DESTATIS
wissen. nutzen.

POSSIBLE IMPUTATION PROCEDURES FOR THE CENSUS 2021

Lydia Spies

UNECE Work Session on Statistical Data Editing

The Hague, 24 – 26 April 2017

Motivation

- German Population Census 2011:
  - Validation and imputation was executed by IT-NRW (the Central statistical and IT services provider of one of the German Federal States, i.e. North Rhine-Westphalia)
  - Extra ad hoc nearest neighbour imputation tool was developed for the respective IT-environment
  - No exact error estimation for most survey estimates
- German Population Census 2021:
  - Validation and imputation will most probably be executed by the Federal Statistical Office
  - Exact error estimation for more than just the population figures is under evaluation

Considered imputation methods and tools

Methods	Tools
▪ Multiple imputation	▪ R package *mice* ▪ SAS 9.4 MI procedure
▪ Nearest neighbour imputation	▪ German ad hoc imputation tool ▪ CANCEIS

Multiple imputation

- Fully conditional specification
- Mixture of polytomous regression models and predictive mean matching
- m=10
- Re-imputation of implausible imputed values (shoot-out)
- Application of the same procedure and models in R and SAS
- Ad hoc implementation of the shoot-out procedure

Advantages and challenges of MI

- Advantage:
  - Enables correct error estimation
    - MI makes only sense if an exact error estimation is applied
- Challenges:
  - Results in m datasets
    - Every analysis has to be performed m times
    - Extra user interface necessary
  - Computationally very expensive
    - Neither R nor SAS are able to generate the multiple imputations in an acceptable time
    - Ad hoc implementation necessary

Advantages and challenges of MI

- Quality of the results not comparable to the quality of other statistics
  - Relative standard deviation rises when considering the variance due to imputation but not because of a lower quality but due to an improved methodology
- Deviation from standard publication rule
  - Results with a relative standard deviation of 15% and higher are usually not published
  - Sticking to this rule would cause a reduction of publishable results
- Edit rules cannot be incorporated into the imputation process
  - Shoot-out procedure can be very time consuming

Conclusions about MI

- MI is only an option if an exact error estimation is applied
- MI comes along with a lot of mostly practical challenges
- Before MI can be applied on the German population census a lot of work still has to be done
- Focusing on MI would bear a high project risk

Nearest neighbour imputation tools

- German ad hoc imputation tool:
  - Two distance measures implemented
    - For categorial variables only a simple comparison is available
  - All variables get the same weight
  - Subsequent validation step and additional adoption of the now implausible values from the same donor
- CANCEIS:
  - Ten distance measures implemented
  - Variable weights can be chosen
  - Near minimum change imputation action

Comparison of nearest neighbour imputation tools

- Advantages CANCEIS:
  - Flexible distance functions
  - Elaborated quality assessment of potential donors
  - Result of over 20 years of improvements
- Advantages German ad hoc imputation tool:
  - Is able to interact with the German standard validation tools
  - Expandable
  - Potential for a German standard nearest neighbour imputation tool

Conclusions about nearest neighbour imputation tools

- Improving the German ad hoc tool might be worth the effort
- Ad hoc solutions of other statistics could be included later on
- But this takes a lot of time and effort
- Simulation study planned to compare the quality of the results of CANCEIS and the German ad hoc tool