UNITED NATIONS ECONOMIC COMMISSION FOR EUROPE
CONFERENCE OF EUROPEAN STATISTICIANS

UNECE Virtual Workshop on Statistical Data Editing 2020

31 August - 4 September 2020

Topic: Quality – assessing data quality and indicators

Topic organizers: Sander Scholtus (Statistics Netherlands) and Pedro Revilla (INE, Spain)

Topic: Quality – points for discussion

(presentation by Statistics Canada)

- Focus on a single variable of interest. What if there are several variables of interest?
  - Distributional accuracy also important for multivariate relations
  - How to visualize results?

- Imputation methods for RCS – besides donor imputation – impute predicted values without added noise (?). Could this explain why tails of distributions were slightly underrepresented in imputed values?
  - See plots for “distributional accuracy”

Topic: Quality – points for discussion

(presentation by Statistics Canada)

- In practice, imputation strategies at NSIs are often complicated, involving several methods and parametrizations applied in sequence. Reproducing this complexity in a simulation study is challenging:
  - Simulating the entire imputation process many times can be time-consuming
  - Production system may not lend itself to use on simulated data
- How to analyse the quality of imputation strategies through simulation, without over-simplifying?
- Experiences of other NSIs?

Topic: Quality – points for discussion

(presentation by Statistics Netherlands)

- Focus on evaluating the variance of estimated frequency tables based on mass imputation
- Context: Dutch decennial virtual population census
- Problem relevant for other applications in which a categorical variable is imputed for an entire target population
- Two approaches tested: analytical approximations and bootstrapping (extension based on pseudo-populations)
- Both approaches have been tested on a small artificial population in a simulation study
- Analytical approach depends on the specific context of mass imputation, not easily generalizable, requires simplifying assumptions
- Bootstrap approach: more flexible, fewer assumptions, computationally intensive

Topic: Quality – points for discussion

(presentation by Statistics Netherlands)

- Multiple imputation: interesting to compare to analytical and bootstrap methods
- Future work: extending variance estimation methods
- The analytical approach could be extended to the case of imputing multiple categorical variables (or a combination of categorical and numerical variables) and to handle imputation methods not based on a parametric model (e.g., hot deck)
- Develop variance estimation techniques that can account for uncertainty in the measurement of register-based variables, including the effects of micro-integration
- Experiences of other NSIs?

Topic: Quality – points for discussion

- Focus in two presentations on evaluating the quality of / estimating the uncertainty due to *imputation*. Any experiences with evaluating the quality of / estimating the uncertainty due to *editing*?
  - Overall effect of editing
  - Contributions of individual editing process steps

- Any interesting experiences with reporting about quality of editing/imputation to end users?