AUTOMATIC INTERACTION DETECTION FOR IMPUTATION – TESTS WITH THE WAID SOFTWARE PACKAGE

Pasi Piela and Seppo Laaksonen
Statistics Finland
e-mail: Firstname.Surname@stat.fi

Abstract

Some types of tree-based methods have been applied during some recent decades. In most cases, the purpose has been to explain or describe the behavior of a certain phenomenon using empirical data. An important feature of this algorithm is that a tree may be built more or less automatically after the target variables and the explanatory variables have been determined. Tree methods have not been very common in data analysis, nevertheless. A reason obviously is that it is easy to criticize this method. On the other hand, any tree is not always easy to interpret well, or, thus, this method is not always reasonable as the final analysis. It is thus more or less a technique for exploratory data analysis. In our paper, this technique is also used as a good helping tool for the final target, that is, for imputing missing values. Our first tests, done in an European research project, give fairly promising results. In this paper, we use regression tree for a metric variable, and classification tree for a categorical variable. Further research is needed.

Keywords: *Classification trees, regression trees, imputation.*

1. Introduction

An imputation process has the two main parts, (i) to construct a good imputation model, and (ii) to replace the missing or other incomplete values with imputed ones. There are a high number of optional imputation models, but some types of parametric regression models have often been preferred. Simple imputation methods such as mean and ratio imputation, for example, use such a regression model, which only consists of a constant term of the model. In this case, the model is deterministic. Respectively, we say such an imputation method to be deterministic. When adding a random term in the imputation model, stochastic imputation methods may be performed. In some cases, an imputation model may be quite implicitly described, and it is even difficult to see how deterministic or stochastic the method is.

The imputation task may be done with various techniques for the same imputation model. Laaksonen (2000) uses the following division. If the imputed value is taken directly from the model, he calls this technique as ‘model-donor’ method. At contrast, if the imputed value has been taken from an available or responded unit, the imputation method is called ‘real-donor’.

We here present some applications using both real-donor and model-donor techniques, on one hand, and deterministic and stochastic, on the other. The speciality in this paper is imputation model, which belongs to the family of tree-based methods. These methods have a rather long history, but recently, these have met a new invasion in various fields, as an approach to data mining methods. These methods have some similarities to neural nets techniques, which also have become very popular in various fields.

Although the family of tree-based methods is not new, many new developments have been done in recent research. New options within these techniques are available, some of these trying to build each particular tree as robustly as possible. In the case of imputations, the terminal nodes of the tree have been used as imputation cells, and the imputation tasks have been done within each node or the cluster of such nodes. This is a fairly easy technique and seems to be promising as the first tests in the EU project called AutImp have shown. The project in which the main partners have been The University of Southampton and Statistics Netherlands has developed the prototype software package 'WAID' (*Weighted Automatic Interaction Detection*). In this paper, we present some results based on the Finnish survey data and on the UK census data, and evaluate the advantages and disadvantages of this methodology. Furthermore, a number of comparisons are done between these and traditional techniques. This naturally raises a question about the appropriate imputation methods after construction of the trees.

The paper is organised so that in Section 2 we give the overview to the tree-methods used, and in Section 3, how imputation has been done within the imputation cells obtained from the model. In Section 4, we give empirical results from the two types of data with missing data. The first example is from the Finnish consumption survey data, in which all the variables needed to impute are metric. Hence, regression tree-methods are used. The second example exploits classification trees, since the variables being imputed are categorical, derived from the sample of the UK population census. Section 5 gives some concluding remarks.

The WAID software and AutImp reports can be downloaded from the project AutImp website: see References.

2. Construction of tree-based imputation models and imputations

We here first present the principles for building regression trees, and then these for classification trees.

Regression tree

Let $y_1, y_2, \ldots, y_n$ denote the values of the response variable Y in a node, with corresponding values $\{x_{ij}\,;\, i = 1, \ldots, n \text{ and } j = 1, \ldots, p\}$ for p categorical explanatory variables $X_1, \ldots, X_p$.

The measure of dispersion is the Weighted Total Sum of Squares (TSSW), defined by

$$\text{TSSW} = \sum_1^n w_i \left(y_i - \bar{y}_w\right)^2 ,$$

where w_i is a (node-specific) weight attached to the i^{th} case in the node, and $\bar{y}_w$ is the corresponding weighted mean of the response variable in the node.

The node chosen for splitting is the one with largest value of TSSW.

Given the weights w_i, a standard ANOVA style decomposition is then used to "pick" the explanatory variable to define the split for this node and the definition of the split in terms of the

categories of this explanatory variable and its monotone or non-monotone nature.

The actual split that is chosen for X_j is one that maximises the Weighted Between Sum of Squares

$$\text{BSSW} = \left(\sum_{i:\, x_{ij} \le c} w_i\right)\left(\frac{\sum_{i:\, x_{ij} \le c} w_i y_i}{\sum_{i:\, x_{ij} \le c} w_i} - \bar{y}_w\right)^2 + \left(\sum_{i:\, x_{ij} > c} w_i\right)\left(\frac{\sum_{i:\, x_{ij} > c} w_i y_i}{\sum_{i:\, x_{ij} > c} w_i} - \bar{y}_w\right)^2$$

where $1 \le c < d_j$ denotes the (ordered) category of X_j that determines the split. Finally, the explanatory variable (and splitting criterion) that determines the actual split applied to the node are defined as the variable that generates the largest such maximum value of BSSW. The terminal nodes can be defined by several restrictions.

Classification tree

WAID uses the standard Gini measure of within node heterogeneity

$$G = 1 - n^{-2} \sum_1^A n_a^2 ,$$

where n is the total number of cases in the node, n_a is the number of cases in the node with $Y = a$. The candidate parent node with largest value of G among the set of all such candidate parent nodes available at any stage is the one that is chosen for splitting at that stage.

The optimal split for X_j is the one that leads to the minimum sum of G values for the two resulting child nodes. For a non-monotone explanatory variable X_j, two ways of deciding an optimal split: GINI Optimal looks all possible binary splits and the other GINI creates pseudo-ordering of the categories of X_j.

3. Imputation techniques of WAID

Our imputation model is thus rather special, but the imputation tasks used are rather standard. In the current WAID, there are four methods for imputation: most common category (mode imputation), mean imputation, random selection of a donor (random hot decking) and nearest neighbor. Our results are based on tree methods so that within each terminal node, either random selection or nearest neighbor technique has been applied. This thus means that a real donor has been drawn randomly or by using nearest observation of the non-missing units within each imputation cell, and the missing value has been substituted with this observed value.

All the options of the WAID have not been attempted. For example, the imputation model and the terminal nodes, consequently, may be estimated from the different data file than that used for imputations. Naturally, the same variables with similar categories should be included in both files. The WAID software also has some mass imputation tools so that a number of variables may be imputed successively. We here, however, only present results for single variables.

WAID gives opportunity to use a different imputation technique for each imputation cell (terminal node), although currently only 4 alternatives are available. We have not tried to exploit this feature.

For comparisons, we present the true results, and also the results based only on available cases. The latter one gives opportunity to follow whether each particular method and its specification is approaching to a true value or not.

WAID gives opportunity easily to build various types of tree models. We present several examples in order to better understand, what options could be best for each imputation task.

4. Empirical findings from the two survey data

We first present some imputation test results based on the Finnish Household Expenditure Survey (HES) data from 1996, and then those based on the UK Census data.

Finnish Household Expenditure survey data (HES)

Tables 1 to 3 and Figure 2 cover our test results based on the HES 1996. The results presented are varying to some extent so that, on one hand, various auxiliary variables have been used, and, on the other hand, somewhat different parameters for WAID-Tree algorithm have been used. So, we can see how well the different assumptions perform. We have not looked in advance the real values, but since we had before these tests already made evaluative tests by other software, it is possible that our understanding has been better than that of an ordinary user. It should be noted that, in all cases, we have not needed to decide which method and its specification would be used in practice. We thus only compare the results obtained and try to look forward to the best method/specification.

DRINKS

Table 1 is concerned alcohol drinks (DRINKS) consumed by a household. When comparing the estimates of the true vales and those of available cases, we see the means and the standard deviations of the latter to be slightly higher but the differences are not very big. Thus, the non-responding households seem to drink only a bit less than the responding households. To understand better this factor, it should be noted that single households (especially men) respond much worse than the bigger ones. Thus, if we would standardize household size, in particular, the change in these estimates would be vice versa. But, this is not any analysis report, we compare simply the estimates between different imputation models, since random hot decking was applied in all WAID imputations.

One approach to look results is to check whether after a particular imputation technique the estimates are at least as good as based on ‘available cases’, and secondly, to hope that the estimates would be closer to true values even. Thus although the estimates would be after imputation at available cases level, we could be satisfied because a higher number of observations would be obtainable. In Table 1, the best estimates seem to be achieved with ordinary least square tree, the minimum number of groups being about 50 (but factually, there are smaller groups), and using six explanatory variables (see Table 1). The results are approximately the same when the group

size is about 75. It is interesting that when adding the number of explanatory variables, the results are worse, the mean being underestimated and the standard deviation overestimated. This is obviously due to some small and non-homogeneous imputation cells (terminal nodes). One test with Huber's min/max seems to provide the worst results, if we do not take into account the last-row OLS result which is an example with the only four explanatory variables. This shows that these four ones, although especially Number of Adults and Gender are fairly good explanatory variables, are not reasonable to explain differences in drinking consumption.

The number of terminal nodes has an influence on the results but not so that the results would be better while this number is increasing. This was also seen in results from the evaluative CART (*classification and regression trees*) software tests by Mesa, et all. (2000). So it seems that there is an optimum between an ideal combination of explanatory variables and the number of terminal nodes, but it is not clear how this will be definitely found.

HP5/KP5 (health)

All the following imputations have been made by using the same explanatory variables (see Table 2). Imputation results for yearly consumption of health of household (KP5) and household member (HP5) are given in Tables 2 and 3. The yearly consumption of a household is simply the sum over the consumption of its members. It seems to be easier to impute at member level than at household level, which can be seen in the imputation results too. However, there are several possibilities to choose auxiliary data for KP5. Here we have simply used values of the breadwinner of household when possible, that is age of breadwinner, sosio-economic status of breadwinner. It is clear that knowing the number of members or children in household is more informative for imputation of KP5 than it is for HP5.

The tree of approximately 80 terminal nodes gives the best results for HP5. The mean is even closer to the true mean than the mean of the available cases although non-response is quite non-ignorable. Also the estimate of standard deviation is quite good. Part of this regression tree (only 11 terminal nodes) is in Figure 2 in the end of the paper. It is interesting to see what is the first split. Namely, one node only consists of pensioners (SOSECON = 70) and second of all the others; naturally pensioners usually have higher consumption of health. Also next splits seem to be sensible.

As noted earlier the optimal tree in all of these cases is not the largest one but somewhere between medium-sized and large. Because simple random hot deck is used as a final imputation method the number of zeros with non-ignorable non-response has always been estimated very well in WAID tables.

UK Census Sample of Households data

In this example case we present results for anonymised sample of UK Census data; the data are from EU/FP5 Euredit project (*The Development and Evaluation of New Methods for Editing and Imputation*). Again there are both member level and household level data, and breadwinners of each household are chosen to impute two variables, namely CARS (*number of cars*) and ROOMSNUM (*number of rooms*). Imputation variables are thus categorical and GINI index will be used with pseudo ordering method for all non-monotone variables.

The data consist of 19179 breadwinners of the households in York and Humb area. Rate of missingness is 2 % for ROOMS and 0.9 % for CARS. The nearest neighbor imputation method is preferred here.

ROOMSNUM (number of rooms)

Values of ROOMSNUM are between 1 and 15, 15 meaning more than 14 rooms in household. Chosen explanatory variables are HHSPTYPE (*household space type*, 1-14 classes), PERSINHH (*number of persons in household*), SEGROUP (*socio-economic group*) and SOCLASS (*social class based on occupation*, 9 classes). Again it can be seen that a high number of appropriate auxiliary variables gives worse imputation results than small number of obvious explanatory variables like four variables here. Moreover, the medium size tree is not the best one, though the differences in the results between trees are small. However, due to the size of the data terminal nodes become easily very large for imputation by using only four explanatory variables. Further, because of the small rate of missingness, it makes no difference to use a full size tree of 131 terminal nodes or a large tree of 80 terminal nodes by nearest neighbor imputation method.

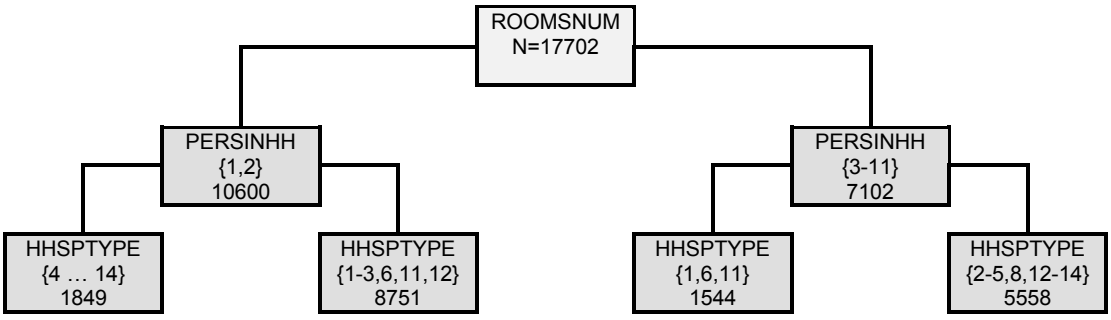
Results are quite good as it is seen in Table 4 of marginal distributions. GINI-values are relatively low in parent nodes: between 0.19 and 0.30.

As Figure 1 shows, the first split is for PERSINHH so that one node consists only 1-2 persons household and thus other node consist larger households with obviously large number of rooms too. Next splits are done for HHSTYPE in both nodes.

CARS (number of cars)

Small number of appropriate explanatory variables gives the best results here too: DISTWORK (*distance to work*), PERSINHH and SEGROUP. Large imputation trees are best. First the data are divided by SEGROUP then other part again by SEGROUP to get manual with 'junior non-manual' workers into same group and then by PERSINH while other side is divided by PERSINH after on split of SEGROUP. One person households are separated into their own node in the both splits of PERSINH. Results are presented in Table 5.

Figure 1. First nodes of the WAID GINI tree for ROOMSNUM.



Method	Mean	Std. Dev.	25 % Quantile	Median	75 % Quantile	95 % Quantile	Number of 0s (%)	Unimputed I. Obs
True values (N = 2250)	211.3	486.0	0	0	175	66.6	66.6	
Available cases (N = 1498)	216.2	503.3	0	0	173	66.8	66.8	
Huber's Min/Max, else as above, 22 terminal nodes	223.0	516.7	0	0	173	66.6	66.6	1
OLS, else as above, 31 terminal nodes	213.6	488.8	0	0	194	66.5	66.5	1
OLS, min = 75, else as above, 18 terminal nodes	214.5	492.0	0	0	191	66.3	66.3	0
OLS, min = 75, as previous but variables C E S added, M excluded, 22 terminal nodes	208.7	494.3	0	0	157	66.8	66.8	1
OLS, min =75, only four variables G S M N, 18 terminal nodes	229.6	533.9	0	0	194	65.7	65.7	0

Table 1. WAID test results for alcoholic drinks (DRINKS), Consumption data. Weighting scheme OLS = Ordinary Least Square Method. Explanatory variables: Classified Age (A), Number of Children (C), Decile of Disposable Income Distribution (D), Education Level of Breadwinner (E), Gender (G), Mobile Phone at Home (M), Number of Adults (P), Electrically Heated Sauna Owen (S) and Degree of Urbanization (U).

Method	Mean	Std. Dev.	25 % Quantile	Median	75 % Quantile	95 % Quantile	Number of 0s (%)	Unimputed I. Obs
True values (N = 6011)	395.7	786.1	4	119	469	1662	21.7	
Available cases (N = 4563)	396.3	810.5	4	114	459	1654	22.2	
OLS, expl. vars: C M	397.2	822.6	4	120	467	1629	21.9	0
OLS, 30 terminal nodes	398.1	800.3	4	117	463	1681	21.5	8
OLS, 60 terminal nodes	393.5	796.2	4	116	459	1624	21.8	12
OLS, 80 terminal nodes	395.4	795.4	4	118	461	1654	21.8	15
OLS, 92 terminal nodes (max)	401.2	830.9	4	120	469	1636	21.9	24

Table 2. Imputation Results for Yearly Consumption of Health of Household Member (HP5), Consumption data. All the following explanatory variables have been used unless mentioned: Number of Household Members (M), Number of Children (C), Decile of Disposable Income Distribution, Classified Age of Member (5 years interval), Sosio-Economic Status of Member. The good minimum target size of the groups has been considered as 50. Max = number of terminal nodes without restrictions.

Method	Mean	Std. Dev.	25 % Quantile	Median	75 % Quantile	95 % Quantile	Number of 0s (%)	Unimputed I. Obs
True values (N = 2250)	1057.1	1338.9	244	655	1364	3310	3.6	
Available cases (N = 1498)	1084.5	1414.7	257	665	1409	3339	3.5	
OLS, expl. vars: C M	1079.8	1414.3	258	666	1416	3308	3.6	1
OLS, 20 terminal nodes	1079.2	1376.0	255	668	1413	3236	3.3	3
OLS, 28 terminal nodes (max)	1088.6	1450.9	259	667	1403	3285	3.8	4

Table 3. Imputation Results for Yearly Consumption of Health of Household (KP5).

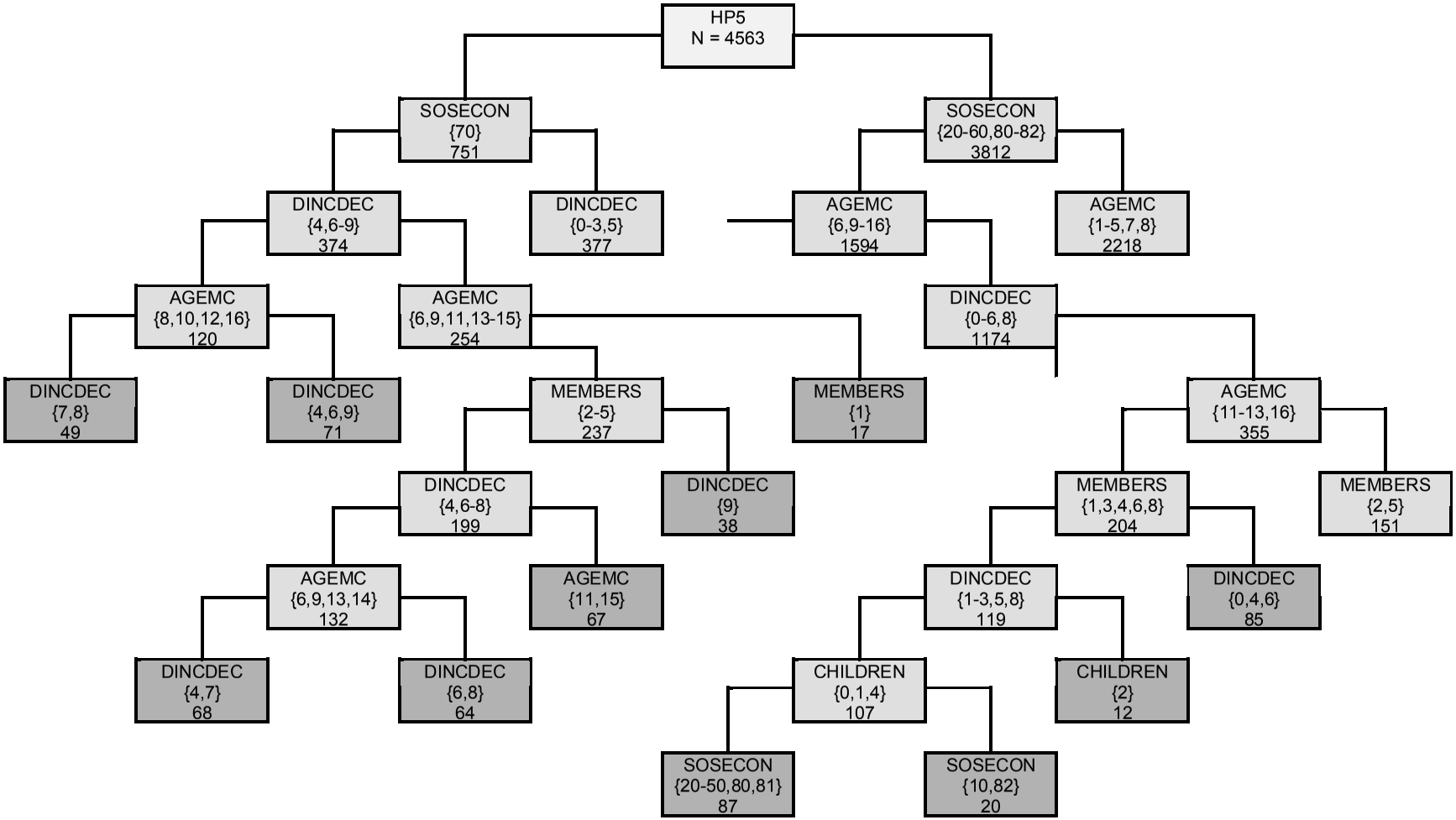
Method	Mean Distance	Number of Rooms (%)					
		1-2	3	4	5	6	7+
True values (N=354)	0	1.4	9.3	23.2	34.2	19.8	12.1
Nearest n. imputation, 30 terminal nodes	1.26	4.0	6.5	24.3	30.5	22.6	12.1
Nearest n. imputation, 60 terminal nodes	1.29	3.4	7.6	22.9	30.5	23.7	11.9
Nearest n. imputation, >80 terminal nodes	1.15	2.8	7.6	25.2	29.1	22.6	12.7
Random imputation, full tree (131 tn.)	1.12	3.1	8.8	21.8	30.2	22.9	13.2

Table 4. Imputation results for Number of rooms of household (ROOMSNUM), marginal distributions, Census data. Note, that only imputed values $\hat{Y}$ and their corresponding true values Y^* are used here. Chosen explanatory variables are Household Space Type (HHSPTYPE), Number of Persons in Household (PERSINHH), Socio-Economic Group (SEGROUP) and Social Class Based on Occupation (SOCLASS). Minimum group size is 50. Mean Distance $= n^{-1} \sum_{i=1}^{n} d(\hat{Y}_i, Y_i^*)$.

Method	Mean Distance	Number of Cars (%)			
		0	1	2	3+
True values (N=167)	0	37.5	45.2	14.3	3.0
Nearest n. imputation, 30 terminal nodes	0.64	37.5	38.7	21.4	2.4
Nearest n. imputation, 60 terminal nodes	0.65	36.9	47.0	13.7	2.4
Nearest n. imputation, full tree (139 tn.)	0.47	36.5	48.5	12.6	2.4
Random imputation, full tree	0.60	35.3	47.9	11.4	5.4

Table 5. Imputation results for number of cars of household (CARS), marginal distributions, Census data. Explanatory variables: Distance to Work (DISTWORK), Number of Persons in Household (PERSINHH) and Socio-Economic Group (SEGROUP).

Figure 2. Part of the WAID regression tree for yearly consumption of health (of household member), HP5.



5. Summaries and Concluding remarks

Tree-based methods for imputation have not been generally used due to unavailable easy tools to continue towards imputations after the construction of trees. The prototype software WAID 4.1 is a new development for this purpose. This software consists of classification tree, on one hand, and of regression tree, on the other. The former method is a good starting technique when needed to impute categorical variables, but it may be used for continuous variables in a limited extent, too. The latter method is available for continuous variables, respectively. Under this technique, there are some options for making a tree building more robust. A limitation of regression tree techniques in WAID is that all explanatory (auxiliary) variables should be categorical or categorized. This means that the method cannot be used successfully applied to many business surveys.

We have tested WAID with two different data sets, one being the Finnish consumption data, and the other the UK census data, respectively. The target variables of the former are consumption item values (continuous), whereas those of the latter are different states of the population (categorical). Both data sets are fairly complex. The complexity of consumption data is due to skew distributions of consumption items, even so that there are often a high number of zeros. On the other hand, these data are of the two levels, both from household level and household member level. At household level about 2200 records are available, whereas at member level more than 6000. The number of observations have, naturally, an impact on the possible number of terminal nodes if the parameters for the tree building have not been changed; the larger data set is, the more terminal nodes will be obviously appearing. For the larger data set, we have performed around 50-100 terminal nodes but around 15-30 for the smaller one. The appropriate minimum for group size has been considered as 50. UK census test data include nearly 20000 households, and two crucial categorical variables have been tried to impute. Hence classification trees have been built.

Terminal nodes or the sub-groups constructed from these trees (regression or classification trees) are interpreted as imputation cells. In an ideal situation, these should be as homogeneous as possible. The second question is, how to impute the missing values within each cell. WAID 4.1 has 4 alternatives, but we have only applied random hot decking (random draw of the real values) and nearest neighbor technique. Both methods necessarily require that there are in each cell a reasonable number of real (neighbor) values, derived from respondents. WAID 4.1 does not give automatically this information or diagnose problematic cells. This could be done using other tools, but we have not done such operations but used WAID rather straightforward. Some not-good test results could be explained with problems in some cells.

Our results in general show that the WAID approach with real-donor methods never gives very poor results. It should be noted that our test data sets do not cover most difficult types of NSI data, such as business survey and longitudinal data sets. Secondly, the current WAID does not seem to be a conformable tool for handling very big data sets. Nevertheless, in our exercises, when a reasonable number of correct explanatory (auxiliary) variables were used, the bias due to selective missingness was reduced essentially. It is still difficult to find an optimum approach in order to decide how many variables should be included in a particular tree model, how these variables should be pre-classified, and how many terminal nodes (imputation cells) should have been tried to use. The diagnostics with graphs and appropriate tabulations would be helping.

References

AutImp (2001). Website of the Project: *http://www.cbs.nl/en/services/autimp/autimp.htm.*
Kalton, G. and Kasprzyk, D. (1986). The Treatment of Missing Survey Data. *Survey Methodology* 12, 1-16.
Laaksonen, S. (2000). Regression-Based Nearest Neighbour Hot Decking. *Computational Statistics* 15, 1, 65-71.
Little, R. and Rubin, D. (1987). *Statistical Analysis with Missing Data*. John Wiley & Sons.
Mesa, D.M., Tsai, P. and Chambers R.L. (2000). *Using Tree-Based Models for Missing Data Imputation: An Evaluation Using UK Census Data*. Forthcoming in Research Papers of Statistics Netherlands. Currently available in the AutImp website.
Plomp, R., de Waal, T. and de Waard, J. (2001). *Manual WAID (4.1)*. Research paper 0118. Statistics Netherlands. See AutImp website.
Rubin, D. (1987*). Multiple Imputation in Surveys*. John Wiley & Sons.
Rubin, D. and the papers and the discussion by B. Fay, J. Rao, D. Binder, J. Eltinge and D.
Judkins (1996). Multiple Imputation After 18+ Years. *Journal of the American Statistical Association* 91, 473-520.
Schulte Nordholt, E. (1998). Imputation: Methods, Simulation, Experiments and Practical Examples. *International Statistical Review, 66, 157-180*.