***Dr Natalie Shlomo**, School of Social Sciences, University of Southampton*

Calibrated Imputation of Numerical Data Under Linear Edit Restrictions

Jeroen Pannekoek[1], Natalie Shlomo[2] and Ton de Waal[1]

[1] Jeroen Pannekoek, Ton De Waal, Statistics Netherlands, PO Box 24500, 2490 HA The Hague, The Netherlands (email: j.pannekoek@cbs.nl and t.dewaal@cbs.nl)
[2] Natalie Shlomo, Southampton Statistical Sciences Research Institute, University of Southampton, Highfield, Southampton SO17 1BJ, United Kingdom (email: n.shlomo@soton.ac.uk)

***Abstract*:** A common problem faced by statistical offices is that data may be missing from collected datasets. The typical way to overcome this problem is to impute the missing data. The problem of imputing missing data is complicated by the fact that statistical data often have to satisfy certain edit rules and that values of variables sometimes have to sum up to known totals. The edit rules are most often formulated as linear restrictions on the variables that have to be satisfied by the imputed data. For example, for data on enterprises edit rules could be that the profit and costs of an enterprise should sum up to its turnover and that that the turnover should be at least zero. The totals of some variables may already be known from administrative data (turnover from a tax register) or estimated from other sources. Standard imputation methods for numerical data as described in the literature generally do not take such edit rules and totals into account. We describe algorithms for imputation of missing numerical data that take edit restrictions into account and ensure that sums are calibrated to known totals. These algorithms are based on a sequential regression approach that uses regression predictions to impute the variables one by one. For each missing value to be imputed we first derive a feasible interval in which the imputed value must lie in order to make it possible to impute the remaining missing values in the same unit in such a way that the imputed data for that unit satisfy the edit rules and sum constraints. To assess the performance of the imputation methods a simulation study is carried out as well as an evaluation study based on a real dataset.

Keywords: linear edit restrictions, sequential regression imputation, Fourier-Motzkin elimination,

benchmarking

[1] Jeroen Pannekoek, Ton De Waal, Statistics Netherlands, PO Box 24500, 2490 HA The Hague, The Netherlands (email: j.pannekoek@cbs.nl and t.dewaal@cbs.nl)
[2] Natalie Shlomo, Southampton Statistical Sciences Research Institute, University of Southampton, Highfield, Southampton SO17 1BJ, United Kingdom (email: n.shlomo@soton.ac.uk)