice of training set and number of intermediate nodes in the case of MLP and BSMLP;
3. Yes, log returns of each time series;
4. Yes, inverse log returns with consistency checking;
5. Yes, except for the LVCF method;
6. No, but all methods impute sequentially over time. The R1, NP100, MARX1 and BSEM impute simultaneously for a set of variables;
7. SVM requires normalisation of scalar independent variables and the target variable if it is scalar. Categorical independent variables may require 1 of n encoding (also known as design variables).
8. Max iterations, and sometimes interactions fitted

Footnotes for Table 5.
2. All edit and imputation methods using POEM have the capacity to also use user-specified edit rules. However, these are not required for the methods to work.
3. Reverse calibration imputation requires specification of outlier robust estimate for variable being imputed.
4. Optimal tuning for univariate WAID error detection requires access to training sample or historical data with target values.
5. Optimal tuning for multivariate WAID error detection requires access to training sample or historical data with target values.

6. All methods require initial transformation of data to linearity to achieve optimal performance. With the ABI and EPE data, this was achieved via log transformation. None of the methods work well where there are many zero or "special" values in the data.