Statistics Finland

Automatic editing and imputation

Activity 2.10 Usage of administrative sources

Ville Tolkki, Statistics Finland

Statistics Finland

Topics

- Needs, purpose and categories for editing and imputation
- Fellegi & Holt: Systematic Approach to Automatic Edit and Imputation
- Outlier detection
- BANFF at Statistics Canada
- Editing and imputation system at Statistics Finland SBS
  - manual editing
  - automated editing
  - imputation methods
- Conclusions

Statistics Finland

What is data editing?

Statistics Finland

What is data editing?

- Data editing refers to activities by which the statistical data are checked and made as correct as possible with respect to both individual values and mutual compatibility between the values for different variables.
- All errors are not corrected

Statistics Finland

What is data editing?

- Needed at each phase, starting from the planning of the data collection all the way to data file formation and data processing and analysis
- Problems occur in every survey and the whole data must be examined with care in order to avoid significant distortions to the survey results!
- Data editing can consume up to 40 per cent of the total resources spent on a survey, particularly for business surveys (Granquist 1977)

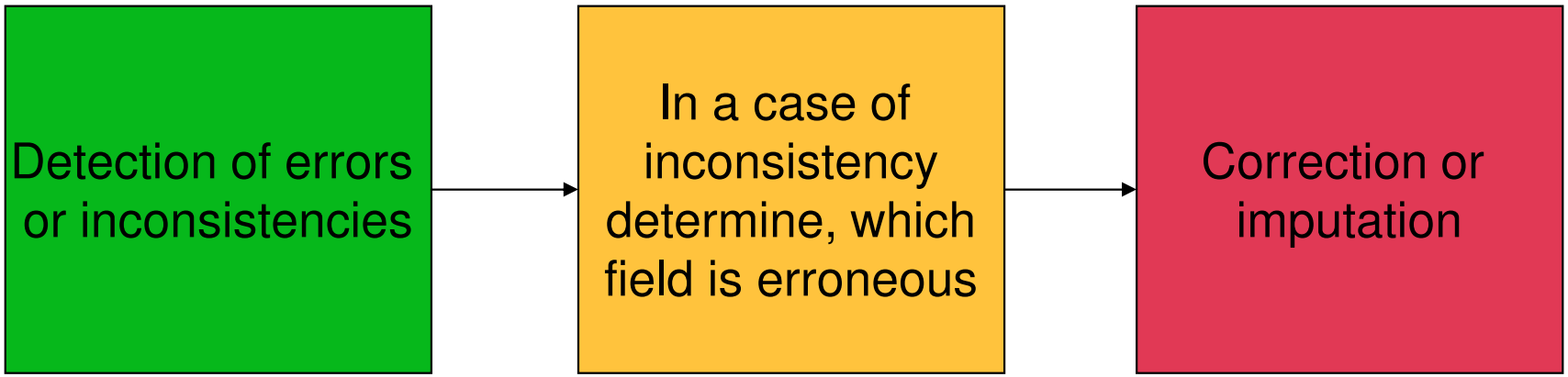
Statistics Finland

Imputation

- = *statistical replacement of missing values*
- usually for correcting item non-response, but can also be applied at the unit response level
- Imputation is an estimation tool
- *Mass imputation* refers to large-scale automated imputation situations

Statistics Finland

Relation of editing and imputation



Statistics Finland

Needs for editing and imputation

- Missing values
  - imputation is used to replace these values
- Inappropriate values
- Inconsistent values
  - implies two or more values
- Distinction should be made between the three concepts: *missing value, zero value and impossible value*

Statistics Finland

Purpose of editing

- The purpose is threefold (Granquist, 1994):

  (i) it creates the foundation for improving of statistal survey data in the future,

  (ii) produces information about the quality of statistical survey data,

  (iii) cleans up the data.
- errors and defects should be analysed for their overall importance

Statistics Finland

Editing categories

- Completeness edits
- Validity and range edits (where only certain codes or ranges of values are permissible)
- Consistency edits (comparison of different answers from the same record to check logical consistency)
- Historical edits (e.g. comparison of response for one survey with a previous response – ratios may be calculated and rules based on percentage variance)
- Statistical edits (checks based on statistical analysis of respondent data where suspicious values are identified - could include historical data)

Statistics Finland

Imputation categories

- There are many kinds of imputation methods, which can be divided into three main groups

1) Logical imputation (deductive) is part of the editing process. Used when *reliable*, explicit solution exists given appropriate assumptions (cf. deterministic imputation)

2) Model based imputation: a model fitted to the data; also unobserved values are possible

3) Real donor imputation: the imputed observation value is borrowed from another respondent

Statistics Finland

Imputation categories

- On the other hand, methods can be divided as below:

1) Deterministic imputation gives a same, unambiguous value when repeated (as in the mean imputation or in the logical imputation)

2) Stochastic imputation has a random element

Statistics Finland

Fellegi & Holt: Systematic Approach to Automatic Edit and Imputation

- Ideas presented in the article for *Journal of the American Statistical Association, March 1976, Volume 71, Number 353 Applications Section*
- Criteria:
  1) The data in each record should be made to satisfy all edits by changing the fewest possible items of data
  2) As far as possible the frequency structure of the data file should be maintained
  3) Imputation rules should be derived from the corresponding edit rules without explicit specification

Statistics Finland

Fellegi & Holt: Definition for editing

- Editing means:
  1) the checking of each field for every record to ascertain whether it contains a valid entry
     - not invalid blanks
     - codes are among valid boundaries
  2) the checking of entries in certain predetermined combination of fields to ascertain whether the entries are consistent with one another
     - are usually specified on the basis of extensive knowledge of the subject matter of the survey

Statistics Finland

Fellegi & Holt: Five theoretical options for correcting records that fails some of the edits

1 Check the original questionnaire (coding errors)

2 Contact the original respondent (to verify the response)

3 Manual editing: Clerical staff ‘correct’ the questionnaire using certain rules

4 Automated editing: Use the computer programs to ‘correct’ the questionnaire using certain rules

5 Drop all records that fail any of the edits - implies weighting

Weighting (5) is not as good procedure as explicit imputation according to the 1st criteria

Statistics Finland

Fellegi & Holt: Five theoretical options for correcting records that fails some of the edits

- 1 & 2 are ‘of course’ best accuracy but expensive, that is these should be used only when appropriate
- For the alternatives 3 & 4 the 4 is strongly recommended
  - data should be corrected using predefined rules
  - computer should be used for the correction rather than people

Statistics Finland

Fellegi & Holt: advantages and disadvantages using computer

- Advantages are
  - timeliness
  - consistency
  - *documentation (not in F-H article)*
- Disadvantages
  - complexity
  - rigidity
  - *development is time consuming (not in F-H article)*

Statistics Finland

Fellegi & Holt: Philosophy

- Edits can be presented as linear equations
  - defined by subject matter specialists
- Define the complete set of edits
- Search the minimum set of fields to be changed to pass all edits
  - starting with one field
- Change the values for identified fields

Statistics Finland

Fellegi & Holt: a trivial example

- Fellegi & Holt (1976): an erroneous record should be made to satisfy all edits by changing the values of the fewest possible number of variables
- implicit edits are logically implied by the explicitly specified edits
- implicit edits can be defined for numerical as well as categorical data
- implicit edits sometimes allow one to see relations between variables more clearly

Statistics Finland

Fellegi & Holt: a trivial example

- Suppose we have four *numerical* variables x_i (i = 1,...,4). The explicit edits are given by:

x_1-x_2+x_3+x_4 >= 0

and

-x_1 + 2x_2 -3x_3 >= 0

The implicit edits are given by:

x_2-2x_3+x_4 >= 0,

x_1-x_3+2x_4 >= 0

and

2x_1-x_2+3x_4 >= 0

Statistics Finland

Fellegi & Holt: a trivial example

- **The explicit edits: x_1-x_2+x_3+x_4 >= 0 and -x_1 + 2x_2 -3x_3 >= 0**
  The implicit edits: x_2-2x_3+x_4 >= 0, x_1-x_3+2x_4 >= 0 and 2x_1-x_2+3x_4 >= 0.
- Suppose we are editing a record with values

(3, 4, 6, 1) (with reliability weights all equal to 1).

I) The first edit is satisfied, the second edit is violated:

-3-4+6+1 >= 0 but -3+8-18 <= 0.

II) Which of the fields should be changed?

III) We see that two implicit edits will fail and that variable x_3 occurs in all three violated edits.

Statistics Finland

Fellegi & Holt: a trivial example

- So we can satisfy all edits by changing the value of x_3 , for example, x_3 could be made equal to 1.
- Changing the value of x_3 is the only optimal solution for this error localisation problem.
  - imputation(!)

Statistics Finland

Outlier detection: ”current method”

- Calculate the first quartile, Q1, the median, M, and the third quartile, Q3, of the variable in question.
- Calculate the distances d_{Q1} and d_{Q3} as given below. They are normally the distances from the median to the first and third quartiles.
  - d_{Q1} = Max (M - Q1, |A*M|),

    d_{Q3} = Max (Q3 - M, |A*M|).
- Impute (or exlude..) if
  - $x_i < M - C*d_{Q1}$ or $x_i > M + C*d_{Q3}$.
  - A and C are user-specified parameters respectively.

Statistics Finland

Outlier detection: ”current method”

- Of course: the well-known alternative is to use means and standard deviations instead of medians and quartiles:
  - IF x_i > AVG + 3*SD or x_i < AVG - 3*SD

    THEN x_i is an outlier - or is it?

  etc...

Statistics Finland

Outlier detection: ratio method

- For each record in which $x_i > 0$ and $y_i > 0$, calculate the ratio $r_i = x_i / y_i$, where Y is the appropriate auxiliary variable.
- Otherwise calculations are similar to those done in the Current Method, but these are performed to *transformed values* due to use of ratios.
- Transformations:

  if $0 < r_i < r_M$ then $s_i = 1 - r_M / r_i$

  if $r_i > r_M$ then $s_i = r_i / r_M - 1$,

  where r_M = median of the ratios.

Statistics Finland

Outlier detection: ratio method

- For each record in which $x_i > 0$ and $y_i > 0$, calculate the ratio $r_i = x_i / y_i$, where Y is the appropriate auxiliary variable.
- Otherwise calculations are similar to those done in the Current Method, but these are performed to *transformed values* due to use of ratios.
- Transformations:

  if $0 < r_i < r_M$ then $s_i = 1 - r_M / r_i$

  if $r_i > r_M$ then $s_i = r_i / r_M - 1$,

  where r_M = median of the ratios.

r_1 = 50 / 140 = 0.36, while
r_2 = 140 / 50 = 2.80.
Assume $r_M = 1$ and $s_M = 0$, then
s1 = 1-1/0.36 = -1.8, and
s2 = 2.80/1-1 = 1.8.

Statistics Finland

Outlier detection: ratio method

- **Transformations:**

  if $0 < r_i < r_M$ then $s_i = 1 - r_M/r_i$

  if $r_i > r_M$ then $s_i = r_i / r_M$,

  where r_M = median of the ratios.

- To give greater importance to small deviations of large units, one can calculate the effect:
  - $e_i = s_i [\ \max (x_i , y_i)\ ]^{\mathbf{exp}}$,

  where exp is between 0 and 1,

  and do the calculations (M-Q1, Q3-M, etc...) and the comparisons using e_i instead of s_i .

Statistics Finland

Outlier detection: historical trend method

- Similar to the ratio method but r_i is defined as

  $r_i = x_{it} / x_{i(t-1)}$.

- So, the ratios are calculated between the current, t , value and the corresponding historical, t -1, value of the same variable.

Statistics Finland

Handling outliers

- *Robust methods* minimize effects of outliers.
- *Trimming:* removing the *n* members having the *n*/2 largest values and the *n*/2 smallest values of a given parameter. *The trimmed mean* is the mean value ignoring the *n* extreme values.
- *Winsorization:* the extreme values are moved toward the centre of the distribution, e.g. by replacing the *n* extreme values by the two remaining extreme values.
- In the sample data, the outlier value can be weighted by a weight related to appropriate auxiliary information.

Statistics Finland

Editing systems
BANFF of Statistics Canada

- a collection of specialised SAS procedures developed at Statistics Canada
- derived from the Generalised Edit and Imputation System (GEIS)

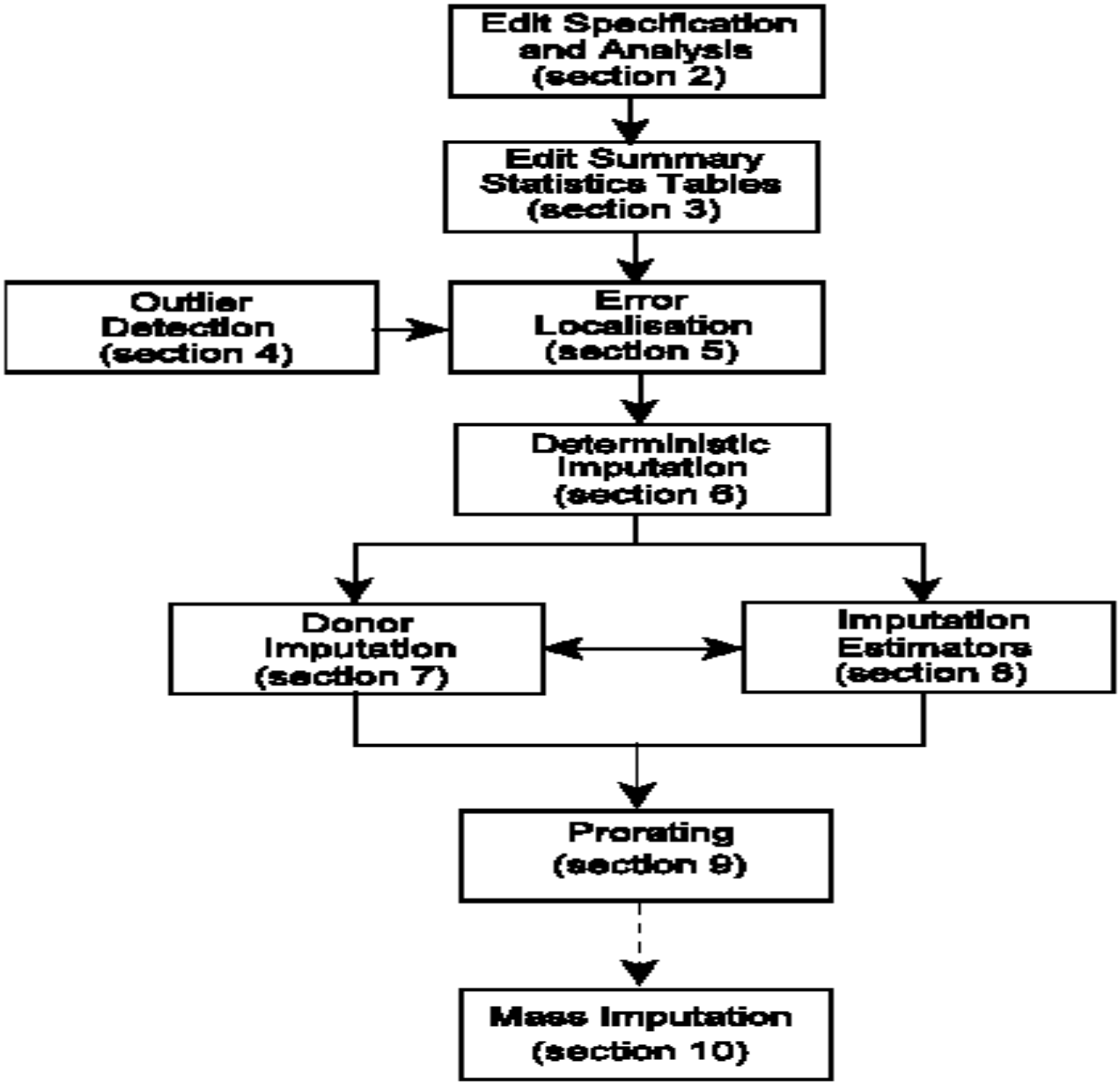
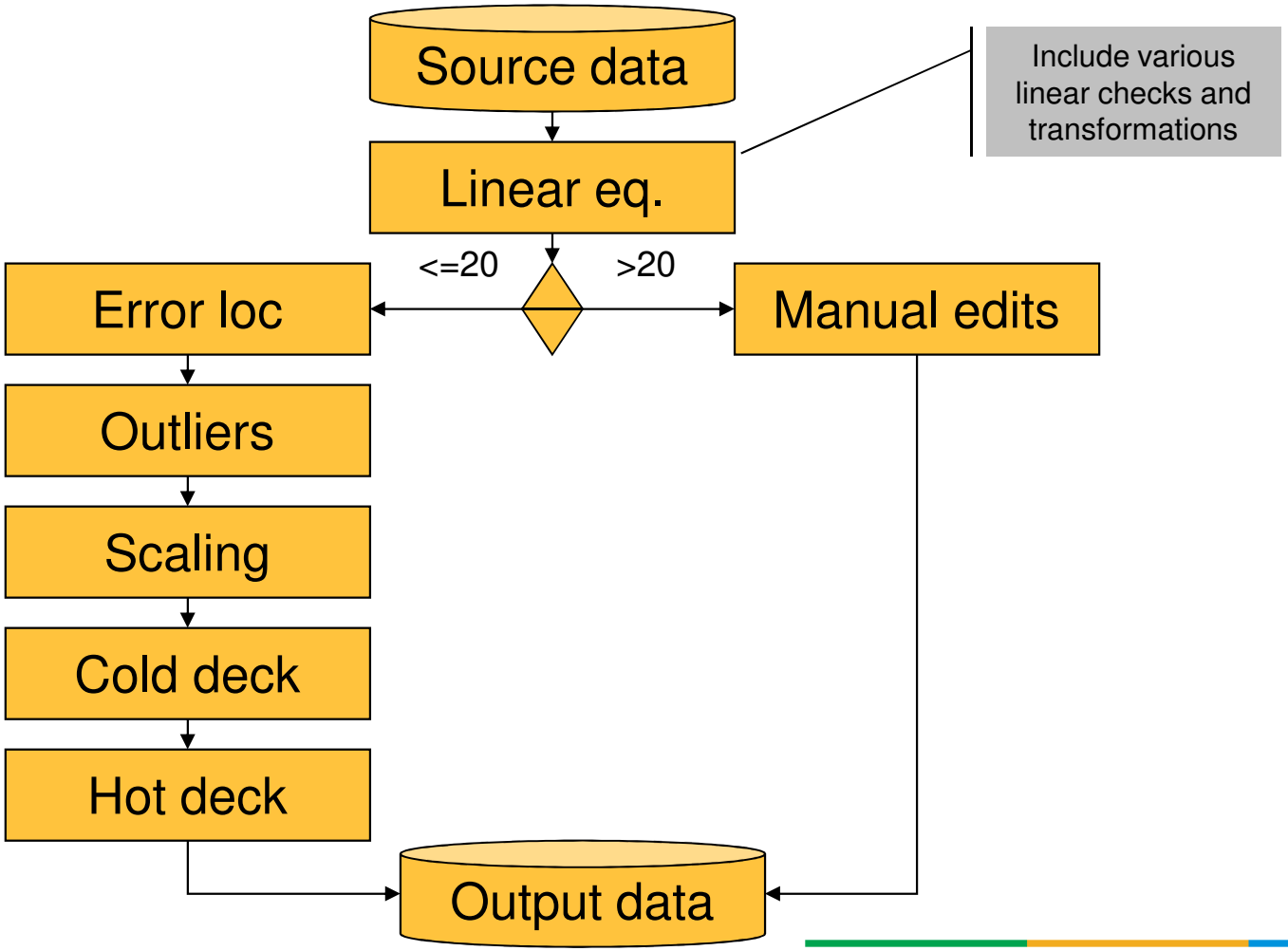


Figure 1.1 The logical order of the Banff procedures in the context of survey processing. Section numbers refer to sections of the document.

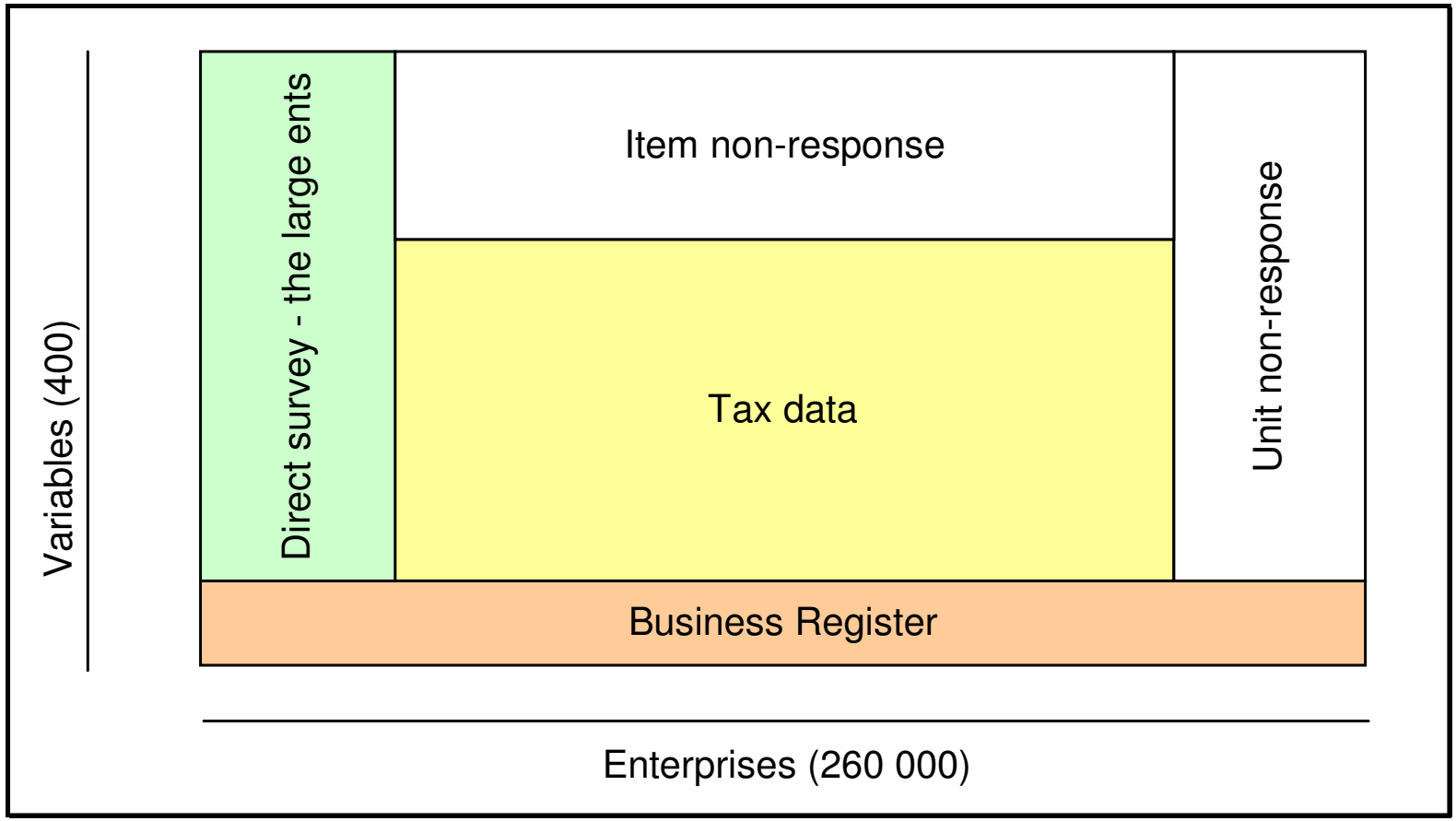
Statistics Finland

Editing process for Structural Business Statistics at Statistics Finland



Statistics Finland

Structural Business Statistics Data Sources and Methods at Statistics Finland



Statistics Finland

Conclusions

- Develop the edits with co-operation of IT and branch stat
- First apply the logical edits with outlier detection and automated corrections
- Do not apply manual edits for the small firms
- Manual edits are only rational when one needs to compare the data with the questionnaire or ask directly from the respondent
- Flag all edited and imputed values (also manual)
- Save all versions of the data
- Document your editing and imputation system