**Handling Missing Data with CHAID: A Statistical Experiment**

Getting to the data analysis, every sociologist faces the problem of missing values. This problem is usually solved either through exclusion of missing values, thus, reducing the available sample, nor through imputation, preserving the necessary amount of observations. However, both alternatives have their drawbacks. This increases the dignity of using methods of analysis that are able to handle missing values without their removal or imputation. Such techniques are included in many decision tree algorithms. For example, within the most common in sociology algorithm of decision trees CHAID, the missing values are considered a single category that joins the most similar node in the model regarding the distribution of the response. Although the ability to handle missing values is defined as the unique advantage of decision trees in the literature, so far in the publications there were no proofs of the correctness of the direct inclusion of missing values in models.

In this study, a statistical experiment was conducted. The study aimed to find out how well CHAID identifies missing values to the nodes and what are the consequences of the inclusion of missing data into the analysis. The experiment was based on three conditions: the location of the variable regarding the root, the accuracy of the tree, and proportion of missing values. A total of 780 experiments were conducted on the sample of 3000 observations.

The results of the experiments show that, overall, CHAID correctly identifies missings to the nodes. In most cases, the nodes, which were filled with missing values, are located at the appropriate depth, combine the right categories of the independent variable and present the correct distribution of the dependent variable; thus, nodes reproduce the actual structure of the relationships in the data. However, the bad results are observed in the case when the proportion of missing values is very high (in this study - 50%) independently of other conditions.

Nevertheless, in most cases, the inclusion of missing data is accompanied by structural changes in the tree. The four possible types of tree damage were identified: structural changes in nodes at the same depth as the node with missing values, the presence of waste nodes, the absence of nested nodes, and a change of the variables` location. In the last case, in the resulting trees, most nodes do not represent the actual structure of the relationships in the data in general. Also, the trees are severely damaged in cases when missing values are in variables that are not located at the maximum possible depth, for example, next to the root or in the middle of a tree. The inclusion of missing values for such variables can entail changes in the structure of all nodes located below. However, if the variable is at the maximum possible depth, the missing values damage the tree slightly, as their effect extends only to parallel nodes.

In most cases, CHAID performs exploratory functions, and researcher does not know what the initial relationship structure is, whether missing values are correctly defined and how damaged tree was obtained. At the same time, the researcher has no indicators of the fact that the tree is damaged: it has been shown that neither high accuracy of the received tree nor its sustainability (not-overfitting) can be the reason to deny possible damage of a tree.

Thus, within deciding whether to include missing values in the CHAID model, the following factors should be considered:

1. *Purpose of the analysis*. If the researcher has only the predictive purpose (classification or regression), then CHAID copes with this goal successfully, including the cases of presence of missing values. However, when it comes to searching for interaction effects, there is a high probability of obtaining artifact results. This fact leads to making erroneous conclusions.
2. *Location of the variable with missing values in the tree structure*. If missing values are not "in the bottom of" a tree but in any other place, then damage of a tree can be severe – up to complete discrepancy to the real structure of relationships. The closer an independent variable is located to the root of the tree - the more intense the tree damages.
3. *The percentage of missing values*. In the case of a high percentage of missing values (50%), the model often does not correspond to the established structure of relationships in the data - more than half the time. The probability to receive the appropriate model on this condition is low.

These results demonstrate that the possibility of handling missing data in decisions trees is defined in the literature as an advantage of these methods in vain: the risk of receiving false and erroneous conclusions exists – as well as in the case of the imputation of missing values. It is important to note that the design of the experiment represented the simple decision tree structure that included only three predictors with a small number of categories and a dichotomous dependent variable. It should be assumed that the complication of the tree structure (which the models in real empirical studies are) would only aggravate the consequences of including them in the analysis.