Euredit WP5.1 Internal Report No.1

DIS Evaluation Report (Final September 2002)

Nargis Rahman - Office for National Statistics
Date September 26, 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 and lfsn_dk2.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.

We first give results for the development data set lfs_dk3.csv. 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. For the development data set 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. This table also shows the imputed (i) and true (t) medians and means for the income variable.

Table 1: Preservation of true values.

	d_{L1}	d_{L2}	$d_{L\infty}$	imedian	imean	tmedian	tmean
Euclidean	56216.88	96516.69	711668.8	163000	177700		
Manhattan	56154.96	96411.28	711668.8	162900	177700	160140.5	175385.5
Regression	56225.80	96284.45	711668.8	163000	177700		

The first three statistics are all distance measures hence a smaller value indicates that the imputed data set is closer to the true data set. The three distance functions used to impute the Danish LFS all give similar imputed data sets. 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. The medians and means obtained from the imputed data sets are similar to those obtained from the true data.

The same set of matching variables were used to impute the evaluation data set. The predictive mean matching option was used to carry out the imputation. The Kolmogorov-Smirnov statistics are close to zero indicating preservation of the distributions. The values for the measures d_{L1}, d_{L2} and $d_{L\infty}$

are 63225.37, 102042 and 869105 respectively. These values are larger than those obtained from the development data indicating that true values for the evaluation data set have not been preserved as well as for the development data set. This may be because the evaluation data set only contains 15579 records so finding a suitable donor is more difficult. The development data set contained 200000 records.

3.2 UK SARS

This data set (newhhold(area 2)new.csv and newhholdm.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.

We give results for the development data (newhhold(area 2)new.csv) first. 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, ltill and bath in Table 2 to Table 6 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}	$d_{L\infty}$	imedian	imean	tmedian	tmean
Set 1	11.98	16.82	89	36	37.92	36	37.45
Set 2	13.56	18.58	92	36	38.02	36	37.45
Set 3a	10.21	15.08	91	36	37.84	36	37.45
Set 3b	10.15	15.05	92	36	37.85	36	37.45

The first three 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. 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. The table also shows the imputed (i) and true (t) medians and means and we can see that the imputed medians and means are very similar to the true medians and means.

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: 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 6: 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 6 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) and bath (set 2) are given in Table 7 to Table 9 respectively.

Table 7: Cross classification of actual vs. imputed counts, sex.

	1	2
1	22500	463
2	646	24094

Table 8: 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 9: Cross classification of actual vs. imputed counts, bath.

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

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 and bath is 67, 68 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 which may have made it more difficult to find a suitable donor. To achieve a high rate of correct imputations it is essential to choose appropriate matching variables.

We now present results for the evaluation data set. The imputation was carried out using matching variables in set 3b. For all variables the imputation process does preserve the distributions. We use the W statistic to assess whether or not the distribution is preserved and in this report we present results for mstatus, sex and relat in Table 10.

We can see from Table 10 that for variables mstatus, sex and relat the W statistic follows the appropriate χ^2 distribution indicating that the distributions have been preserved. The imputation method preserves true values if $\epsilon = 0$. This was the case for variables insidewc, ltill, mstatus, qualnum, residst, termtim and workplce. The imputation results were greatly improved for the evaluation data set compared to the development data set. When the development data set was imputed very few distributions were preserved. One reason for this could be that the evaluation data set is 10 times

Table 10: Evaluation criteria statistics, evaluation data set.

	W	D	ϵ	p-value
mstatus	1.90	0.15	0	> 0.1
sex	0.67	0.41	0.21	> 0.1
relat	17.88	0.48	0.29	> 0.1

larger (492472 records compared to 47773 records for the development data) and hence a larger donor pool is available.

3.3 UK Annual Business Inquiry

This data set (sec297(y2).csv, sec298(y2).csv and sec198(y2).csv) 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.

A combined set of matching variables was chosen using the same method as for the SARS data set in Section 3.2. For the development data sets 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 first provide results for the development data (sec297(y2).csv and sec298(y2).csv). 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 $\mathrm{d}_{L\infty}$ are given in Table 11, Table 12 and Table 13 respectively. The imputed and true medians and means are given in Table 14.

Table 11: Preservation of true values, d_{L1} 1997 data.

	Set 1	Set 2	Set 3
turnover	2251.78	4252.31	2320.87
emptotc	286.83	267.29	359.97
purtot	4017.28	9856.29	2282.69
taxtot	62.91	87.05	97.72
assacq	73.03	67.89	63.29
assdisp	17.86	11.89	11.28

For the 1997 data we can see from Table 11 to Table 13 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. Table 14 shows the true medians and means and the medians and means from the three imputed data sets. The means and medians from the imputed data sets are similar to those obtained from the true data set.

Table 12: Preservation of true values, d_{L2} 1997 data.

	Set 1	Set 2	Set 3
turnover	12202.78	21050.79	18915.68
emptotc	3752.35	2542.05	3802.47
purtot	27433.55	69607.42	14912.95
taxtot	1452.28	1600.16	1617.11
assacq	232.36	227.67	217.73
assdisp	93.62	34.08	35.85

Table 13: Preservation of true values, $d_{L\infty}$ 1997 data.

	Set 1	Set 2	Set 3
turnover	17890.26	30342.65	30342.65
emptotc	6652.73	4400.12	6652.73
purtot	6816.77	17306.48	3701.83
taxtot	2633.05	2633.05	2633.05
assacq	149.45	282.67	282.67
assdisp	41.44	16.30	20.61

Table 14: True and imputed medians and means, 1997 data set.

	median	median 1	median 2	median 3	mean	mean 1	mean 2	mean 3
turnover	2500	2482	2481	2500	34970	35070	35300	35290
purtot	1859	1835	1835	1835	28290	28140	28070	28210
taxtot	10	10	10	10	2425	2411	2416	2417

For the 1998 data set the results for the measures d_{L1}, d_{L2} and $d_{L\infty}$ are given in Table 15, Table 16 and Table 17 respectively. The imputed and true medians and means are given in Table 18.

Table 15: Preservation of true values, d_{L1} 1998 data.

	Set 1	Set 2	Set 3
turnover	20949.21	22213.26	22534.57
emptotc	169.81	170.56	123.95
purtot	2378.68	2359.12	2404.61
taxtot	4351.05	4351.86	4354.03
assacq	107.41	109.78	77.76
assdisp	30.67	37.39	29.50

Table 16: Preservation of true values, d_{L2} 1998 data.

	Set 1	Set 2	Set 3
turnover	306537.6	313847.6	311310.3
emptotc	1599.79	1603.13	681.16
purtot	18295.61	18225.41	18165.45
taxtot	30802.05	30801.99	30802.51
assacq	1424.60	1431.61	1186.31
assdisp	710.94	744.68	711.51

Table 17: Preservation of true values, $d_{L\infty}$ 1998 data.

	Set 1	Set 2	Set 3
turnover	231182	236057.4	234153.8
emptotc	2957.55	2957.55	900.48
purtot	6857.24	6857.24	6857.24
taxtot	11833.68	11833.68	11833.68
assacq	2472.12	2472.12	2253.10
assdisp	1588.92	1588.92	1588.92

For the 1998 data we can see from Table 15 to Table 17 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. Table 18 shows the true medians and means and the medians and means from the three imputed data sets. The means and medians from the imputed data sets are similar to those obtained from the true data set.

For the evaluation data set, we look at two versions y2 and y3. The y2 version contains just missing values and all other values are assumed to be correct. The y3 version contains missing values and errors. Results for the y2 evaluation data set are better than those obtained from the development data. The Kolmogorov-Smirnov statistic is close to zero for all variables except assacq, assdisp and capwork. So for these variables the distributions have not been preserved. The statistics d_{L1}, d_{L2} and $d_{L\infty}$ indicate that the imputed values are similar to the true values for most variables. Table 19 shows results for the variables turnover, purtot and taxtot.

For the y3 data set the imputation method does not give good results for most variables. This is as

Table 18: True and imputed medians and means, 1998 data set.

	median	median 1	median 2	median 3	mean	mean 1	mean 2	mean 3
turnover	2344	2343	2346	2341	28550	27600	27510	27530
purtot	1750	1752	1756	1756	23510	23370	23360	23340
taxtot	11	10	11	10	1791	1798	1798	1797

Table 19: Evaluation criteria statistics, evaluation data.

	d_{L1}	d_{L2}	$d_{L\infty}$	d_{KS}
turnover	1113.56	47506.73	79181.99	0.11
purtot	70.54	1647.71	2547.33	0.06
taxtot	8.03	97.29	127.22	0.07

expected since the y3 data set contains errors that have not been corrected by any editing process. In practice, dirty data is cleaned by an editing procedure prior to applying the imputation process.

3.4 Swiss EPE.

This data set (epe93a(y2).csv and epe93na(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. Again for the development data 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.

We first give results for the develoment data. 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 20, Table 21 and Table 22 respectively. The imputed and true medians and means are given in Table 23.

Table 20: Preservation of true values, d_{L1}.

	Set 1	Set 2	Set 3
totinvto	517.83	713.88	1520.29
totexpto	1001.16	1104.08	457.74
subtot	15	120	15
rectot	249.18	417.20	416.86

Table 21: Preservation of true values, d_{L2}.

	Set 1	Set 2	Set 3
totinvto	1687.71	1970.37	2832.83
totexpto	2218.12	2171.44	1395.59
subtot	15	159.45	15
rectot	904.45	1102.85	1102.42

Table 22: Preservation of true values, $d_{L\infty}$.

	Set 1	Set 2	Set 3
totinvto	954.56	919.92	670.13
totexpto	1672.97	1418.50	1418.50
subtot	7.85	117.80	7.85
rectot	1423.03	1423.03	1423.03

Table 23: True and imputed medians and means, Swiss EPE data set.

	median	median 1	median 2	median 3	mean	mean 1	mean 2	mean 3
totinvto	0	0	0	0	1026	1028	977.9	1070
totexpto	15.5	12	12	12	1850	1716	1704	1752
subtot	0	0	0	0	44.13	43.68	44.88	43.68
rectot	0	0	0	0	222.3	218.9	210.3	210.5

From Table 20 to Table 22 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 23 shows the true medians and means and the medians and means from the three imputed data sets. We can see that the means and medians from the imputed data are similar to those obtained from the true data.

For the evaluation data set we present results for totinv, totexpto, subtot and rectot. Table 24 shows results for the statistics d_{L1}, d_{L2}, $d_{L\infty}$ and the Kolmogorov-Smirnov statistic (d_{KS}).

Table 24: Evaluation criteria statistics, evaluation data.

	d_{L1}	d_{L2}	$d_{L\infty}$	d_{KS}
totinvto	127.44	323.82	332.22	0.43
totexpto	39.31	117.92	200.70	0.13
subtot	1.44	2.08	1.92	0.5
rectot	41.49	132.69	232.69	0.81

The Kolmogorov-Smirnov statistic indicates that for most variables that distributions have not been preserved. This data set contains many observations that are zero hence it is difficult to find suitable donors. Also it is a small data set, only 1039 records, so finding a suitable donor is more difficult.

3.5 German Socio-economic Panel Data.

This data set (clgsoep(m).csv and gsoep(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.

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

	Development data				Evaluation data			
	inc91	hinc91	inc96	hinc96	inc91	hinc91	inc96	hinc96
d_{L1}	21952.23	44980.45	23652.08	42615.04	13171.87	37529.73	22640.82	47495.84
d_{L2}	46892.6	73382.96	37622.35	66577.29	22321.61	51859.89	40047.33	70178.13
$d_{L\infty}$	389900	454634	202000	404296	249150	595045	533200	532290
d_{KS}	0.09	0.10	0.07	0.05	0.021	0.10	0.076	0.15

From Table 25 we can see that the Kolmogorov-Smirnov statistic is close to zero for the development data and for the evaluation data indicating that distributions have been preserved. However, the donor imputation method did not 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.

Table 26 shows the time it took for the GeDaM program to run for the development (D) data sets and the evaluation (E) data sets.

Table 26: Run times for the GeDaM program.

	Number of records	GeDaM run time (minutes)
Danish LFS (D)	200000	180
Danish LFS (E)	15579	2
SARS (D)	47773	60
SARS (E)	492472	5760
ABI98 (D)	5594	1
ABI98 (E)	6233	1
EPE (D)	200	1
EPE (E)	1039	1
GSOEP (D)	704	1
GSOEP (E)	5383	2

For all data sets the ApplyEdits program took 1 minute to run. 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, SARS, UK ABI, Swiss EPE and GSOEP data sets, the preparation of the options files took 1 hour, 1.5 hours, 2 hours, 3 hours and 4 hours 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 and when a large donor pool is available. 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. DIS performs well for data sets such as SARS and the Danish LFS but for business data sets DIS does not perform very well.

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.