SIMULATING MULTIPLE IMPUTATION

APPLICATION TO THE GERMAN AGRICULTURAL CENSUS 2010

Lydia Spies, Sven Schmiedel, Katrin Schmidt
UNECE Work Session on Statistical Data Editing
Paris, 28 – 30 April 2014

© Federal Statistical Office of Germany | Department for Mathematical-Statistical Methods
The Aim

- Implementation of a software tool in R for an automation of
  - the generation of missing data by different mechanism
  - the imputation of missing values by different techniques
  - the creation and the comparison of the (combined) results
- Comparison of the performance of varying imputation approaches in practice
- Assessment of the impact of different underlying missing mechanism
- Acceleration and simplification of the individual decision for a final imputation
Test Application

- German Agricultural Census 2010:
  - One incomplete item: water consumption in m³
  - 16% missing or implausible values
- 1000 random samples
- MAR (missing at random) scenario implemented through a logistic regression
- Bayesian multiple imputation under the normal linear model
  - with a hot-deck component
  - Predictive mean matching
- Number of imputations: 5, 20, 200
- Transformations: no transformation, logarithm, cubic root
Performance measures

- Bias
- Standard deviation
- Coefficient of variation
- Root mean square error
- Coverage rate
- Confidence interval width

Parameter of interest: mean water consumption
Main Results

- Predictive mean matching:
  - Plausible values
  - Good results

- Hot-deck approach:
  - Implausible values possible with no or cubic root transformation
  - Serious bias observed with log transformation
- Performance depends on the number of imputations
- The overall estimate and its total variance stabilize very slowly
Next Steps

- Investigation of the factors that influence the number of imputations needed until the estimates stabilize
- Implementation of further imputation methods
- Simulation with non-random (e.g. stratified) samples