

Dataset Preparation

Mariet Tetty Nuryetty
mariet@bps.go.id

Session 6

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

Brunei, 22–24 August 2017

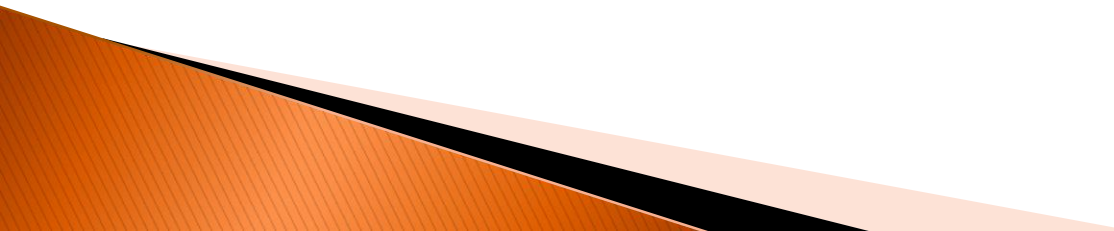
Content

1. Record
Linkage

2. How to do it?

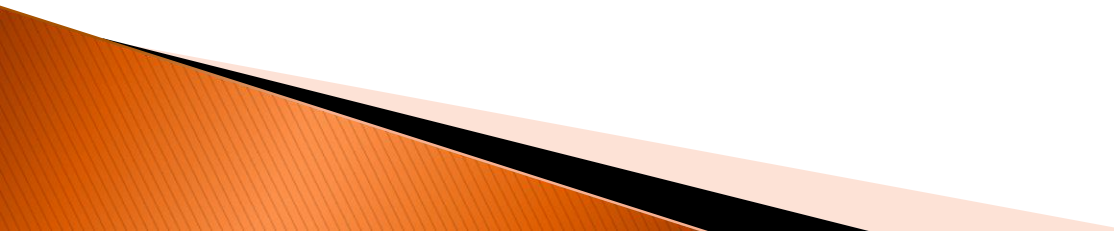
1. Record Linkage

As a rule only **one linkage** variable at a time is used when two or more sources are combined – an identity number or an address code.



Need for Linkage between multiple sources

Data are organized into *statistical registers*, which are data compilations with identification keys in which the population and the data content of each register are correlated.



Links are created based on

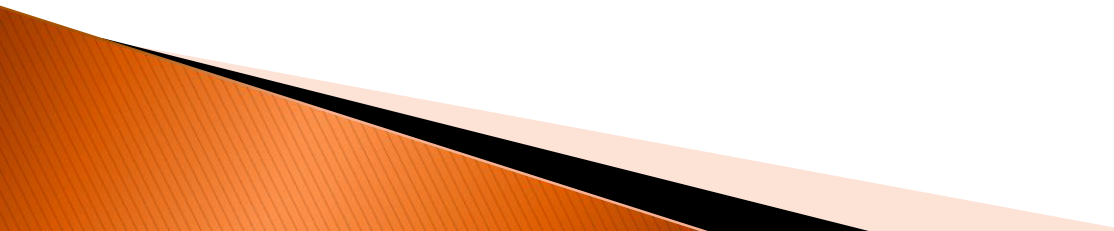
Relation :

1. Between persons, enterprise and local area

For instance: person works/studies at an enterprise/organization located in a local area

2. Between person and property/dwelling
 3. Between local area and property
 4. Between person/enterprise and vehicle
- 

How to do it?

- A link consists of “*one or several*” common *linkage variables*.
 - *The link* contain the information to *identify relations* between different types of units.
 - This takes a form of *matching (Exact or Probabilistic)*
- 

Exact matching

Information between administrative sources is match (could be using a *common identifier*)

Probabilistic matching

- No *common identifiers* exist
- *Common identifiers have* poor quality
- Could be solved by using common variables (name, address, DOB, occupation) to sources
→ *Matching keys*
- Matching keys may be derived.

Probabilistic matching (2)

Choice of variables to be used for matching should take into account **the distinguishing power** (*uniqueness of the values of the matching key*).

- High distinguishing power: *reference number, full name, full address*
- Low distinguishing power: *sex, age, city, nationality*

Probabilistic matching (3)

Basic matching techniques:

1. Match: $A=A \rightarrow$ same entity in reality
2. Non-Match: $A \neq B \rightarrow$ two different entities in reality
3. Possible match: $A=a? \rightarrow$ no enough information to determine match or non-match
4. False match: $A=B \rightarrow$ a pair wrongly designated as match
5. False Non-match: $A \neq A \rightarrow$ match in reality, but is designated as a non match

Matching techniques

Clerical

Requires significant human input



- ☐ Expensive
- ☐ Inconsistent
- ☐ Slow
- ☐ But; Intelligent

Automatic

minimize human intervention



- ☐ Cheap
- ☐ Consistent
- ☐ Quick
- ☐ But; of limited intelligence

Best is a mixed method with minimizing clerical

Step of automatic matching

1. Standardization → used for text variables
 - Spell out abbreviations (e.g. “mfg” → manufacturing , "ltd" → “limited”,)
 - Standardize common variations of names (*of places, persons*). E.g. "Brussel" /"Bruxelles", "John"/"Johny")
 - Remove “noise” words (*street, road*). E.g. "road"/"stree" in addresses.
 - Postal code, DOB, etc. E.g. "# 22 January 2008"/"220108".

This might reduce data quality or even chance of matching.
The best is keeping the original variables.

Maximizing automatic matching (2)

2. Parsing

text converted from a form recognisable by human to more logical for computer processing.

For instance:

- Letters with similar sounds to a common string (*“f”, “v”, and “ph” to “f”*)
- Remove silent letters (*“h” from “Johnson”*)
- Vowels to single character (*“ee” in “steel”*)
- Remove vowels from the end (*“ea” from “Andrea”*)
- Replace double letters to single (*“nn” in “Anna”*)

Maximizing automatic matching (3)

3. Blocking

break the large files into smaller "blocks" to save processing time

- can improve the cost-efficiency of the matching process.
- E.g. if the record has an address in a certain town, match it against the block containing other records from that town rather than all records for a whole country.

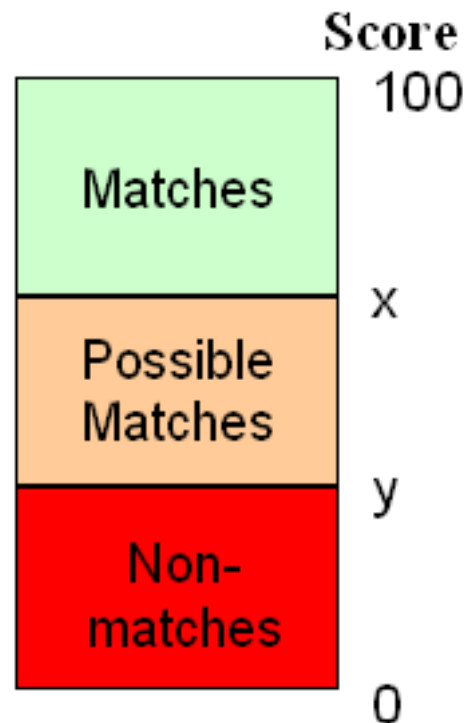
Maximizing automatic matching

(4)

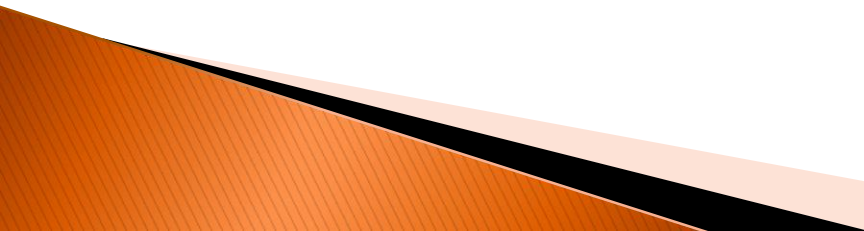
4. Scoring

Assess the likelihood for matching

- whether a pair of records is considered to be a **definite match**, a **possible match** or a **non-match**.



References

1. Session 6b: From administrative data to register-based statistics; Data Linkage and Matching. Presented in the Regional Training on Producing Register-based Population Statistics in Developing Countries, 27–31 October 2013.
 2. United Nations (2011): Using Administrative and Secondary Sources for Official Statistics: A Handbook of Principles and Practices.
 3. Wallgren, A. And Wallgren, B. (2007), Register-based Statistics: Administrative Data for Statistical Purposes, John Willey & Sons, Ltd, Chichester, UK.
- 

Thank you