

# Microediting and Software. Introduction

Matthias Templ<sup>1,2</sup>

<sup>1</sup> Department of Methodology, Statistics Austria

<sup>2</sup> Department of Statistics and Probability Theory, TU WIEN, Austria

UNECE Worksession on data editing  
Ljubljana, Mai 10, 2011

## Session (iv): Microediting and Software

- New topic at the work session with many contributions
- Microediting gets more and more important as well as software solutions are required.
- We have three presentations with strong focus on R and three with focus on SAS.
- The session is biased from the subjective view of the session chair, i.e. the focus lay more on R since it is the standard software used by more users than SAS or any other statistical software, and it is used by all innovative large big players like facebook, google, pfister, etc.

## Session (iv): Microediting and Software

- New topic at the work session with many contributions
- Microediting gets more and more important as well as software solutions are required.
- We have three presentations with strong focus on R and three with focus on SAS.
- The session is biased from the subjective view of the session chair, i.e. the focus lay more on R since it is the standard software used by more users than SAS or any other statistical software, and it is used by all innovative large big players like facebook, google, pfister, etc.

## Session (iv): Microediting and Software

- New topic at the work session with many contributions
- Microediting gets more and more important as well as software solutions are required.
- We have three presentations with strong focus on R and three with focus on SAS.
- The session is biased from the subjective view of the session chair, i.e. the focus lay more on R since it is the standard software used by more users than SAS or any other statistical software, and it is used by all innovative large big players like facebook, google, pfister, etc.

## Session (iv): Microediting and Software

- New topic at the work session with many contributions
- Microediting gets more and more important as well as software solutions are required.
- We have three presentations with strong focus on R and three with focus on SAS.
- The session is biased from the subjective view of the session chair, i.e. the focus lay more on R since it is the standard software used by more users than SAS or any other statistical software, and it is used by all innovative large big players like facebook, google, pfister, etc.

## Session (iv): Contributions

- R** Model-based Imputation of complex data using robust methods (Templ, Kowarik, and Filzmoser 2011b, Templ, Alfons, and Kowarik 2011a)
- R** Imputation using hot-deck methods from R-package StatMatch (D'Orazio, Di Zio, and Scanu 2006, D'Orazio 2011)
- R** Outlier detection at microdata level with missing values (Todorov, Templ, and Filzmoser 2011, Todorov 2010), based on statistical estimation.
- SAS** Using Priority Scores for Business Surveys: Fellegi-Holt and H-B soft edit in Banff, deterministic rules (Canada 2009).
  - (Innovative use of significance editing in USDA's National Agricultural Statistics Service)
- SAS** The Edit (Anders Norberg), hard and soft edits.
- SAS** Germany - Redesigning a data editing concept, hard balanced edits, deterministic rules.

## Session (iv): Contributions

- R Model-based Imputation of complex data using robust methods (Templ et al. 2011b,a)
- R Imputation using hot-deck methods from R-package StatMatch (D'Orazio et al. 2006, D'Orazio 2011)
- R Outlier detection at microdata level with missing values (Todorov et al. 2011, Todorov 2010), based on statistical estimation.
- SAS Using Priority Scores for Business Surveys: Fellegi-Holt and H-B soft edit in Banff, deterministic rules (Canada 2009).
  - (Innovative use of significance editing in USDA's National Agricultural Statistics Service)
- SAS The Edit (Anders Norberg), hard and soft edits.
- SAS Germany - Redesigning a data editing concept, hard balanced edits, deterministic rules.

## Session (iv): Contributions

- R Model-based Imputation of complex data using robust methods (Templ et al. 2011b,a)
- R Imputation using hot-deck methods from R-package StatMatch (D'Orazio et al. 2006, D'Orazio 2011)
- R Outlier detection at microdata level with missing values (Todorov et al. 2011, Todorov 2010), based on statistical estimation.
- SAS Using Priority Scores for Business Surveys: Fellegi-Holt and H-B soft edit in Banff, deterministic rules (Canada 2009).
  - (Innovative use of significance editing in USDA's National Agricultural Statistics Service)
- SAS The Edit (Anders Norberg), hard and soft edits.
- SAS Germany - Redesigning a data editing concept, hard balanced edits, deterministic rules.



## Session (iv): Contributions

- R Model-based Imputation of complex data using robust methods (Templ et al. 2011b,a)
- R Imputation using hot-deck methods from R-package StatMatch (D'Orazio et al. 2006, D'Orazio 2011)
- R Outlier detection at microdata level with missing values (Todorov et al. 2011, Todorov 2010), based on statistical estimation.
- SAS Using Priority Scores for Business Surveys: Fellegi-Holt and H-B soft edit in Banff, deterministic rules (Canada 2009).
  - (Innovative use of significance editing in USDA's National Agricultural Statistics Service)
- SAS The Edit (Anders Norberg), hard and soft edits.
- SAS Germany - Redesigning a data editing concept, hard balanced edits, deterministic rules.

## Session (iv): Contributions

- R Model-based Imputation of complex data using robust methods (Templ et al. 2011b,a)
- R Imputation using hot-deck methods from R-package StatMatch (D'Orazio et al. 2006, D'Orazio 2011)
- R Outlier detection at microdata level with missing values (Todorov et al. 2011, Todorov 2010), based on statistical estimation.
- SAS Using Priority Scores for Business Surveys: Fellegi-Holt and H-B soft edit in Banff, deterministic rules (Canada 2009).
  - (Innovative use of significance editing in USDA's National Agricultural Statistics Service)
- SAS The Edit (Anders Norberg), hard and soft edits.
- SAS Germany - Redesigning a data editing concept, hard balanced edits, deterministic rules.

## Session (iv): Contributions

- R Model-based Imputation of complex data using robust methods (Templ et al. 2011b,a)
- R Imputation using hot-deck methods from R-package StatMatch (D'Orazio et al. 2006, D'Orazio 2011)
- R Outlier detection at microdata level with missing values (Todorov et al. 2011, Todorov 2010), based on statistical estimation.
- SAS Using Priority Scores for Business Surveys: Fellegi-Holt and H-B soft edit in Banff, deterministic rules (Canada 2009).
  - (Innovative use of significance editing in USDA's National Agricultural Statistics Service)
- SAS The Edit (Anders Norberg), hard and soft edits.
- SAS Germany - Redesigning a data editing concept, hard balanced edits, deterministic rules.

## Session (iv): Contributions

- R Model-based Imputation of complex data using robust methods (Templ et al. 2011b,a)
- R Imputation using hot-deck methods from R-package StatMatch (D'Orazio et al. 2006, D'Orazio 2011)
- R Outlier detection at microdata level with missing values (Todorov et al. 2011, Todorov 2010), based on statistical estimation.
- SAS Using Priority Scores for Business Surveys: Fellegi-Holt and H-B soft edit in Banff, deterministic rules (Canada 2009).
  - (Innovative use of significance editing in USDA's National Agricultural Statistics Service)
- SAS The Edit (Anders Norberg), hard and soft edits.
- SAS Germany - Redesigning a data editing concept, hard balanced edits, deterministic rules.

- Statistics Canada. Functional description of banff - the generalized edit and imputation system. Technical report, Statistics Canada, 2009.
- M. D'Orazio. StatMatch: Statistical Matching, 2011. URL <http://CRAN.R-project.org/package=StatMatch>. R package version 1.0.1.
- M. D'Orazio, M. Di Zio, and M. Scanu. Statistical Matching: Theory and Practice. John Wiley & Sons, 2006.
- M. Templ, A. Alfons, and A. Kowarik. VIM: Visualization and Imputation of Missing Values, 2011a. URL <http://CRAN.R-project.org/package=VIM>. R package version 1.4.3.
- M. Templ, A. Kowarik, and P. Filzmoser. Iterative stepwise regression imputation using standard and robust methods. Computational Statistics & Data Analysis, In Press, Accepted Manuscript:–, 2011b. ISSN 0167-9473. URL <http://www.sciencedirect.com/science/article/B6V8V-52R7YYH-2/2/2c6c9ed7138d50c4197e991d8ffb8a1f>. DOI: 10.1016/j.csda.2011.04.012.
- V. Todorov. rrcovNA: Scalable Robust Estimators with High Breakdown Point for Incomplete Data, 2010. URL <http://CRAN.R-project.org/package=rrcovNA>. R package version 0.3-00.
- V. Todorov, M. Templ, and P. Filzmoser. Detection of multivariate outliers in business survey data with incomplete information. Advances in Data Analysis and Classification, 5:37–56, 2011.