Review of Ph.D. dissertation:

„Comparison of data missing management processes and imputation"

Miklós Péter, Máder

Tutor: Dr. Mária, Székelyi DSc.

I. thesis (publication: Máder 2005)

The data missing occurs for various reasons. If it occurred randomly, they do not have significant effect on the observed distribution of cases. Systematic data missing mechanisms have serious biased effect on the distribution of cases, so the conclusions won't be correct as well. There are three levels of data missing mechanisms; they are the MCAR, the MAR and NOTMAR mechanisms. The effectiveness of imputation is different by these three missing data mechanisms.

II. thesis (publication: Máder 2005)

The increasing data missing demands increasing level of data management. The higher the level of data missing and the more systematic pattern the data missing has, the level of bias is increasing.

III. thesis (publication: Máder 2004)

I completed the Oravecz-defined five missing data patterns, by introducing of sociological explanation approach. I showed at the MAR data missing mechanism, that some data missing have uniformed data missing pattern, so we can define the attitude of respondents.

IV. thesis (publication: Máder 2005)

The imputation procedures can reduce the biased of the given estimations of incomplete databases. The databases with MAR or NOTMAR missing data mechanisms are distorted. However, there are sophisticated imputation methods, which can correctly impute

values, so the heterogeneity and mean of completed database can be restored. The HOT DECK imputation method was the most effective; this method estimated the distribution of the deleted values well.

V. thesis (publication: Máder 2005)

Laaksonen divided the imputation methods into four main categories. I showed that the model-donor imputation and donor imputation were effective approaches, but the naive approaches had insufficient result.

VI. thesis (publication: 5.2nd chapter of the dissertation)

The imputation methods can be based on both explicit and implicit function, or on a mixture of these functions. Implicit models use the range of observed values, which can be found in the database. The explicit models are similar to the implicit models, the algorithms use the information of the database, but they impute the values from the model. The imputed value is intended to extrapolate on the basis of this estimation algorithm. The differences between them:

• At MCAR missing data mechanism the implicit models are more efficient than the explicit methods

• At MAR missing data mechanism the explicit models were more effective because of the correlation with the data missing oriented these procedures

VII. thesis (publication: 8.5th chapter of the dissertation)

The characteristic of regression analysis is to extrapolate beyond the original range; it can estimate values even outside of the range of the given distribution. These extreme values should recode into the theoretical maximum or minimum values.

VIII. thesis (publication: 10.1st chapter of the dissertation)

The NOTMAR missing data mechanism can be handle by three steps, as a first step, other independent surveys' results should be taken to receive information by the extent and characteristics of the biased of the distribution, the second step is to correct the distribution by an deflationary function and finally we need to impute the missing values in this adjusted database.