TNO report
PG/VGZ/99.045

Flexible multiple imputation by chained equations of the AVO-95 Survey

TNO Prevention and Health

Public Health
Wassenaarseweg 56
P.O.Box 2215
2301 CE Leiden
The Netherlands

Tel + 31 71 518 18 18
Fax + 31 71 518 19 20

Date

Oktober 1999

Author(s)

Karin Oudshoorn
Stef van Buuren
Jan van Rijckevorsel

TNO Prevention and Health contributes to the improvement of quality of life and to an increased healthy human life expectancy. Research and consultancy activities aim at improving health and health care in all stages of life.

Netherlands Organization for Applied Scientific Research (TNO)

Author

K. Oudshoorn

Project number

40657

ISBN-number

90-6743-615-1

This report can be ordered from TNO-PG by transferring *f* 21,= (incl. VAT) to account number 99.889 of TNO-PG Leiden. Please state TNO-PG publication number PG/VGZ/99.045

ERRATUM

TNO Rapport PG/VGZ/99.045

Table 5, page 22 should be replaced with

Method	β	se	Increase in se
List-wise deletion	1.056	0.0179	-
Hot-deck imputation	0.824	0.0184	2.8 %
Multiple Imputation	1.007	0.021	17.3 %

Executive summary

In this report we consider multiple imputation of missing data in the National Services and Amenities Utilization Survey (AVO-95), conducted by the Dutch Social and Cultural Planning Office in 1995. The imputation is done by a new approach, where for each incomplete variable a separate imputation model is used. The focus is on the process of deriving the imputation models for mortgage, the current selling price of the house, and some predictor variables with missing values as well. This appears to be a critical and complex step in Multiple Imputation. Next, we describe the methodology used and practical issues encountered in obtaining the imputations. Furthermore, we give some characteristics of the obtained imputed values. Finally, the added value of the multiple imputed dataset is compared with the listwise deletion and the hot-deck imputed dataset as published now under responsibility of the Social and Cultural Planning Office. The main conclusion is that flexible multiple imputation by chained equations is an extremely suitable algorithm to obtain completed versions of incomplete large national public use datasets.

Keywords and phrases: Multivariate Multiple Imputation, Gibbs Sampler, Missing at Random, Item Nonresponse, Survey.

Contents

1 Introduction

Missing data is a returning problem in large national public use datasets. In recent years multiple imputation has been introduced as a useful, consistent and, straightforward solution. With multiple imputation the responsibility of correctly dealing with the missing items, in order to avoid biased estimates and overestimation of the precision is laid in the hands of the data collectors. One of the main attractive reasons to use multiple imputation is the fact that it results in completed data and therefore it allows for standard statistical complete-data techniques afterwards.

The idea of (multiple) imputation is to draw several times from the predictive distribution of the missing values (cf. Rubin (1987) or Schafer (1997)). From these draws several completed datasets are made by replacing the missing values with the imputed values. By applying the standard statistical techniques to the completed datasets and combining the results afterwards by pooling, both the uncertainty due to missing data and the variability in the complete data itself are taken into account. These types of variability are called within imputation variance and between imputation variance. Imputation methods are based on the assumption that the data is missing at random (MAR), that is, the response mechanism does not depend on unobserved information or on the variable(s) with missing values itself.

To impute the AVO-95[1] survey we use a new approach of multiple imputation where for each incomplete variable a separate imputation model with a set of predictor variables is used. Our imputation model consists of a set of predictor variables, the so-called donors, and a statistical model which characterizes the relationship between the imputed variable and its donors. One can think of a multiple regression model (based on a multivariate normal distribution), a logistic regression model or a multinomial logit regression model. The algorithm, which is based on Gibbs sampling, imputes each incomplete column of the dataset in an iterative fashion, variable by variable. Donors may have missing values themselves which are imputed based on a particular imputation model for this donor. The algorithm has been investigated by Brand (1999). Van Buuren & Oudshoorn (1999) describe MICE, an implementation of this algorithm in the statistical package S-PLUS. MICE stands for Multivariate Imputation of Chained Equations. This implementation is programmed in such a way that it is very easy to add imputation models not yet included. At the moment linear regression, predictive mean matching, nearest neighbor imputation, logistic regression, multinomial logit regression and discriminant analysis are included (cf. van Buuren and Oudshoorn (1999)). A consequence and an advantage of the variable by variable approach is that for each variable, a different imputation model can be used and different assumptions can be made. Therefore a set of variables can have mixed measurements levels such as both continuous as well as categorical measurement levels.

The selection of the imputation models is crucial for the quality of the imputations. Nowadays algorithms exists where different models for the data can be chosen as special models for multivariate continuous or categorical data (Schafer (1997)). However, up to our knowledge, such algorithms use the same set of donors for each impute variable. This means that, due to memory

[1] This dataset was obtained through Steinmetz Archives, Amsterdam.

problems or ill conditioning problems one has to make compromises in the size of the set of donors. This disadvantage is less present in the variable by variable approach.

In this report we discuss multiple imputation of the Dutch National Services and Amenities Utilization Survey (in Dutch: Aanvullend Voorzieningen Onderzoek 1995, abbreviated by AVO-95). The Dutch Social and Planning Office conducted this survey in 1995 (Spit (1996)).

The National Services and Amenities Survey includes 6421 households and contains information about housing, education and services (cf. Spit (1996)). The aim of the survey was to get insight in the use of a large number of social and cultural services by the Dutch population. It gives information, among other things, about how households are composed, their income positions, the type of houses they live in and how these houses are financed.

The non-response on some items is substantial. For example, from 49.9 % of the 2792 households with a mortgage, the monthly paid mortgage interest is missing. From 3427 of the households, owning a house, 11.1 % did not report the current selling price of the house.

We focus in this report on deriving imputation models and the imputed values for the yearly payment of interest and the current selling price of the house. The derived models contain variables like yearly installment of mortgage, type of house and mortgage, and others. In Section 3 is described how the imputation models are derived. Since some predictors of the yearly payment of interest and of the current value of the house also have missing items, these predictors have to be imputed as well. We deal with this problem in Section 4. In Section 2 the algorithm itself is explained. Section 5 discusses the calculation of the completed datasets. And finally, in Section 6, the added value of the multiple imputed dataset is compared with the listwise deletion and a hot-deck imputed dataset.

2 The algorithm explained

Below we briefly describe the multiple imputation algorithm used here. For a detailed discussion we refer to Brand (1999) and for the implementation in S-PLUS to van Buuren & Oudshoorn (1999).

Denote by Y the impute variable. In our case this is either the yearly mortgage interest or the current selling price of the house. Let Y_{obs} be the complete part of Y and Y_{mis} the incomplete part. Let X be the matrix consisting of predictor variables for Y. X itself may be partly observed as well. When X is completely observed the posterior predictive density of Y_{mis} can be written as

$$P(X_{mis} \mid Y_{obs}, X) = \int_{\Theta} P(Y_{mis} \mid Y_{obs}, X, \theta) P(\theta \mid X, Y_{obs})\, d\theta, \qquad (2.1)$$

with $P(Y_{mis} \mid Y_{obs}, X, \theta)$ the predictive distribution of the missing data given θ. It will be assumed that the complete data distribution of Y is parameterized by parameter $\theta \in \Theta$. Multiple imputation of Y with completely observed X consists of the following steps (Rubin (1987)):

1) (Estimation task) Estimate θ, based on the observed data (Y_{obs}, X).

2) (Imputation task)

   a) Draw a value θ^* from $P(\theta \mid X, Y_{obs})$.

   b) Draw a value Y^*_{mis} from the conditional distribution of Y^*_{mis} given $\theta = \theta^*$, thus from $P(Y_{mis} \mid X, Y_{obs}, \theta = \theta^*)$.

3) Repeat these steps for more imputations.

Concrete algorithms for normally distributed Y and categorical Y can be found in Rubin (1987). Rubin also gives (rather technical) conditions under which a multiple-imputation procedure will yield valid statistical inferences (for the frequentists) without reference to any specific parametric model. Such an imputation method is said to be proper. In Schafer (1997) the emphasis is laid on the concept of proper multiple imputations in the Bayesian sense, which means that all the imputations are independent realizations of the posterior predictive distribution of the missing data under some compete-data model and prior distribution. In contrary with Rubin's definition of proper, assumptions for Bayesianly properness invokes no condition on the response mechanism, but this is due to the extra condition of ignorability (i.e. MAR plus distinctness of the data model parameters and the response model parameters) , that is assumed throughout the book of Schafer.

In practice, however, non-response is not limited to impute variables. Some columns of data matrix X may have missing cases also. For these columns a set of predictors is selected and added to the matrix X. This process is repeated until for each column of X a model is determined and the variables that are not already present are added to X. This process terminates always since in the worst case we end up with the total dataset. On the other hand it is assumed that modelling of Y is more important than modelling the covariates of Y. Therefore in practice less effort is put into forming models for predictors of a predictor of Y.

Changing notation somewhat, we denote from now on with $Y = (Y_1,\ldots,Y_k)$, the matrix with columns of X that have missing values, and, as previous denoted, the main impute variable Y included. With X we denote the remaining columns of X, having no missing values. If all variables of the dataset have missing values then matrix X will be empty. The crucial point is now that for each component of Y, conditional on the values of all other components of Y, and the complete covariates X, we have a univariate situation. This idea is exploited in the Gibbs sampling algorithm. We just work through the matrix Y with the Gibbs sampling algorithm, by drawing, for each component, a next iteration of the posterior distribution of Y_i given all other components:

$$\theta_1^{(t)} \sim P(\theta_1 \mid Y_{obs,1}, X, Y_2^{(t-1)},\ldots, Y_k^{(t-1)})$$

$$Y_{mis,1}^{(t)} \sim P(Y_{mis,1} \mid Y_{obs,1}, Y_2^{(t-1)},\ldots, Y_k^{(t-1)}, X, \theta_1^{(t)}, \theta_2^{(t-1)},\ldots, \theta_k^{(t-1)})$$

$$\theta_2^{(t)} \sim P(\theta_2 \mid Y_{obs,2}, X, Y_1^{(t)}, Y_3^{(t-1)}\ldots, Y_k^{(t-1)})$$

$$Y_{mis,2}^{(t)} \sim P(Y_{mis,2} \mid Y_1^{(t)}, Y_{obs,2}, Y_3^{(t-1)},\ldots, Y_k^{(t-1)}, X, \theta_1^{(t)}, \theta_2^{(t)}, \theta_3^{(t-1)}\ldots, \theta_k^{(t-1)})$$

$$\ldots$$

$$\theta_k^{(t)} \sim P(\theta_k \mid Y_{obs,k}, X, Y_1^{(t)},\ldots, Y_{k-1}^{(t)}, Y_k^{(t-1)})$$

$$Y_{mis,k}^{(t)} \sim P(Y_{mis,k} \mid Y_1^{(t)},\ldots, Y_{k-1}^{(t)}\, Y_{obs,k},, X, \theta_1^{(t)},\ldots, \theta_k^{(t)})$$

Brand (1999) validates the algorithm by simulations, based on complete datasets which are artificially made incomplete by a MAR mechanism and subsequently completed by imputation. His results confirm that 1) the variables with missing data are adequately recovered, 2) the relations between imputed variables and predictors are adequately recovered and 3) the extra uncertainty due to missing data is correctly reflected. These are properties of imputations that are also established by any proper imputation method.

We mentioned before that the first step, the selection of the imputation model is very important. It turns out that obtaining the imputation model costs a lot of effort and time. The imputation model consists of major choices: the donor variables and the statistical model itself. The donor or predictor variables are chosen according to the following strategy (cf. Brand (1999)):

1) Select all variables that are relevant in the complete data model.

2) Select in addition the variables that are related to (in terms of e.g. explained variance or correlation coefficients) the impute variable(s).

3) Include in addition all variables that are related to the response model.

4) Remove from the list all variables with too many missing values within the subgroup of incomplete cases of the impute variables or variables for which the number of cases within the subgroup of complete cases of the impute variables is small.

These steps must be sufficient to obtain a list of donors for the impute variable(s). Of course, for donors with missing data, one has to repeat this procedure again. After these steps, in particular with large datasets containing many variables, the list of donors may still be quit long. In stead of selecting predictors one by one, based on e.g. a Pearson correlation coefficient one can shorten the list by using multivariate techniques such as multiple regression with several predictors simultaneously in step two.

3 Derivation of the imputation models

The National Services and Amenities Survey, AVO-95, conducted in 1995 in the Netherlands, was held originally under 9232 households. The survey response was 68.8 %, resulting in 6421 households in the dataset. These 6421 households consist of, in total, 14489 persons who where all interviewed, depending on the age, using the adult or youth questionnaire. In each household the head of the household or his/her partner was asked to fill in the household questionnaire.

Lifestyle was one of the topics of the household questionnaire. Of the households in the sample, 53.4 % (3427 households) live in a house of their own. The item "Present market value of the house " (in the survey called *V536*) is missing in 11.1 % (382 households) of the cases. From these 382 households, 69 (2.0 %) did report not knowing the value and 313 (9.1 %) did not want to answer the question. The observed present market values of the house range from 16000 (Dutch guilders) to "9998000 or more". It is common to transform such amounts with a logarithm, since this tends to linearize the relationship with the predictors. From now on we focus therefore on the variable *logV536* the logarithm of the present market value of houses.

From the 3427 households who own a house, 81.5 % (2792 households) reported to have a mortgage on the house. One of the variables concerning the mortgage on the house is *V768*, the payment of interest. The period of the payment of interest by the respondents (V766 in the questionnaire) varies between 'yearly', 'half yearly ', 'quarterly' and 'monthly'. The payment of interest, *V768*, is missing for 1387 of the households, and *V766* is missing for 6 households within the group of not missing *V768*. This means that the yearly payment of interest is missing for 1393 households in total, which is 49.9 %. Below we continue with the variable *logV768j*, the logarithm of the yearly payment of interest. For households with a mortgage, *logV536* is missing for 228 cases (8.2 %).

We focus on multiple imputations for *logV536* and *logV768j*. The datamatrix of households, with a house of their own, can be split into two parts: households with and households without a mortgage on their house. Since the goal of this report is to describe how the imputation process works for a large dataset and this action is similar for the two parts of the datamatrix, we restrict ourselves to the cases with a mortgage on the house. We follow the strategy as described in the previous section to get a set of donors for the impute variables, *logV536* and *logV768j*. For predictor variables with missing values we specify also an imputation model that is described in the next paragraph.

The first step consists of including all variables in the set of donors that will be used later on in complete data analyses of *logV536* and *logV768j*. This means that possibly existing relations between predictor variables and impute variables are kept. The AVO-95 dataset is public and is or will be used by many people. This yields that it is beforehand not known what research questions, with respect to *logV536* and *logV768j*, will be investigated with the completed datasets. We therefore include all variables that are assumed important for *logV536* and *logV768j*, based on consultation with SCP[2]. These donor variables and some characteristics of them related to the impute variables, are given in Table 1 for *logV768j* and in Table 2 for *logV536*. In the last column of these tables it is indicated at which step the predictor is added to the set of donors. The term *%Usables* stands for the percentage of cases in the datamatrix that have a missing value for the impute variable but not for the donor variable. Usable cases of predictors are needed to predict accurately the missing cases of the impute variable. *%Relevants* is the percentage of cases where both the impute variable and the donor is observed. A sufficiently large number of relevant cases of predictor variables are needed to fit a model that describes the relationship between the predictors and the impute variable. The variable *R_logV768j* and *R_logV536* are respectively the response indicators of *logV768j* and *logV536*. The response indicator is one for cases where the imputed variable is observed and zero otherwise.

Table 1: Donors for logV768j. For continuous donors the correlation with logV768j and R_logV768j (the response indicator of logV768j) is given. For categorical variables the relationship is measured with ANOVA (for logV768j) or with a Chi-square test (for R_logV768j).

Donors for *LogV768j*	Description	Correlation and p-value with *logV768j*	Correlation and p-value with *R_logV768j*	% Usables	% Relevants	Added in Step	Removed in Step
LogV536	Current value house[3]	0.363	- 0.004	86.7 %	96.9 %	1	-
V006	Age head households	-0.254	0.080	100.0 %	100.0 %	1	-
V529	Year house is built[4]	p[5] < 0.0001	p[6] = 0.065	100.0 %	100.0 %	1	-
LogV767j	Yearly mortgage pay off[7]	0.851	- 0.015	95.3 %	55.7 %	1	-
V715	Type of mortgage[8]	p[9] < 0.0001	p[10] = 0.0005	88.9 %	97.5 %	2	-
V716	Mortgage interest	0.219	- 0.037	74.7 %	95.4 %	2	-
LogV536	Current value house	0.363	- 0.004	86.7 %	96.9 %	2	-
LogV717j	Unpaid amount of mortgage	0.858	- 0.051	48.0 %	90.9 %	2	-
V40504	Other income?	p[11] =0.00134	p[12] = 0.62	100.0 %	100.0 %	2	-
V765	Remaining years of mortgage	0.349	- 0.057	84.9 %	92.2 %	2	-
V515	Respondent head household?	p[13] = 0.749	p[14] < 0.0001	100.0 %	100.0 %	3	-

[2] We thank J. Spit, M. Ras and I. Stoop from SCP, The Netherlands, for the coorperation in the definition of the project.

[3] Log transform is taken,

[4] Seven categories: before 1930; 1930-1944; 1945-1969; 1970-1979;1980-1989, 1990 + and missing,

[5] F-value = 21.2 with df = (5,1388),

[6] $X^2 = 10.4$, df = 5,

[7] *ln767j* is the yearly installment of the mortgage in contrary with *logV768j* that is solely the yearly payment of interest,

[8] The number of categories are reduced to four, to avoid nearly empty cells,

[9] F-value = 61.2, df = (4,1359),

[10] $X^2 = 19.8$, df = 4,

[11] F-value = 10.3, df = (1,1397),

[12] $X^2 = 0.25$, df = 1,

[13] F-value = 0.10, df = (1,1397),

[14] $X^2 = 51.9$, df = 1.

Table 2: Donors for logV536 (with mortgage). For continuous donors the correlation with logV536 and R_logV536 (the response indicator of logV536) is given. For categorical variables the relationship is measured with ANOVA (for logV536) or with a Chi-square test (for R_logV536).

Donors for*logV536*	Description	Correlation and p-values with *logV536*	Correlation and p-values with *R_logV536*	% Usables	% Relevants	Added in Step	Re-moved in Step
V530	Type of house[15]	p[16] < 0.0001	p[17] < 0.001	100.0 %	100.0 %	1	-
AS03n	House in Randstad yes/no	p[18] = 0.0030	p[19] = 0.075	100.0 %	100.0 %	1	-
AS07	Net Income	p[20] < 0.0001	p[21] = 0.0003	57.5 %	91.0 %	1, 3	-
AS08	Type of household[22]	p[23] < 0.0001	p[24] = 0.4759	100.0 %	100.0 %	1	-
LogV767j	Yearly mortgage pay off	0.341	0.029	33.8 %	79.3 %	1	-
LogV768j	Yearly mortgage interest	0.363	0.012	18.9 %	52.9 %	1	4
logV717j	Unpaid amount of mortgage	0.376	0.018	26.8 %	73.3 %	2	-
V529	Year house is built[25]	p[26] < 0.0001	p[27] = ????	100.0 %	100.0 %	2	-
V110	Religious affiliation[28]	p[29] < 0.0001	p[30] = 0.1009	100.0 %	100.0 %	2	-
V40504	Other income?	p[31] < 0.0001	p[32] = 0.1339	100.0 %	100.0 %	2, 3	-
V776	Number of cars[33]	p[34] < 0.0001	p[35] = 0.4080	100.0 %	100.0 %	2	-
V271	Own business[36]	p[37] < 0.0001	p[38] < 0.0001	100.0 %	100.0 %	2, 3	-
V751	Number of rooms[39]	0.399	-0.016	100.0 %	100.0 %	2	-
V752	Area living room[40]	p[41] < 0.0001	p[42] = 0.0789	97.8 %	99.5 %	2	-
V33603	Visit Ballet performance?[43]	p[44] < 0.0001	p[45] = 0.9173	100.0 %	100.0 %	2	-

The second step includes all variables related to the impute variables. When the imputation models are too large, the algorithm will be unstable due to ill-conditioning. To avoid this we first look for all variables with reasonable high correlation with the (observed part of the) impute

[15] The number of categories (including the category unknown or missing) is reduced to four, to avoid almost empty cells,

[16] F-value = 246.9, df = (3,2560),

[17] X^2 = 46.6, df = 3,

[18] F-value = 8.8, df = (2562,1),

[19] X^2 = 3.2, df = 2,

[20] F-value = 126.4, df = (2328,4),

[21] X^2 = 20.8, df = 4,

[22] Some categories are joint to avoid nearly empty cells,

[23] F-value = 47.2, df = (2560,3),

[24] X^2 = 2.5, df = 3,

[25] The missings are recoded to the category "unknown",

[26] F-value = 19.3, df = (6,2557)

[27] X^2 = 48.0, df = 6 *** expected values smaller than 5,

[28] 21 cases have missing values, the category "other groups" are joint with the missings (as unknown's),

[29] F-value = 9.2, df = (4,2559)

[30] X^2= 7.8, df = 4,

[31] F-value = 94.0, df = (1,2562),

[32] X^2 = 2.25, df = 1,

[33] 4 cases were missing, the mean was imputed (i.e. one car).

[34] F-value = 94.06, df = (2,2561),

[35] X^2 = 1.8, df = 2

[36] V271 was missing for 85 cases, 66 with known logV536; these are coded unknown,

[37] F-value = 85.6, df = (2,2561),

[38] X^2 = 32.1, df = 2

[39] 8 cases have missing entries, these are imputed with mean imputation,

[40] Some categories are joint and the missings are imputed with the modus category,

[41] F-value = 90.1, df= (7,2556)

[42] X^2 = 12.7, df = 7

[43] 103 cases have missing values; these are recoded to "unknown".

[44] F-value = 8.1, df= (3,2560)

[45] X^2 = 0.5, df = 3

variables. Correlations are calculated in terms of the Pearson correlation coefficient for continuous predictors. For categorical predictors we used ANOVA to analyse the dependency between the predictor and the impute variable. Next we check with multiple regression which variables contributed significantly to the multiple correlation coefficient. It turns out that *V715* (type of mortgage), *V716* (interest rate), *logV717j* (unpaid amount of mortgage), $(logV717j)^2$, $(logV767j)^2$ and *V765* (remaining years of mortgage) are important predictors for *logV768j*. In Table 1 some characteristics of these variables are given. One can see that these variables have missing entries as well. For *logV536* (with mortgage) the variables *logV717*, *V529* (year house is built), *V110* (religious affiliation), *V40504* (other income), *V776* (number of cars), *AS03n* (living in Randstad?), *V271* (own business), *V751* (number of rooms), *V752* (area living room) and *V33603* (do you visit ballet performances?). Characteristics of these variables are displayed in Table 2.

Next, the third step adds all variables that are related to the response mechanism of the impute variable to the set of donors. A few variables correlate significantly with the response indicator of *logV768j*, but except for one, the correlation coefficients are less than 0.1. Variable V515, "the relation of the person who filled in the questionnaire with respect to the head of the household", correlates with a Pearson correlation coefficient of –0.1466 (in terms of Chi-squared test: X^2=59.1, p<0.0001, df = 1). Thus people who are not head of the household have a larger tendency not to respond. It seems logically that these persons have less knowledge of the mortgage and its amounts, so this response is caused by not knowing and not due to not willing to tell in general. It is worth to note that income correlates very weakly ($\rho = 0.0458$, p-value is 0.023) with the response indicator of *logV768j*. The response indicator of *logV536* is related to *AS07*, *V40504* and *V271* (See table 2).

Step four removes all predictor variables with too many missing values from the donor list. In practice it is wise to exclude those variables with less than 30 % usable cases or less than 50 % of relevant cases. This means that *logV768j*, though significantly correlated with *logV536* is removed from the donor list.

Through step one to four we have built now for both impute variables, *logV536* and *logV768j*, imputation models. The model of *logV536* consists of, added in step one: *V530*, *AS03n*, *AS07*, *AS08*, *logV767j* and, added in step two, *logV717*, $logV717^2$, *V539*, *V110*, *V40504*, *V776*, *V271*, *V751*, *V752* and *V33603*. The donors in the model of *logV768j* are, added in step one: *logV536*, *V006*, *V529*, *logV767j*, added in step two: *V715*, *V716*, *logV536*, *logV717*, *V40504*, *V765* and added in step three: *V515*.

4 Imputation models for 2nd level impute variables

Some predictor variables for *logV536* and *logV768j* have missing entries as well. These variables are called *2nd level impute variables* since they have to be imputed due to the fact that they are incorporated in the model for the main impute variables, the first level impute variables. In our case, it turns out that we only have to impute second level impute variables and we do not have to go a level beyond. In other words, the impute models for the second level impute variables do not contain any variables that have missing values themselves and are not yet donor for the first level impute variable.

For donor variables of the first level impute variables were the nonresponse was minimal we use mean imputation for imputing the missing cases. Knowing that mean imputation is, in general, a bad choice as imputation technique, due to underestimation of the variance (Little and Rubin (1987)), this effect is negligible when the nonresponse is very small. In our case we used only mean imputation when the nonresponse is less than 15 cases, that is in 0.54% of the cases, yielding the variables *V751* and *V776*. The missing cases of the categorical variable *V752* were imputed with the modus category.

There are categorical variables in the set of donors of the first level impute variables *logV768j* and *lnV536* for which the nonresponse is between 1.0 % and 5.0 %, namely *V529*, *V530*, *V110*, *V271* and *V33603*. For these variables it is assumed that building a separate model is not important in terms of the effect on the imputed values for the first level impute variables. For these variables we defined the cases with missings as a separate category. Furthermore, in order to avoid ill-conditioning problems when calculating the imputed values, we reduced the number of categories, when needed, by joining categories of a categorical variable that have similar values for the impute variables.

For other variables with a reasonable amount of missing cases (this concerns the variables with the number of *usables* and *relevants* less than 100.0 % in Table 1 and 2), step 1 to 4 were followed as well. From Table 3 one can extract the imputation models for these variables, that is for *V715*, *V716*, *logV717*, *V765*, *logV767j*, and *AS07*.

To conclude the description of the derivation of the imputation models one can say that the process of obtaining the imputation models is very labour intensive. For every variable with missing entries one has to look for donors that are highly correlated with the (one or two level impute variables). Variables with (almost) empty cells have to be redefined and checked again in relation with the impute variables. Until now the process of obtaining the imputation models is not yet automized in the `mice` algorithm. Although it hasn't been proved yet we think that generally going beyond second level impute variables only brings an enormous amount of extra work that isn't paid back by increased quality of the impute variables.

5 Calculation of the impute values by `mice`

The imputed datasets were generated with the S-PLUS routine `mice`. For a detailed description we refer to van Buuren and Oudshoorn (1999). In the Appendix the syntax is given, to show that, after the imputation models are obtained, calculating the imputed values is easy.

In Table 3 all used imputation models are given. This table reads as follows: the variable name of a column is the impute variable. The zeros and ones below the impute variable indicate which donors are used for this impute variable. A one stands for inclusion in the donor set and zero otherwise. For example, in the column with name *logV536* one can read the imputation model for *logV717* consisting of variables *logV536, logV768j, V006, V715, V716, logV767j,* $(logV767j)^2$, *V765*, *V776* and *V530*. Since $(logV767j)^2$ is a function of *logV767* it will be (passively) imputed with the corresponding imputed values of *logV767*. The same holds true for $(logV717)^2$.

We used linear regression as univariate imputation method for *logV536*, *logV768j*, *V716*, *logV717*, *logV767j*, *V765* and *V41702*. For variables *V715* and *AS07* multinomial logit regression was used.

Table 3: The imputation models for the first and second level impute variables. Entry (i, j) *equals one if the variable of row* i *is a predictor variable for column* j.

	LogV536	*logV768j*	*V715*	*V716*	*logV717*	*(logV717)*^{2}	*logV767j*	*(logV767j)*^{2}	*V765*	*AS07*	*V41702*
LogV536		1	1	1	1	0	0	0	1	1	1
LogV768j	0	0	1	1	1	0	1	0	1	1	1
V006	0	1	1	0	1	0	1	0	1	1	1
V715	0	1	0	1	1	0	1	0	1	1	1
V716	0	1	1	0	1	0	1	0	0	0	0
logV717	1	1	1	1	0	1	1	0	0	1	1
logV717^2	1	1	0	0	0	0	1	0	0	0	0
logV767j	0	1	1	0	1	0	0	1	1	0	0
logV767j^2	0	1	0	0	1	0	0	0	0	0	0
V534	0	0	0	1	0	0	0	0	1	0	0
V529	1	0	1	0	0	0	0	0	0	0	0
V765	0	1	0	0	1	0	1	0	0	0	0
AS07	1	0	0	0	0	0	1	0	0	0	1
V110	1	0	0	0	0	0	0	0	0	0	0
V40504	1	1	1	0	0	0	0	0	0	1	1
V776	1	0	1	0	1	0	0	0	0	1	1
AS08	0	0	1	0	0	0	0	0	0	1	1
AS03n	1	0	1	0	0	0	0	0	0	0	0
V271	0	0	1	0	0	0	0	0	0	0	0
V751	1	0	0	0	0	0	0	0	0	0	0
V752	1	0	0	0	0	0	0	0	1	0	0
V33603	1	0	0	0	0	0	0	0	1	0	0
V530	1	0	1	1	1	0	1	0	1	0	0
AS22	0	0	0	0	0	0	0	0	0	1	1
V41702	0	0	0	1	0	0	0	0	0	1	0
V515	0	1	0	0	0	0	0	0	0	0	0

6 Quality of the imputed values

Only for artificially made datasets with missing data where the real values of the missing entries are known, one can really check by statistical methods whether the imputed values are, with large probability, drawn from the correct posterior predictive distribution of Y_{mis}. In real life of course the fact that cases are missing is just the very problem and by definition the values of these cases are not known. Therefore we can only hope that by choosing the right imputation model we get a situation that, given the value of the donors, the response mechanism is MAR. Under the assumption of MAR it has been proved, for various multivariate models, that multiple imputation yields drawing from the correct posterior predictive distribution (Rubin (1987)) and pooling the results of inferences of the completed datasets yields valid results.

To get, at least an idea whether the imputed values are in line with the data we checked some characteristics of the imputed values. We compare the imputed values obtained by multiple imputation with the dataset based on list-wise deletion and with the hot-deck imputed dataset of AVO-95 as distributed by the Steinmetz Archives. From Table 4 one can conclude that there are no peculiar values found for the mean, minimum and maximum of *logV536* and *logV768j* using the different completed datasets. The largest deviation is found for the minimum of *logV768j*. This is not surprising since there are only few values of *logV767j* in the observed part below 5.0.

Table 4 The mean, maximum and minimum for logV768j *and* logV536; *In the case of Multiple Imputation and Hot-deck imputation only the imputed values are used.*

	logV768j			*logV536*		
	Listwise deletion	Multiple Imputation	Hot-deck imputation	Listwise deletion	Multiple Imputation	Hot-deck imputation
Mean	8.75	8.79	8.81	12.37	12.45	12.54
Maximum	12.34	12.40	12.48	15.76	14.13	15.76
Minimum	2.48	5.15	2.48	10.31	11.27	11.29

From Figure 1 and 2 one can see that the distribution of the imputed values of *logV768j* and *logV536* is comparable to the distribution of observed part of these variables.

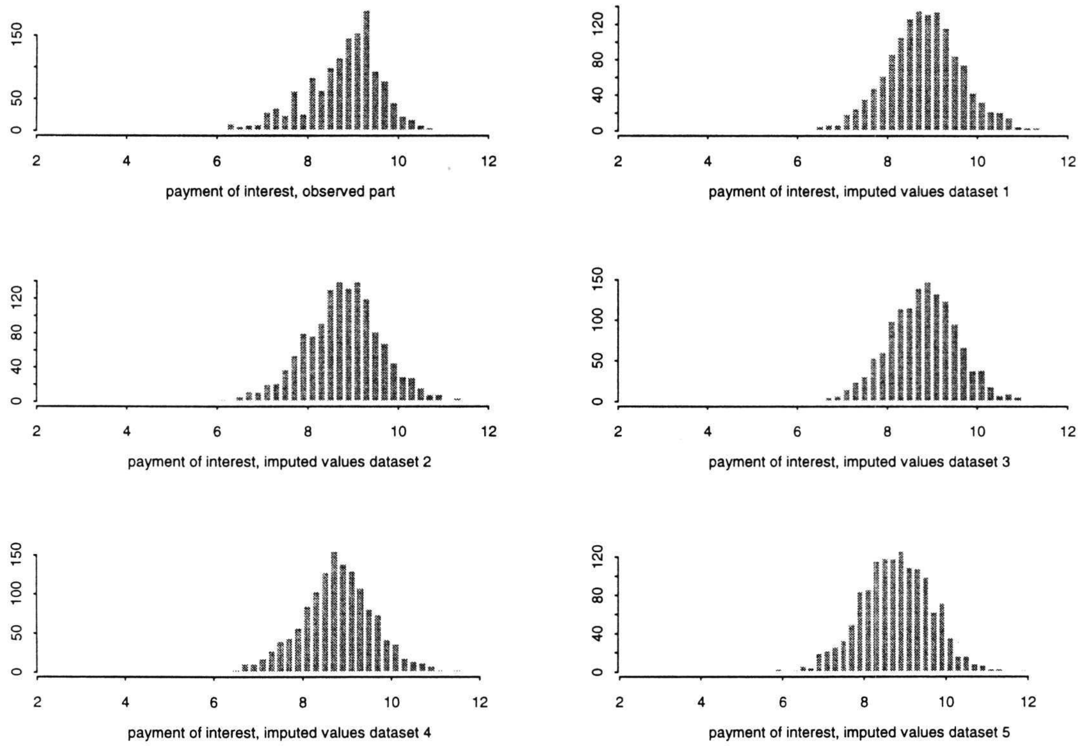

Figure 2: Distribution of lnV768j, *the observed part and the imputed values of the five completed datasets*

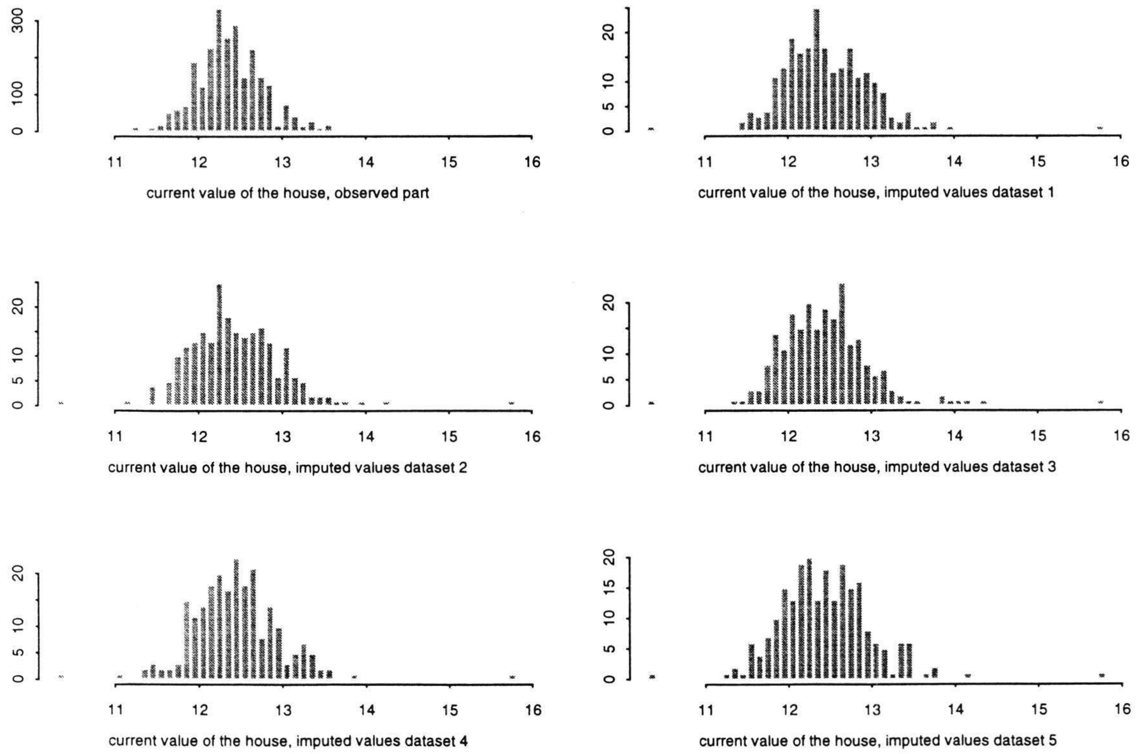

Figure 1: Distribution of lnV536, *the observed part and the imputed values of the five completed datasets*

Variables *logV768j* and *logV767j* are highly correlated ($\rho = 0.85$) in the completely observed part of the dataset. This correlation is 0.81 for the imputed values, so there is only a tiny reduction of the correlation coefficient. The correlation of the hot-deck imputed values of *logV768j* and *logV767j* equals 0.50, which is a reduction of 59 %. This fact becomes even clearer from Figure 3, the scatterplot of *logV768j* against *logV767j* for the three different methods. In case of multiple imputation the first imputed dataset is displayed.

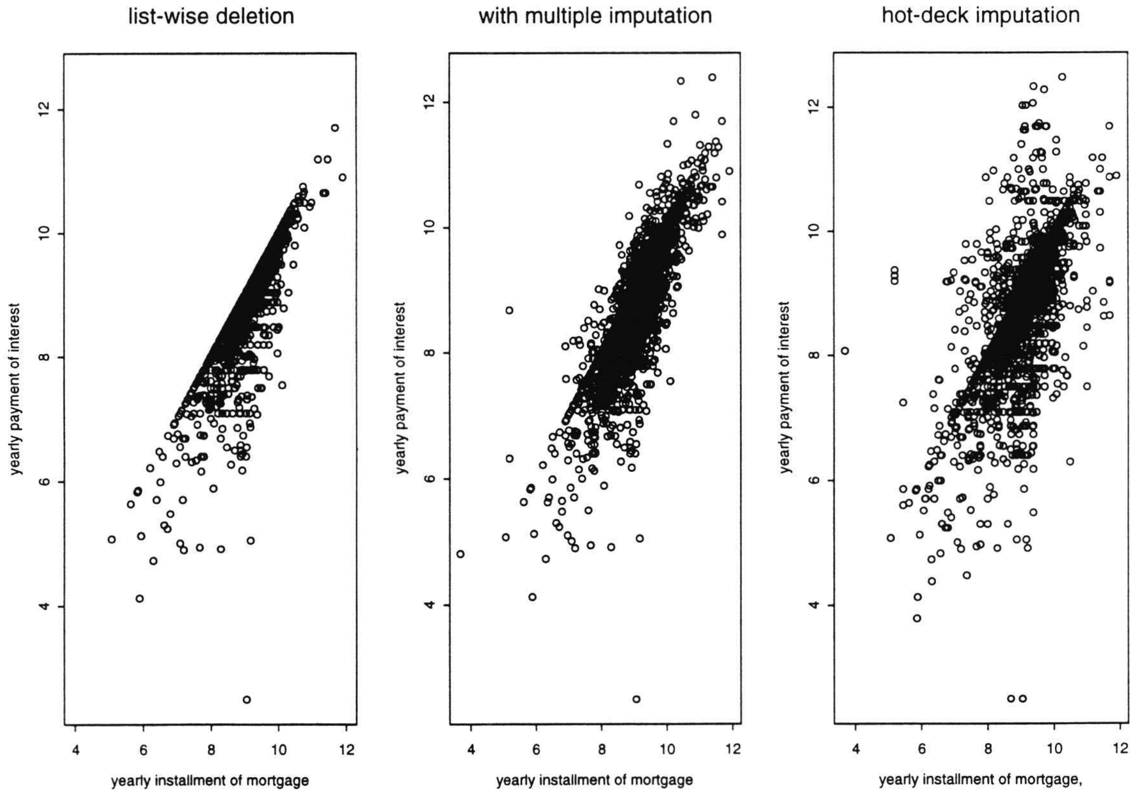

Figure 3: Yearly installement of the mortgage against yearly payment of interest.

By definition the value of the yearly installment of the mortgage should be larger or equal to the yearly payment of interest, as is the case for the observed values. This restriction is violated both in the multiple imputed datasets and in the hot-deck imputed dataset. But from Figure 3 we conclude that this restriction is far worse exceeded in the hot-deck imputed dataset. Note however that side conditions like this can easily be implemented in MICE. For reasons of comparison this is not applied here.

Besides the fact that the strong correlation between *logV768j* and *logV767j* is reduced in the hot-decked imputed dataset, standard errors are underestimated. This is a well-known fact for single imputed datasets. When using multiple imputation, in stead of single imputation, standard errors will not be underestimated. The point is that the variation caused by the missingness of the data is correctly taken into account when multiple imputation is used. For example, consider the model

$log V768j = \alpha + \beta log V767j$. In Table 5 the estimates of β can be found. Observe that the increase in standard error for the hot-deck imputed dataset, compared to the list-wise deleted dataset, is only 2.8 % whereas for the multiple inputed dataset the increase is 17.3 %. This does not mean that the hot-deck outcome is better because the standard error of the regression estimate is smaller. To the contrary, the standard error of the regression estimate is underestimated considerably.

Table 5: Regression estimates of β obtained with the different imputation methods.

Method	β	se	Increase in se
List-wise deletion	1.056	1.079	-
Hot-deck imputation	0.824	1.084	2.8 %
Multiple Imputation	1.007	0.021	17.3 %

7 Conclusion

From this case study we conclude that handling missing data of the AVO-95 survey with multiple imputation by chained equations works out better than any other customary method like list-wise deletion or hot-deck imputation. The imputed datasets are, according to certain characteristics, in line with the observed data.

Due to the flexible implementation of multiple imputation by chained equations in S-PLUS there is a lot of freedom in choosing the imputation models. Most of the job of obtaining the imputed values has to be done before actually drawing the imputed values. Especially deriving the imputation models gives a lot of work. The more levels of impute variables are involved, the more effort it takes and the more complex it is to model the impute variables.

For large national public use datasets it is inevitable to use complex imputation models. These datasets are usually used for various reasons, and by numerous people. Therefore it is important to include all related variables with the impute variables in the model to ensure that existing relations are kept. This is not an easy task. And since in practice the relations that will be studied by use of the dataset are not known by the imputer of the data the more important donors are included the larger the probability that no important relations are destroyed. Schafer (Schafer (1997), page 383-384) states that "... *Models for multivariate data from complex surveys can (and undoubtedly should) be quite complex, and more work needs to be done to formulate flexible models and algorithms applicable to a variety of survey datasets*". Based on this case study of multiple imputation with MICE of the AVO-95, which is a rather complex survey, we conclude that MICE is an algorithm that not only suits the demands of Schafer but is flexible as well. It is flexible because one can incorporate different models and the algorithm is applicable to a variety of survey datasets.

References

BRAND JPL. Development , implementation and evaluation of multiple imputation strategies for the statistical analysis of incomplete datasets. Academic thesis, Erasmus University Rotterdam, 1999.

BUUREN S van, BOSHUIZEN HC, KNOOK DL. Multiple imputation of missing blood pressure covariates in survival analysis. Statistics in Medicine 1999;18:681-94.

BUUREN S van, OUDSHOORN K. Flexible multivariate imputation by MICE, Contributed paper, International Conference on Survey Nonresponse, Oct. 1999, Portland.

RUBIN RB. Multiple Imputation for nonresponse in surveys. New York: Wiley, 1987.

SCHAFER J. Analysis of incomplete multivariate data. London: Chapman & Hall, 1997.

SPIT J. Onderzoeksverantwoording AVO'95 (in Dutch), SCP, 1996.

Appendix A

The command to calculate imputed values in S-Plus is as follows (van Buuren and Oudshoorn (1999)):

```
Imp<-mice(data = avo,
            index = index.avo,
            methods = methods.avo,
            functions = functions.avo,
            pred.mat = avo.pred.mat,
            iter.max = 100, seed = 65432, nimp = 5 )
```

where `avo` is the data matrix to be imputed, so this is the data matrix containing the cases with a mortgage on the house; `index`, `methods` and `functions` are vectors (see Table 6). `index` is a vector with the numbers of the columns of `avo` that are imputed. `methods` is a vector with the univariate methods used for imputation of the incomplete columns. For the `avo`-dataset we used the methods `norm` (Bayesian normal imputation, Rubin (1987), page 168) for continuous variables, `polyreg` (multinomial logit regression, see Brand (1999), page 94) for the categorical variables and `passive` for the variables that are a function of another variable. The method `passive` applied to a certain column basically means that the column is a function of another (so called mother-) column and is imputed with use of corresponding values of the mother column and the function that defines the relation between the two columns. The matrix `pred.mat` is a matrix specifying the set of donors to be used for each incomplete column by zeros and ones. The rows in the matrix correspond to target variables and a 1 for entry *(i,j)* means that the variable of column *j* is a donor for the variable in row *i*. In Table 3 a transposed part of this matrix is given, that is, the rows for the completely observed variables are omitted (the rows for these variables consist of zeros only). The output of `mice` is stored in the object `Imp`, that is of class `mids` (abbreviation of multiple imputed dataset).

Table 6: The inputvectors **`index.avo,`** **`methods.avo`** and **`functions.avo`** for **`mice`** to impute *logV536* and *logV768j*.

Variables	`Index.avo`	`methods.avo`	`functions.avo`
LogV536	1	norm / pmm	
LogV768j	2	norm / pmm	
V715	4	polyreg	
V716	5	norm / pmm	
LogV717	6	norm / pmm	
$(logV717)^2$	7	Passive	logV717^2
LogV767j	8	norm/ pmm	
$(logV767j)^2$	9	Passive	log767^2
V765	12	norm/ pmm	
AS07	13	Polyreg	
V41702	25	norm / pmm	