Topic (ii): New and Emerging Methods

Introduction

Maria Garcia (USA)
Jeroen Pannekoek (Netherlands)
Topic (ii): New and Emerging Methods

- Papers under this topic cover new and emerging methods or other methods for improving and/or optimizing the process of data editing and imputation.

- There are eight papers addressing improvement in the following areas.
  - Development of E&I for new kinds of data sets
  - Improving imputation by selecting better methods
  - Detection of outlying, suspect or erroneous values
  - Evaluation methods/measures for E&I
E&I for new kinds of data sets

- Imputation with multi-source data: the case of Italian SBS – Istat.
  Construction of a micro-data file from different administrative sources with imputation to fill in the gaps of non-observed values. Many constraints on the variables.

- Data editing and scanner data – France
  Compute CPI using scanner data instead of data collected by persons. Vast amount of data is a performance challenge for automatic editing.
Improving imputation for particular applications by better methods

- A new imputation methodology for the Agricultural Resource Management Survey – USA/NASS
  Old method: Condition mean
  New method: Iterative Sequential Regression, ISR

- Multiple imputation methods for imputing earnings in the SIPP - USA Census Bureau.
  Old method: Randomized hot-deck
  New method: Sequential Regression Multivariate Imputation (SRMI).
Detection of outlying, suspect or erroneous values

- Development of outlier treatment in HCSO – Hungary
  - Overview of outlier detection methodology.
  - Current method: much work by experts, not reproducible.
  - Alternative: develop a quicker more automated method.

- A generalized Fellegi-Holt paradigm for automatic editing – Netherlands
  - Automatic error localization of random errors.
  - Generalizes the FH-paradigm with the purpose of making automatic editing more similar to manual editing.
Evaluation methods/measures for E&I

- Simulating multiple imputation – Germany
  Aim is to support the choice of imputation model by a tool in R. Simulating different missing data mechanisms and comparing different multiple imputation techniques using a number of performance measures.

- Implementation and Evaluation of Automatic Editing – Netherlands
  Implementation of automatic editing as a sequence of process steps performing different tasks. Evaluation measures across process steps. To optimize their configuration.
Enjoy the presentations!

Agenda

- **Italy** - Imputation with multi-source data: the case of Italian Structural Business Statistics
- **Hungary** - Presentation and development of outlier treatment in HCSO
- **Germany** – Simulating Multiple Imputation of Water Consumption in the Agricultural Census
- **United States Census Bureau** – Multiple Imputation Methods for Imputing Earnings in the Survey of Income and Program Participation

10.40 *Break* (30 minutes)

- **United States/NASS** – Assessing the Impact of a New Imputation Methodology for the Agricultural Resource Management Survey
- **Netherlands** – A generalized Fellegi-Holt paradigm for automatic editing
- **France** – Data editing and scanner data
- **Netherlands** – Implementation and evaluation of automatic editing

11.50 *Lunch*

14.00 *Discussion*
Topic (ii): New and Emerging Methods

Summary
Summary

- **Italy**
  - Imputation of a multi-source admin. data set for SBS.
  - Imputation involved a sequence of 4 methods to take the properties of variables into account.
  - Largest difference with traditional method is due to sampling error, not measurement error.

- **Hungary**
  - Reviews the current outlier detection method and considered a number of alternative indicators.
  - Aim is to improve automation, reproducibility and analysis
  - Considers development of a tool in R
Summary

- Germany
  - Development of a new R software tool for repeated simulation of different types of nonresponse and different imputation techniques.
  - Test application: Multiple imputation. Linear regression with PMM or hot-deck residuals. PMM performs better.

- USA Census Bureau
  - Use the Sequential Regression Multivariate Imputation (SRMI) method in a population survey.
  - Results show SRMI is a feasible alternative to the Hot deck and it has the potential to improve estimates.
  - Imputations are set up in a multiple imputation framework.
Summary

- **USA National Agricultural Statistics Service**
  - Use Iterative Sequential Regression for imputation in an agricultural survey
  - Results show the new method preserves important relationships and distributions of response data while providing a better estimate of uncertainty.

- **Netherlands – Generalization of FH**
  - Extension of the definition of the error localization problem: finding the minimum number of edit operations rather than finding the minimum number of fields to impute.
  - Simulation study shows improvements over FH.
Summary

- **France**
  - Uses large volume scanner data for the CPI.
  - Confidence intervals used in data editing can be more precise, but the performance of computations with “Big data” is a problem.

- **Netherlands – Evaluation of E&I**
  - A step-by-step implementation and evaluation of the E&I process.
  - Feedback on edit rules, deterministic and model based methods and data.
Summary

- EUROSTAT (Tabled paper)
  - Imputation of continuous variables using four separate data mining procedures
Topic (ii)  Discussion ...............
Topic (ii): Points for discussion

How difficult is it to select and specify the correct imputation method/model. And to re-configure when variables or sources change. Can this be done in routine production, or must it be done by methodology staff.

We see several applications of Multiple Imputation, but MI is still not much applied in standard production at NSI’s. What are the reasons for or against using MI? Are other methods for variance estimation also considered (replication, model-assisted)?

Some problems are reported with outliers when using imputations drawn from a distribution (Germany, USA/NASS). Is there remedy?
Topic (ii): Points for discussion

The Fellegi-Holt approach to error localization (minimum number of fields to change) does not always identifies the errors that human editors do. Are there experiences with shortcomings of the FH-approach? Alternatives for error localisation?

What are the differences between the methods ISR (Iterative sequential regression) used by USA/NASS and the SRMI (Sequential regression multiple imputation) used by the Census Bureau.