Imprecise Imputation for Statistical Matching

Eva Endres EVA.ENDRES@STAT.UNI-MUENCHEN.DE
Paul Fink PAUL.FINK@STAT.UNI-MUENCHEN.DE
Thomas Augustin AUGUSTIN@STAT.UNI-MUENCHEN.DE
Department of Statistics, Ludwig-Maximilians-Universität München (LMU Munich)
Munich (Germany)

Poster Abstract

The task of statistical matching, where two (or more) data samples with a partially overlapping set of variables should be merged, can be interpreted as a special, block-wise pattern of missing values with the following properties: (i) the missing values are missing completely at random and, (ii) there does not exist a single observation with information on all variables (D'Orazio et al., 2006, p.6). While the former property is desirable in the context of missing values, the latter leads to a drastic identification problem in the estimation of parameters concerning those variables which have not jointly been observed.

One way of dealing with this problem is the imputation of all missing values. Hot deck imputation is a method that originates from the missing data problem, where the missing values of each (*recipient*) record are substituted by an observed value from a similar (*donor*) record in the sample (e.g. Little and Rubin, 1987, p.62). This method is frequently used in practice, comparatively easy to apply and nonparametric (e.g. Andridge and Little, 2010). However, it is a well-known issue that any single imputation is not able to reflect the uncertainty which arises from the missingness.

In our poster, we propose an imprecise single imputation approach, based on a generalization of the random hot deck single imputation with donation classes (e.g. D'Orazio et al., 2006, p.37). It enables us to take the uncertainty of the statistical matching problem into account. We substitute every missing value with a *set* of suitable values observed within the same donation class. Thus, we produce a coarse complete data set with partly set-valued observations, which are used to estimate lower and upper bounds for the parameters of not jointly observed variables. We discuss different strategies to determine the set of values to be imputed. The extreme case where the variable domains are imputed, resembles the approach of Ramoni and Sebastiani (2001).

References

R. Andridge and R. Little. A review of hot deck imputation for survey non-response. *International Statistical Review*, 78(1):40–64, 2010.

M. D'Orazio, M. Di Zio, and M. Scanu. *Statistical Matching: Theory and Practice*. Wiley, Chichester, United Kingdom, 2006.

R. Little and D. Rubin. *Statistical Analysis with Missing Data*. Wiley, New York, NY, 1987.

M. Ramoni and P. Sebastiani. Robust learning with missing data. *Machine Learning*, 45(2):147–170, 2001.