



EESW (Session 7): Data Processing and Validation

Pete Brodie

Office for National Statistics

Overview

- Paper 1
 - Missing data treatment in administrative fiscal sources in French structural business statistics production system
 - **Deroyon**
- Paper 2
 - Automatic data editing functions for establishment surveys
 - **Pannekoek et al.**
- Paper 3
 - The Italian new survey COEN 2011: an innovative editing procedure
 - **Seri et al.**

Paper 1 – Deroyon

- Ability to identify the missing businesses is crucial. How successful was the identification exercise?
- The "micro-imputation" method used for the very smallest businesses is very basic. Is there any reason why the "mean imputation" method for the larger businesses is not extended to these businesses?
- Why was average growth used rather than, for example, a "ratio of means" approach? This is widely used and often less biased.
- Possible bias involved when using continuing businesses to model what is happening with dying businesses.

Paper 2 – Pannekoek et al.

- This looks like a very useful decomposition of the editing process. Could this be used more generally for any editing process , or is there some reason for concentrating on automatic editing?
- Are the R packages used for a wide range of applications or specific to the methods carried out in the example?
- Were there any measures of the quality impact of automatically correcting all failures? For example by manually checking a subsample.

Paper 3 – Seri et al.

- Interesting that businesses with 3+ employees are used. Is the number 3 crucial or just convenient?
- How successful is the mixture modelling used to identify unit errors? Is it better than deterministic thresholds?
- The analysis of raw and edited data has led to changes to the survey process. Does this include questionnaire improvements (field testing)?
- Selemix uses a slightly non-standard approach to selective editing which is useful where standard approach cannot be applied. How does it compare?

Conclusions

- Useful to note that many similar challenges are being faced across NSIs.
- The main thrust seems to be at improving the efficiency of the editing process to concentrate on where value can best be added. Measuring quality is key.
- The new challenges of using administrative data are forcing us to question conventional methods and extend automatic approaches.
- Important to apply the lessons learned from using administrative data and “back fit” to conventional data editing.