STATISTICAL COMMISSION and Working Paper No. 22
ECONOMIC COMMISSION FOR EUROPE English only

CONFERENCE OF EUROPEAN STATISTICIANS

UN/ECE Work Session on Statistical Data Editing
(Cardiff, United Kingdom, 18-20 October 2000)

Topic III: New techniques and tools for editing imputation

EVALUATION OF SOLAS 2.0 FOR IMPUTING MISSING VALUES

Submitted by Statistics Denmark[1]

Contributed paper

I. INTRODUCTION

1. Imputation means filling missing values in a data set. At Statistics Denmark we collect a lot of data, some of which have missing values. For this reason we have been evaluating SOLAS versions 2.0. SOLAS is a programme for imputation, but SOLAS also has many facilities for statistical analyses. We have limited ourselves to an evaluation of the imputation facilities, because SOLAS is one of the only commercial programmes with imputation tools.

2. In surveys with many questions, respondents will often forget (or be unable) to answer one or more questions. Ideally we would not have any non-response, but when faced with non-response, we have to decide if the missing values should be imputed or weighted by post-stratification. If there is only one missing value in a questionnaire with 1000 questions it would be natural to impute this value, but if only one out of 1000 questions had been answered, it would probably be useless to impute the remaining values. The purpose of the data collection should also be taken into consideration when imputing. If imputation is done in a good way, it leads to easier statistics and better conclusions, but it can also lead to confusion and biased results.

3. If the probability that a response is missing depends not only on the observed data, but also on the unobserved data, the imputation algorithms here cannot be used. If the missing data have an entirely different structure than the observed data, normal statistics and conclusions about the missing data cannot be made easily. For example, if you ask people about their smoking habits, smokers will be more reluctant to answer this question. It is also much easier for a non-smoker to answer how many cigarettes he smoked last week, because he knows that he smoked none. There can thus be a great difference in willingness and difficulty in answering a question.

II. DEDUCTIVE IMPUTATION

4. Before considering the imputation algorithms available in SOLAS, in some cases one should consider whether it is possible to perform "deductive" (or logical) imputation. If this is the case, it should be done prior to using other imputation algorithms.

5. The most basic case of imputation is the case that the missing value can be determined by certainty or almost certainty from the observed data, e.g. if we know a girl is 5 years old we can be certain she has no children. Likewise, if a total is missing but the subtotals are not missing the total can

[1] Prepared by Bjørn Steen Larsen and Birger Madsen.

easily be generated. In these cases the rules for handling missing values should be implemented directly without using SOLAS.

6. If the girl were 25 years old instead of 5 years old there would not normally be a perfect imputation algorithm, but we might have some idea that 0,1,2,3 would be a likely number of children, but 10 would be a very unlikely number and 100 would be an impossible number. It is in cases like this where the missing value cannot be determined from the auxiliary information, that imputation tools in SOLAS can be used.

III. IMPUTATION ALGORITHMS

7. In the following, we will describe the imputation algorithms available in SOLAS 2.0. Most standard imputation algorithms can be divided into two groups:

(1) Most likely/mean imputation algorithms.

This is where the missing value is replaced by an estimate of the missing value without adding noise. In some cases this is the best imputation method. In SOLAS this is implemented by the following algorithms:

- **Group Mean Imputation** - Imputed values are set to the variable's group mean (or mode in the case of categorical data).

- **Predicted Mean Imputation** - Imputed values are predicted using an ordinary least-squares multiple regression algorithm when the variable to be imputed is continuos or ordinal. When the variable to be imputed is a binary or categorical variable, a discriminant method is applied.

- **Last Value Carried Forward** - The last observed value of a longitudinal variable is imputed. This algorithm is only for longitudinal surveys.

(2) Random imputation algorithms.

This is where one tries to replace the missing values by values drawn from the assumed distribution of the missing values. This will often be a better way to impute, if the imputed data later have to be basis for an analysis, because it is less biased. In SOLAS this is implemented by the following algorithms:

- **(Random) Hot Decking** - Imputed values are selected (usually at random) from respondents who are similar with respect to a set of auxiliary variables. A problem with hot-decking occurs if many auxiliary variables are available. It is often impossible to find a donor which matches all the auxiliary variables, so the user must choose the order of importance of the auxiliary variables, which is not always easy either. On the other hand, statistical analyses can aid in the decision of the order of importance of the auxiliary variables.

- **Predictive Model Based Method – Multiple Imputation** - This is the Multiple Imputation analogy of the Predicted Mean Imputation algorithm. Each parameter is randomly drawn from the posterior distribution. This ensures that the uncertainty of the unknown true model is reflected. The Predictive Model Based Method can be a good way to avoid the problems mentioned above of many auxiliary variables.

- **Propensity Score Method – Multiple Imputation** - The system applies an implicit model approach based on Propensity Scores and an Approximate Bayesian Bootstrap to generate the imputations. The propensity score is the estimated probability that a particular element of data is missing. The missing data are filled in by sampling from the cases that have a similar propensity score. The multiple imputations are independent repetitions from a posterior predictive distribution for the missing data, given the observed data.

These algorithms are now described in turn. Only short descriptions are given here. For complete descriptions see the "SOLAS 2.0 User Reference".

IV. MULTIPLE IMPUTATION VS. SINGLE IMPUTATION

8. By repeating an imputation a number of times the effect of the chosen imputed value on the variance can be taken into account. Other imputation algorithms always impute the same value, thereby ignoring the variance associated with the imputation process. Multiple imputation is the only way of estimating the total variance, which includes the variance within imputations.

V. DATASET FOR EXPERIMENTS

9. The **Danish Labour Force Survey (LFS)** is designed as a sample of approximately 15,600 individuals each quarter, who are sampled using the Danish Central Population Register. Interviewing is carried out mostly by phone, but a follow-up using postal interviews is also used. A sample of **individuals** is chosen (instead of a household sample), because we have access to register data on individuals. In this way, the clustering effect of selecting a household can be avoided, thus yielding estimates with a smaller variance. The panel design of the Danish LFS can be described as a **rotating panel**, where the sample in each quarter is divided into three rotation groups, which allow changes between successive quarters as well as between successive years will be estimated with a smaller variance, compared to independent samples. This is a stratified sample with **two strata**. Approximately 5,000 individuals are sampled from persons who were **registered unemployed** in the last previous quarter for which such information was available. The remaining persons were drawn from persons **registered employed** in the last previous quarter. This procedure gives an adequate number of unemployed in the sample.

10. In the evaluation of SOLAS we have ignored the stratum structure in the imputation process. The rotating panel structure has also be ignored because we only use data from one quarter. The following auxiliary variables and categories were used:

A	Age	Continuous, i.e. 16-67 years			
B	Business (Economic activity)	Construction	Other private	Public	N.A.
C	Children living at home	No	Yes		
E	Education (Vocational)	None	Apprentice	Higher education	
G	Geography/Urbanisation	Metropolitan area	Provincial cities	Rural areas	
I	Income (gross) in 1000 DKK	Continuous, 0-1000			
S	Partner: Yes or no	No	Yes		
M	Married: Yes or no	No	Yes		
U	Unemployed: Yes or no	No	Yes		
N	Interviewed: Yes or no	No	Yes		
X	Sex	Female	Male		

11. Of the original 15,600 persons who had been selected for the **LFS,** 21 had emigrated or died before interviewing started, so Statistics Denmark needed interviews from 15,579 persons. Our interviewers managed to get 11,404 interviews and 4,175 non-response. There are a few very extreme

incomes, so income above a 1,000,000 DKK per year has been truncated to 1,000,000 DKK. Less than 20 values are truncated in this way.

12. From registers, we have auxiliary information about income for all 15,579 persons. We have in this example imputed income for the 4,175 non-response persons, so we have defined the income variable as missing in all cases where there was no interview. The knowledge about who was interviewed is used to get a more realistic knowledge about non-response than just to simulate (artificially) non-response.

V.1 Imputations

13. We have performed imputations with the following set-up:

- **Group Mean.** SOLAS allows only one grouping variable. We have chosen Sex.

- **Predicted Mean Imputations**. We used the following variables as auxiliary variables in the linear regression: Sex, Age, Geography, Education, Business (Economic activity), Married, Partner, Children, Unemployed

- **Random Hot-Decking.** At most five criteria could be selected in SOLAS for selecting donors. We chose: Sex, Education, Business (Economic activity), Married and Children

- **Predictive Model Based Method.** Sex, Age, Geography, Education, Business (Economic activity), Married, Partner, Children, Unemployed

- **Propensity Score Method.** Sex, Age, Geography, Education, Business (Economic activity), Married, Partner, Children, Unemployed

- **Last Value Carried Forward (LVCF).** This algorithm is only used for longitudinal data. Since we only used data from one quarter, we did not test this algorithm.

- **Evaluation process.** To make evaluations of the imputation from the different algorithms we have looked at the distribution of the imputed income compared with the real income. We have also calculated the correlation between the actual value and the imputed value as well as the covariance between the auxiliary information and the imputed values.

14. If two independent samples are drawn from the same population, we would expect that they are different but that the overall distributions and correlations should be maintained.

V.2 Results

Table 1. Mean and Standard Deviation. Correlation between imputed and actual value. (Income in/1000 DKK)

	Mean	Std	Correlation
Actual	145	104	--
Group Mean	179	32	0,16
Predicted Mean Imputation	160	67	0,51
Hot Decking	164	130	0,16
Predictive Model Based Method	158	109	0,29
Propensity Score	164	108	0,03

15. All algorithms compute values close to the real value. The imputed values of the two most likely/mean imputation algorithms have much lower Standard Deviations than those of the real income

distribution. Of the three random imputation algorithms, the Predictive Model-Based Method performs best, because it has the highest correlation with the actual values, a realistic standard deviation and smaller bias.

Table 2. Covariance between the auxiliary variable and the actual/imputed incomes.

Variable	Actual income	Group Mean	Predicted Mean	Propensity score	Hot Decking	Predictive model-based method
Sex	-8,60	-15,70	-12,60	2,44	**-10,20**	-12,20
Age	367,00	-32,40	409,00	24,00	93,00	**393,00**
Employment	4,75	-0,06	7,50	2,95	6,26	**6,16**
Children	1,21	-1,51	**0,60**	4,44	-3,66	-0,42
Partner	5,75	-0,75	**5,40**	9,65	-0,53	5,26
Married	9,12	-1,34	12,20	7,05	0,65	10,21
Education	12,90	0,99	32,80	5,31	**14,90**	31,00

16. In table 2 we have emphasised (in bold) the results which have a covariance closest to the actual covariance. In 3 of the 7 cases, the predictive model-based method performs best. The Hot-Deck method performs well on the variables used for grouping, but it performs poorer than the predictive model based method on all other variables.

VI. CONCLUSION

17. In both tables the Predictive Model-Based Method performs very well. It would have been interesting if it had been possible to combine Propensity Score Method and Predictive Model-Based Method when choosing donors.