UNITED NATIONS ECONOMIC COMMISSION FOR EUROPE
CONFERENCE OF EUROPEAN STATISTICIANS

Expert Meeting on Statistical Data Editing
7-9 October 2024, Vienna Austria

12 July 2024

Full conditional distributions for handling restrictions in the context of automated statistical data editing

Christian Aßmann, Ariane Würbach, Younes Saidani, & Florian Dumpert (Leibniz Institute for Educational Trajectories)

christian.assmann@lifbi.de

Abstract

Reported survey data typically contain inaccuracies due to respondent errors, as reported values may be missing or may not comply with logical restrictions. When such logical restrictions involve more than one variable, it is also unclear which variable or variables are in fact faulty. Statistical offices have often established edit-imputation routines following the Felligi-Holt paradigm to correct data and ensure data coherence, thereby employing an easily computable heuristic that does not necessarily use all information available in the observed data. In contrast, Bayesian methods for edit-imputation incorporate all available information in the full conditional distributions of missing values and correctly reflect the uncertainty arising from the process of replacing erroneous values. While for categorical and continuous data, Bayesian approaches based on parametric models are available in the literature, this article lays out a method for specifying full conditional distributions using classification and regression trees while taking into account nested balance restrictions, i.e. linked restrictions involving several variables. The CART algorithm was chosen, because it provides flexible univariate approximations to the full conditional distributions of the variables while reducing the computational intensity of the overall Bayesian approach. The feasibility of the suggested approach is documented in terms of a simulation study and an empirical application based on survey data from the Federal Statistical Office of Germany. Simulation results suggest that compared to complete case analysis, average root mean squared error of moment estimates can typically be reduced by 20 to 30 percent when using the non-parametric Bayesian approach and the corresponding specification of full conditional distributions using the CART algorithm.