Euredit WP5.1 Internal Report No.1

DIS Evaluation Report (Draft March 2002)

Nargis Rahman - Office for National Statistics
Date April 9, 2002

Contents

Appendices

1 Introduction.

As part of the EUREDIT project the currently used methods for imputation will be tested and evaluated under work package WP5.1 with reference to data sets and variables selected in work package WP2. A functional evaluation of a number of edit and imputation packages has been reported in (Statistics Canada, 1999).

During the late nineties, the UK Census Office developed and tested a hot decking based editing and imputation system known as Donor Edit and Imputation System (DEIS) and the system was reported to show promising results in the context of the census data. It is planned to carry out a comprehensive evaluation of the DEIS and this will form a large part of ONS's contribution to work package WP5.1. As the system developed previously was very much census focussed, it is being re-developed and enhanced to be applicable to the wide variety of data sets and variables selected in EUREDIT. Due to a large amount of time spent on developing the imputation part of DEIS and the loss of a member of staff only the imputation part of DEIS will be evaluated and redeveloped as part of the EUREDIT project.

The purpose of this report is to describe the progress and evaluation of the development of the donor imputation system (DIS) at ONS. The next section gives a description of the donor imputation system being developed. Section 3 gives details of the data sets and the imputations carried out together with results. A summary is provided in Section 4.

2 Brief description of donor imputation system.

The donor imputation system is a variant of the hot decking method which searches and uses donors for imputing missing variables. The basic principle underlying the DIS is to search and use a single donor for all the missing variables of a recipient record. The method searches for a donor using a set of matching variables which are related to the missing variable(s) of the recipient record. The matching variables are used to calculate a statistical distance between recipient and donor records.

A donor is selected based on a statistical distance function. The donor is the one with minimum distance. If at the end of this stage a donor has not been found for a recipient, then the categories of each matching variable are collapsed and the search is repeated. If missing values are still present for recipient records then non-significant matching variables are removed in turn until only one matching variable remains.

There are two main stages in the implementation of DIS and these are:

searching and establishing a pool of suitable donors;

selection of the donor.

Several possibilities exist when more than one donor is available for a recipient. The simplest is to just use the first donor in the list, or one can randomly choose a donor from the available list. Multiple use of donors can be reduced by incorporating a penalty function for each use, see for example, (Yar, 1998).

To summarise, the donor imputation search algorithm is given by,

1. search for the donor using a set of matching variables;
2. search for the donor using the same set of matching variables but with collapsed categories of the variables;
3. remove non-significant matching variables one at a time and search for the donor as in steps 1-2.

As soon as a donor with minimum statistical distance has been found, the search process will be stopped. In the search algorithm, progression from a lower level to a higher level will take place only

if a donor with minimum statistical distance has not been found.

3 Evaluation of DIS

3.1 Danish LFS

This data set (lfs_dk3.csv) consists of administrative records with one record per individual. The data set consists of 14 variables of which only the income variable needs imputing. Missing values for the income variable were created for those individuals that did not respond to a social survey. The income variable is continuous while the matching variables are mostly categorical.

Bivariate scatter plots between the income variable and all potential matching variables were looked at to give an indication of the relationship between income and the other variables. Also, the Pearson correlation coefficient was calculated. Based on the results of the scatter plots and the correlation coefficient the following matching variables were chosen, business, age, marriage, sex, children, unemploy, cohabit, area and education.

In this version of DIS there are two measures of distance available for matching variables that are continuous. They are Euclidean distance and Manhattan distance. For categorical matching variables three types of distance, they are, simple matching, scaled rank difference and user defined distance matrix. Predictive mean matching (regression distance) can be used for any imputation variable that is continuous. See Section 3 in the software documentation and Appendix A in this report for details. The selected matching variables contain one continuous variable, age, the others are all categorical. Imputation was carried out using the Euclidean and Manhattan distances for matching variable age and simple matching for the other variables. Since the imputation variable, income, is continuous we also use the predictive mean matching option.

We applied the imputation performance measures for a scalar variable (Chambers, 2001). In this report we are only looking at measures for assessing the preservation of true values. We calculate the measures d_{L1} (absolute difference), d_{L2} (square root of the squared difference) and $d_{L\infty}$ (maximum absolute difference). The values for the measures d_{L1}, d_{L2} and $d_{L\infty}$ for Euclidean distance, Manhattan distance and predictive mean matching are given in Table 1. From the true data set the minimum, maximum, median and mean values of the income variable are given by 0.0, 769159.6, 127847.8 and 143348.1.

Table 1: Preservation of true values.

	d_{L1}	d_{L2}	$d_{L\infty}$
Euclidean	56216.88	96516.69	711668.8
Manhattan	56154.96	96411.28	711668.8
Regression	56225.80	96284.45	711668.8

These statistics are all distance measures hence a smaller value indicates that the imputed data set is closer to the true data set. The measures obtained for each distance function compare well with the ranges obtained from the true data. To assess preservation of distribution we look at the Kolmogorov-Smirnov distances KS, KS_1 and KS_2. For the three distance measures (Euclidean, Manhattan and regression) the Kolmogorov-Smirnov distances are 0.022, 0.007 and 0.00009. These values are close to zero indicating that the imputation method does preserve the distribution for the income variable.

3.2 UK SARS

This data set (newhhold(area 2)new.csv) is a 1% sample of households from the 1991 UK population census. All variables are catogorical with the exception of the variables age and hours which are continuous. For each record more than one imputation variable may exist. This data set includes responses which are 'Not applicables' for some variables.

The principal behind DIS is to use a single donor for all imputation variables, hence it is necessary to select a group of matching variables that will lead to the selection of a suitable donor record for all imputation variables. By assessing bivariate scatter plots and Pearson correlation coefficients, matching variables for each of the SARS variables are selected. A combined set of matching variables is selected from the individual sets by choosing the most frequently occuring variables amongst house hold variables and person specific variables. We look at three sets of matching variables. The first set (set 1) consists of persinhh, age, sex, relat, mstatus, isco2, qualevel, hhstype, roomsnum and tenure. Set 2 consists of persinhh, age, sex, mstatus, relat, isco1, hours, qualevel, isco1, hhstype, roomsnum and tenure. Set 3 consists of persinhh, sex, age, mstatus, relat, econprim, hhstype, roomsnum and tenure.

Imputation was carried out using the three sets of matching variables. For continuous matching variables we use Euclidean distance and for categorical variables we use simple matching. We also use the user defined distance option for matching variable mstatus in set 3. We apply the evaluation criteria (Chambers, 2001) for assessing the preservation of the marginal distribution for a categorical variable. For the continuous variables we assess the preservation of the true values as in Section 3.1. Results are presented for the variables age, sex, mstatus, relat, ltill, tenure and bath in Table 2 to Table 8 respectively. For matching variables in set 3 there are two sets of results, one using simple matching (set 3a) and one using user defined distances (set 3b).

Table 2: Age, preservation of true values.

	d_{L1}	d_{L2}	$d_{L\infty}$
Set 1	11.98	16.82	89
Set 2	13.56	18.58	92
Set 3a	10.21	15.08	91
Set 3b	10.15	15.05	92

The statistics in Table 2 are distance measures and a smaller value indicates that the imputed data set is closer to the true data set. From Table 2, we can see that the best imputation results for the variable age are achieved using matching variables in set 3, that is, persinhh, sex, age, mstatus, relat, econprim, hhstype, roomsnum and tenure. The minimum, maximum, median and mean values from the true data are 0, 95, 36 and 37.45 respectively. We can assess the preservation of distribution by looking at the Kolmogorov-Smirnov statistics. For this variable the Kolmogorov-Smirnov statistics, KS, KS_1, KS_2, are 0.13, 0.06 and 0.006 respectively. These values are close to zero and indicate that the imputation method does preserve the distribution.

Table 3: Sex, preservation of the marginal distribution.

	W	D	ϵ
Set 1	69.97	0.36	0.33
Set 2	30.20	0.33	0.30
Set 3a	59.56	0.36	0.33
Set 3b	33.93	0.34	0.31

Table 4: Mstatus, preservation of the marginal distribution.

	W	D	ϵ
Set 1	235.98	0.33	0.30
Set 2	233.58	0.35	0.32
Set 3a	212.32	0.32	0.29
Set 3b	135.08	0.30	0.28

Table 5: Relat, preservation of the marginal distribution.

	W	D	ϵ
Set 1	73.99	0.34	0.31
Set 2	51.86	0.32	0.29
Set 3a	64.96	0.32	0.28
Set 3b	57.56	0.30	0.27

Table 6: Ltill, preservation of the marginal distribution.

	W	D	ϵ
Set 1	19.06	0.21	0.17
Set 2	19.64	0.22	0.19
Set 3a	11.28	0.19	0.15
Set 3b	13.55	0.19	0.16

Table 7: Tenure, preservation of the marginal distribution.

	W	D	ϵ
Set 1	103.71	0.59	0.56
Set 2	83.03	0.59	0.56
Set 3a	94.72	0.58	0.56
Set 3b	84.31	0.58	0.56

Table 8: Bath, preservation of the marginal distribution.

	W	D	ϵ
Set 1	1.92	0.0065	0
Set 2	0.16	0.007	0
Set 3a	0.31	0.008	0
Set 3b	0.5	0.007	0

For an imputation variable with m+1 categories, the statistic W follows a chi-square distribution with m degrees of freedom. From Table 8 we can see that the marginal distribution for the variable bath is preserved for all sets of matching variables. The best imputation results are achieved using matching variables in set 2. For the other categorical variables the W statistic suggests that the marginal distributions have not been preserved. One reason for this could be that this data set contains a large number of responses which are "Not Applicable" which can make it difficult to find suitable donors.

We also present the cross classification of actual versus imputed counts. The results for variables sex (set 2), mstatus (set 3), tenure (set 2) and bath (set 2) are given in Table 9 to Table 12 respectively.

Table 9: Cross classification of actual vs. imputed counts, sex.

	1	2
1	22500	463
2	646	24094

Table 10: Cross classification of actual vs. imputed counts, mstatus.

	1	2	3	4	5
1	18786	137	24	43	25
2	187	19368	152	72	31
3	18	224	2559	14	6
4	66	61	11	2211	17
5	99	192	31	58	3311

Table 11: Cross classification of actual vs. imputed counts, tenure.

	1	2	3	4	5	6	7
1	9171	208	14	21	9	11	96
2	73	23490	22	30	13	10	134
3	9	25	1294	2	0	1	8
4	26	36	1	1289	3	0	19
5	4	21	2	1	887	1	6
6	14	24	2	3	1	871	21
7	135	207	16	25	3	14	9430

In the above tables rows represent the imputed data and columns represent the true data. The percentage of correct imputations for the variables sex, mstatus, tenure and bath is 67, 68, 41 and 99 respectively. In general, the donor imputation system performs reasonably well for household variables such as bath but less well for individual variables. This is probably due to using a combined set of matching variables for the imputation. To achieve a high rate of correct imputations it is essential to choose appropriate matching variables.

3.3 UK Annual Business Inquiry

This data (sec297(y2).csv and sec298(y2).csv) set contains responses to selected questions from the UK Annual Business Inquiry for two sectors for the years 1997 and 1998. There are two questionnaires, the short version only asks for summary information. Values for variables from questions that are not on the short form are set to -9 for businesses that answered the short questionnaire. All variables are continuous and there are many imputation variables.

Table 12: Cross classification of actual vs. imputed counts, bath.

	1	2	3
1	47517	8	5
2	7	110	0
3	6	0	50

A combined set of matching variables was chosen using the same method as for the SARS data set in Section 3.2. Again we look at three sets of matching variables. Set 1 consists of purins, purtele, empni, assacq, stockend, turnover, purhire, purtrans, purothse, employ, stockbeg and empwag. Set 2 consists of purhire, empni, empens, purins, stockend, turnove, stockbeg, assacq, purtele and purothse and set 3 consists of stockend, empwag, turnover, purins, purhire, assacq and empni. For the 1997 data set there are a total of 31 variables of which 25 require imputing and for the 1998 data set there are a total of 34 variables of which 28 require imputing.

We carry out imputation using Euclidean distance for the three sets of matching variables and apply the evaluation criteria for assessing the preservation of true values. We present results for the variables turnover, emptotc, purtot, taxtot, assacq and assdisp. For the 1997 data set the results for the measures d_{L1}, d_{L2} and $d_{L\infty}$ are given in Table 13, Table 14 and Table 15 respectively. The minimum, maximum, median and mean values from the true data set is given in Table 16.

Table 13: Preservation of true values, d_{L1} 1997 data.

	Set 1	Set 2	Set 3
turnover	12612.76	21688.21	21179.84
emptotc	2364.83	1832.13	2237.18
purtot	3557.36	6975.07	2314.61
taxtot	746.74	951.77	1016.77
assacq	240.80	219.93	230.56
assdisp	51.17	32.50	35.32

Table 14: Preservation of true values, d_{L2} 1997 data.

	Set 1	Set 2	Set 3
turnover	50500.85	90303.96	90592.93
emptotc	16028.92	10674.08	15997.11
purtot	21814.07	54853.24	12018.87
taxtot	6014.43	6337.86	6364.34
assacq	562.59	673.63	698.84
assdisp	154.34	75.46	89.34

For the 1997 data we can see from Table 13 to Table 15 that matching variables in set 1 give the best imputation results for variables turnover, taxtot and assacq. Matching variables in set 2 give the best imputation results for variables emptotc and assdisp and matching variables in set 3 give the best imputation results for variable purtot. For each variable the Kolmogorov-Smirnov statistics are close to zero indicating that the imputation method preserves the distributions of these variables.

For the 1998 data set the results for the measures d_{L1}, d_{L2} and $d_{L\infty}$ are given in Table 17, Table 18 and Table 19 respectively. The minimum, maximum, median and mean values from the true data set is given in Table 20.

Table 15: Preservation of true values, $d_{L\infty}$ 1997 data.

	Set 1	Set 2	Set 3
turnover	453406	768996	768996
emptotc	153911	101797	153911
purtot	197732	502004	107378
taxtot	56333	56333	56333
assacq	3933	7439	7439
assdisp	1126	443	560

Table 16: Ranges from the true 1997 data set.

	min	max	median	mean
turnover	0	7486000	2500	34970
emptotc	0	555200	294.5	2026
purtot	0	7467000	1859	28290
taxtot	0	3297000	10	2425
assacq	-9	105200	25	431.9
assdisp	-9	63470	0	106.3

Table 17: Preservation of true values, d_{L1} 1998 data.

	Set 1	Set 2	Set 3
turnover	63224.78	75983.93	75642.07
emptotc	1295.93	1309.97	710.59
purtot	3960.47	3703.73	3272.80
taxtot	2290.01	2294.29	2307.10
assacq	926.74	983.13	754.85
assdisp	352.07	454.83	372.47

Table 18: Preservation of true values, d_{L2} 1998 data.

	Set 1	Set 2	Set 3
turnover	505153.83	524120.2	519399.56
emptotc	6267.68	6283.58	2519.58
purtot	18670.33	17556.97	16214.83
taxtot	21432.42	21432.26	21437.28
assacq	5461.42	5481.61	4809.24
assdisp	2987.59	3128.19	2992.35

Table 19: Preservation of true values, $d_{L\infty}$ 1998 data.

	Set 1	Set 2	Set 3
turnover	5106831	5214528	5172477
emptotc	62026	62026	18885
purtot	156627	156627	156627
taxtot	221317	221317	221317
assacq	52250	52250	47621
assdisp	36127	36127	36127

Table 20: Ranges from the true 1998 data set.

	min	max	median	mean
turnover	0	6679000	2344	28550
emptotc	0	161100	298	1741
purtot	0	5197000	1750	23510
taxtot	0	4586000	11	1791
asasacq	-9	72980	24	373.1
assdisp	-9	45440	0	75.86

For the 1998 data we can see from Table 17 to Table 19 that matching variables in set 1 give the best imputation results for variables turnover, taxtot and assdisp and matching variables in set 3 give the best imputation results for variables emptotc, purtot and assacq. Again for each variable the Kolmogorov-Smirnov statistics are close to zero indicating that the imputation method preserves the distributions of these variables.

3.4 Swiss EPE.

This data set (epe93a(y2).csv) consists of a questionnaire distributed in 1993 to enterprises in Switzerland. The enterprises were chosen according to class of economic activity. The data set consists of information on expenditure relating to environmental issues. The data set contains 70 variables which are responses to the questionnaire plus additional general business questions. There is a mixture of continuous and categorical variables.

As in Section 3.2 we obtain a combined set of matching variables. We look at three sets of matching variables given by, set 1: rectot, totinvwp, totinvap, totinvot, totinvto, totexpwm, totexpnp,totexpto, netinv and curexpto, set 2: recot, totinvwm, totinvnp, totinvto, totexpwp, totexpap, totexpot, totexpto, exp93 and curexp and set 3: rectot, recot, curexpto, curexp, totexpto, and totinvto. Out of the 70 variables 51 required imputing. We use Euclidean distance for coninuous matching variables and simple matching for categorical matching variables.

For the continuous variables we assess the preservation of true values using the distance measures d_{L1}, d_{L2} and $d_{L\infty}$. In this report we present results for the variables totinvto, totexpto,subtot and rectot. The results for the measures d_{L1}, d_{L2} and $d_{L\infty}$ are given in Table 21, Table 22 and Table 23 respectively. The minimum, maximum, median and mean values for each variable from the true data set is given in Table 24.

From Table 21 to Table 23 we can see that matching variables in set 3 give the best imputation results for variables totinvto, totexpto and subtot, while matching variables in set 1 give better imputation results for variable rectot. Note that for variable subtot the performance measures using matching variables in set 1 and set 3 are equal possibly indicating that both sets of matching variables lead to equally good imputed data sets.

Table 21: Preservation of true values, d_{L1}.

	Set 1	Set 2	Set 3
totinvto	1722.05	1802.11	1505.16
totexpto	916.42	978.61	693.97
subtot	15	120	15
rectot	660.09	742.73	741.27

Table 22: Preservation of true values, d_{L2}.

	Set 1	Set 2	Set 3
totinvto	3679.57	3606.60	2877.93
totexpto	2722.15	2633.14	2272.19
subtot	15	159.45	15
rectot	1818.27	1882.24	1881.69

Table 23: Preservation of true values, $d_{L\infty}$.

	Set 1	Set 2	Set 3
totinvto	11820	11391	8298
totexpto	12820	10870	10870
subtot	15	225	15
rectot	6000	6000	6000

Table 24: Ranges from the true Swiss EPE data set.

	min	max	median	mean
totinvto	0	90260	0	1026
totexpto	0	190500	2	2001
subtot	0	5000	0	44.13
rectot	0	37540	0	222.3

3.5 German Socio-economic Panel Data.

This data set (clgsoep(m).csv) is a selection from the German household survey for people who participated in the survey over the years 1991 to 1996. For each year there are 30 education and employment variables for each participant plus identification variables. Out of the 30 variables, 4 require imputing. Note that not all of the 4 variables are missing in all six years.

Matching variables were obtained for each of the 4 variables after assessing bivariate scatter plots and the Pearson correlation coefficients. We wish to exploit the lonitudinal aspect of this data set by using the previous years data to match on if it is available. For example if income in 1996 is missing but is present for all previous years then we would use the previous years income variables as matching variables in the search for a donor. For this reason a single donor to impute all missing variables in a record is not appropriate, so for this data set we impute using individual donors for each imputation variable. The most common matching variables are wegen, ausb, erwz, betr, oeffd, iscoh, branch, sex, bilzeit and PBB02. The variables that require imputing are continuous. For this data set we only consider one set of matching variables for each imputation variable.

For the continuous variables we assess the preservation of true values using the distance measures d_{L1}, d_{L2} and $d_{L\infty}$. We present results for variables income91, income 96, houseinc91 and houseinc96. Results for the variables income and houseinc are given in Table 25. The ranges from the true data set for variables income and houseinc are given in Table 26.

Table 25: Preservation of true values, German Panel Data.

	d_{L1}	d_{L2}	$d_{L\infty}$
income91	27245.04	55573.48	419384
income96	23826.93	37943.22	167400
houseinc91	43929.58	71447.53	420884
houseinc96	45196.39	62227.63	224844

Table 26: Ranges from the true GSOEP data set.

	min	max	median	mean
income91	0	97920	20390	24850
income96	0	480000	20400	32250
houseinc91	0	495900	56250	56980
houseinc96	0	480000	56000	61320

From Table 25 and Table 26 we can see that the measures d_{L1}, d_{L2} and $d_{L\infty}$ compare well with the ranges from the true data confirming that the imputation method does preserve true values for this data set.

3.6 Preparation and run time

All programs were run on a Dell Precision 420 Pentium III machine. An imputed data set is produced in two stages. The first stage involves identification of the donor values and the second stage involves replacing the missing values with the donor values. There are two programs from NAG that carry out the two stages. The donor values (stage 1) are found using program GeDaM and replacement of missing values (stage 2) is via the program ApplyEdits.

For the Danish LFS data set GeDaM took 15 minutes to run. The ApplyEdits program took less than 1 minute to produce the imputed data set. Before the programs can be run details of the variables and distance functions to use need to be specified in the options file. For the Danish LFS data set the preparation of the options file took 1 hour. For the SARS data set GeDaM took 1 hour to run and the ApplyEdits program took 1 minute. The preparation of the options file took 1.5 hours. For UK ABI, Swiss EPE and GSOEP data sets GeDaM and ApplyEdits both took only 1 minute to run. The preparation of the options files took 2 hours, 3 hours and 4 hours for the UK ABI, Swiss EPE and GSOEP data sets respectively.

Further preparation is needed before the options file can be set up. It is necessary to select the matching variables which often requires a good knowledge of the data set. Basic statistical analysis such as scatter plots and calculation of correlation coefficients may be necessary. The user also has to select the distance measure for each variable and weights/scaling factors. Depending on the number of variables and the complexity of the data set these preparations may take more than one day.

At present the options file is time consuming to set up. Improvements may be necessary to speed up the process.

4 Summary

The current DIS system finds a single donor for all imputation variables in a record but also has an option for allowing a different donor for each imputation variable. There are a choice of distance functions for categorical and continuous matching variables. Current results indicate that the donor imputation system gives good results when a suitable set of matching variables is used. Comprehensive statistical analyses of the data set may be necessary to obtain a good set of predictors for each imputation variable. Good knowledge of the data set is also necessary.

References

Chambers, R. (2001). Evaluation criteria for statistical editing and imputation. *National Statistics Methodological Series. 28.*

Rahman, N.J. and Morgan, G. (2001). *DIS Software Documentation.* Office for National Statistics and NAG.

Statistics Canada (1999). *A functional evaluation of edit and imputation tools*, UN/ECE Work Session on Statistical Data Editing. Statistics Canada.

Yar, M. (1998). The development of the donor imputation system (DIS). Technical report, Office for National Statistics.

A Distance functions

In the following definitions y^r represents a matching variable from the recipient record and y^d represents a matching variable from a potential donor record.

A.1 Euclidean distance

$$\mathrm{d} = (y^r - y^d)$$

A.2 Manhattan distance

$$\mathrm{d} = |y^r - y^d|$$

A.3 Regression distance

The regression distance obtains predictions from the regression model built using the matching variables as covariates. At present only a linear model is available. Predictions are obtained for non-missing and missing variables. The prediction for each missing variable is compared with the predictions for the non-missing variables to find a match. The imputed value is then the true value from the matched record.