Editing and Imputation
Agenda

• opening presentation on the E&I subworkpackage.
• presentations given by E&I team members with discussion (around 15 + 5)
  • around 09:20 presentation by Claus
  • around 09:40 presentation by Bart & Anneleen
  • around 10:00 presentation by Roberta & Fabiana
  • around 10:20 BREAK
  • around 10:40 presentation by Sebastian
  • around 11:00 presentation by Romina & Fabrizio
  • around 11:20 presentation by Florian
  • around 11:40 "Questions for tomorrow"
Beginning: May 2019

• Participants from
  • Belgium
  • Germany
  • Italy
  • UK
  • Poland
  • Italy
  • Switzerland (later)
  • Australia

• Definitions
  • Editing: the task to find missing and problematic data (i.e. unplausible values, contradictions in records, ...)
  • Imputation: altering incorrect values and inserting missing values
Editing
find missing and problematic data

Imputation
insert missing values
Editing
find missing and problematic data

Imputation
insert missing values

correct incorrect data
Value added? (editing 1)

ML may discover rules that have only been “known” by intuition before. This may help

• to conserve knowledge over time and changes in editing teams

• to formalize the knowledge and enables us to improve the automated detection of “problematic cells” in data sets

• human editing staff to focus on validating “important”, or in some sense “large” records
Value added? (editing 2)

ML may offer a new – not rule based – perspective on editing. This would help

• to detect “problematic cells” which can hardly get found by intuition or rules

• possibly to be faster in detecting “problematic cells”

• to use not only logical but also statistical aspects in the editing process
Value added? (imputation)

• ML may improve prediction tasks within already existing imputation schemes. This would possibly lead to better imputation results.

(--> WP2: When is imputation “good”?).

• ML may be faster in doing imputation if a machine learning method could avoid calculating distances between data points in hot deck imputation.
Pilot Studies

Editing

- UK: LCF survey data
- Italy: Editing in the Italian register of the public administration

Imputation

- Poland: Participation of Polish residents in trips
- Belgium (VITO): Early estimates of energy balance statistics
- Italy: Attained level of education in base register of individuals
- Germany: Simulation study on ML methods for imputation

+ Australia (see tomorrow) and Switzerland (talk by Christian Ruiz 02/2020)
How we’ve worked so far

• Meetings in person at the two sprints (some of us)

• Skyping (editing subgroup)

• Emails (mostly bi- or trilaterally)

• Preparing and reading/listening/commenting presentations and papers
Output (roughly estimated)

- 5 (+1) presentations (with additional background papers) for the Belgrade sprint

- one contribution per pilot study for the report at the end of 2019

- one draft for a final report per pilot study

- 6 (+1) presentations for this virtual sprint

- several presentations (already held or in preparation) in NSOs, at workshops and conferences

- connecting people, sharing ideas, transferring knowledge