Data Validation and Processing

Mariet Tetty Nuryetty
mariet@bps.go.id

Session 8

Population and Housing Censuses; Registers of Population, Dwelling, and Buildings

Brunei, 22-24 August 2017

Outline

1. Content

2. Editing

3. Use of multiple sources

Data Processing

- Registers data processing can be more complex compared with a traditional census, although good quality results can be obtained.
- The use of administrative sources in statistical production process requires a close collaboration between the administrative authorities and the national statistical offices.
- During data preparations, all parties must agree on date(s) of delivery and the content of the data. This implies a bilateral agreement at a high hierarchical level on a detailed data set description, scheduled delivery dates, and the statistical reference period.
- Validation techniques appropriate to administrative data should be applied, including checks on the completeness, reference periods, and plausibility.

Data Processing (2)

- Register information may contain errors (e.g. records showing people as being implausibly old, invalid occupations, information about migration that is not consistent with other data). Edit rules may be defined to highlight inconsistent or implausible information.
- Correction or imputation of records with errors can be attempted in different ways: first, if possible, using another data source (register) that also has information about that specific record and topic; or second, carrying out probabilistic imputation based on available information that is thought to be reliable.

1. Content

Administrative registers must contain data covering the most important subject areas (many variables relating to the relevant units) in a statistical system for elucidating patterns and trends in society.

Units and identifiers

Three central units are essential to the structuring of the statistics: persons, enterprises/establishments and dwellings.

Time references

- Time dimension plays a very important role in statistics, revealing patterns and trends in society, and in all areas it is necessary to be able to make comparisons over time.
- The most important is the dates of changes or events*:* "birth" and "death" of units or the real point in time of event took place →change in a data item, e.g. date of a removal.
- So, dates of events need *registration dates*, i.e. an indication of when the data value in question was entered in the register. The ideal situation is any item of information in the administrative register should be accompanied by two dates.

Cooperation with register-keepers

It is desirable for statisticians to exert a certain influence on data content, but it must not be forgotten that registers are kept for quite specific administrative purposes and that the task of the register keepers is to serve those purposes in the best way possible. They cannot therefore pay too much attention to demands that "only" serve statistical purposes. For this reason, statisticians should not expect to make very extensive demands for additional data, different definitions and the like, and to have those demands met.

1. Editing

By register owners

- Purpose: improve data for administrative purposes
- Focus on variables important for administrative use
- Variable included for statistical purposes only may not be controlled as carefully
  - Statisticians must have good knowledge of these editing procedures

What is edited and how . What is not edited

2. Editing (2)

Editing in NSO

- Focus on statistical use of data
- Each register first controlled by using internal logical checks
  - No external sources
- Discover system errors. Data does not meet formal requirements
  - "Blocks" of data missing or incorrect
  - May result in contacting register owners for new delivery of corrected data
- Isolated errors may be detected and corrected
  - Possibility of contacting individuals to collect information very limited
- Main method: computerised logical editing

2. Editing (3)

Editing in NSO (2)

- Time references must be processed to obtain information for desired point in time or periods
  - Dates of changes or events
  - Registration dates
  - Distinguish real events from corrections

- Close contact with register owners important
  - Systematic errors detected should be corrected in the administrative registers
  - Beneficial for both parties

3. Use of multiple sources

- Administrative data must be processed to meet statistical needs
- Creation of statistical registers often involves linkage of data on an individual level
- Find best estimator for statistical variables
- The solution may be derived variables

3. Use of multiple sources

- Administrative data must be processed to meet statistical needs
- Creation of statistical registers often involves linkage of data on an individual level
- Find best estimator for statistical variables
- The solution may be derived variables

Derived variables

- Creation of new variables
  - Example: Current activity status
- Prioritisation of information for a given variable
  - Depending on quality and coverage
- Adjustment or correction of information
  - Obtain maximum consistency

Use of combined data over time

- Often requires a complex set of rules
- Detailed documentation necessary
- Allow for change over time
- Vulnerable to administrative changes
  - For instance a variable may be omitted in a register or changed in a major way
- Even improvement of data may give rise to discontinuity

Updating

Key principle for registers

- Need to process data only when units or attributes are changed
- No changes: No processing

References

1. Tonder, Johan Kristian, 2008. **The Register-based Statistical System Preconditions and Processes. Paper presented in the** International Association for Official Statistics Conference, October 14 – 18, 2008, Shanghai.
2. United Nations (2011): Using Administrative and Secondary Sources for Official Statistics: A Handbook of Principles and Practices.

Thank You