Statistics Finland

Recommended Practices Manual EDIMBUS

BiH, Sarajevo

Pekka Tamminen

Statistics Finland

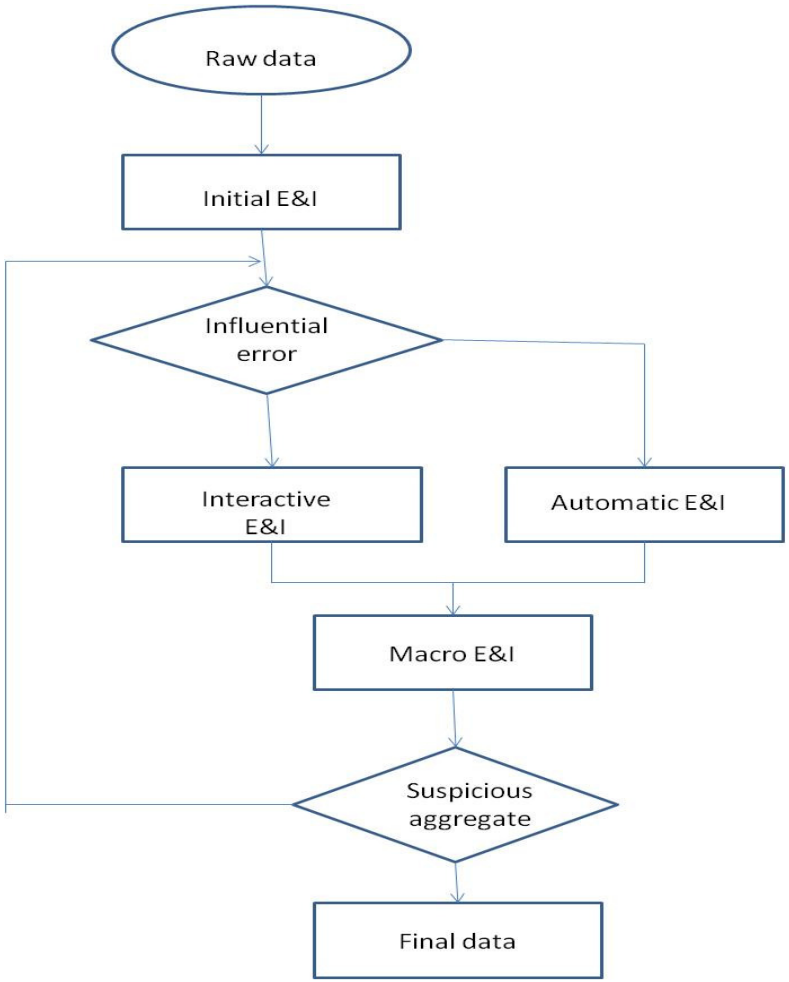
Contents of EDIMBUS RPM

- General Framework
- Designing and tuning E&I
- Detection of errors
- Treatment of errors
- Subsequent analysis and estimation
- Documenting E&I
- Summary

- Defines concepts
- Presents methods that can be used
- Gives practical recommendations on many levels

Statistics Finland

General flow of E&I process



Statistics Finland

Editing and imputation practices

- Detection of errors, detection of influential errors
  - Edit rules, hard and soft edits
  - Selective editing
  - Macroediting
  - Outlier detection
- Treatment of errors
  - Rule based imputation
  - Deductive imputation
  - Model based imputation
  - Donor-based imputation
  - Interactive treatment

Statistics Finland

Rule-based imputation

- The values will be imputed are determined by rules based on the values of the other fields and/or the erroreous values to replaced
- For imputing obvious systematic errors
- Usually based on IF-THEN rules and is often not separated from the error localization procedure
- Example: If number of employees = 0 and worked hours > 0 then
  - Imputation -> worked hours=0.

Statistics Finland

Rule-based imputation

- Rule based imputation is approppriate when, in the presence of systematic errors whose nature is known, the imputation action is quite obvious
- For instance in the case of systematic unity measure error
- Can be used for both cathegorical and numerical variables
- Simple to implement
- Allows to recover the true value when the error source is easy to identify
- Can lead to a severe bias in the estimates if the error cannot be identified with cernatinty in all cases
- It’s generally difficult to set up a set of rules that ensures the consistency of the imputed data with respect to a large set of edits

Statistics Finland

Deductive imputation

- Performed when, givening specific values of other fields, and based on a logical or mathematical reasoning, a unique set of values exists causing the imputed satisfy all the edits
- For instance when items must sum up to a total
- Can be used in any context
- Useful when for several observations the constraits lead to a unique set of values, values that allow the record to pass the edits
- Typical in edits of profit and loss account and balance

Statistics Finland

Deductive imputation

- Simplest and cheapest method of imputation (just like rule-based)
- Is often viewed as reliable method because the result is deterministically defined at unit level and based on logical reasoning
- Leads to true values with certainty if errors in the data has been perfectly localized
- Does not preceive the consistency between the variables

Statistics Finland

Model based imputation

- The predictions of missing values are derived from explicit models
- An imputation model predicts a missing value using a function of some auxiliary variables
- The auxiliary variables are typically from the sampling frame (size class, branch of activity etc), historical information and administrative data
- Most common types of model based imputation is regression imputation , ratio imputation and mean imputation
- For cathegorical variables predictions usually results from logistic or log-linear model

Statistics Finland

Donor based imputation

- Donor based imputation can handle variables that are difficult to treat by explicit modelling
- Under certain conditions donor-based imputation can preserve population distribution
- The consistency of the imputed observations with respect to edit rules is generally not ensured
- Concistent data can be enforced by adjusting the imputed values by a separate algorithm or by restrict the donor pool to donors which result in concistent imputation

Statistics Finland

Donor based imputation

- The way to choose the donor differs among several types of donor imputation
- For instance:
  - Random donor imputation: The donor is chosen randomly from the donor pool
  - Nearest neighbour imputation: The donor is chosen in such a way that some measure of distance between the donor and the recipient is minimized
- Random donor imputation is usually performed inside imputation cells
  - Imputation cells: grouping (”stratifying”)units by auxiliary information
- A substantial number of donors in the donor pool is needed to ensure good performance