DATA LINKAGE

Linking Population Census data with Post-Enumeration survey data

MURENZI Ivan
Deputy Director General – National Institute of Statistics of Rwanda

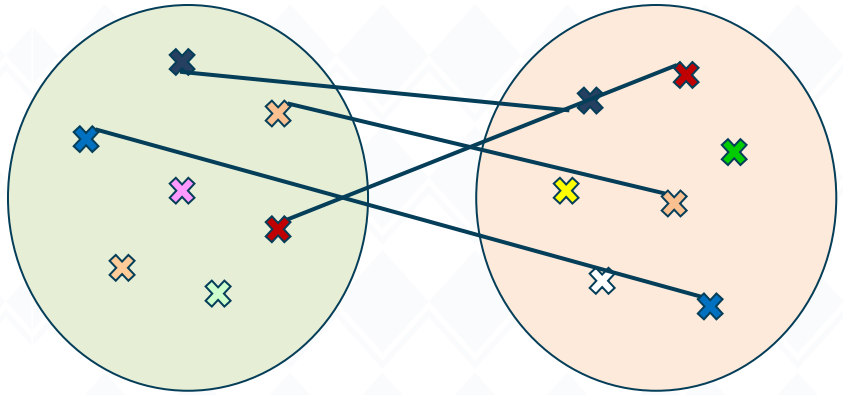
Outlines:

1. What & Why Data Linkage
2. Description of the project
3. Steps of Data linkage
4. Outcomes and next steps

1. DATA LINKAGE

WHAT?

Data linkage is the process of trying to establish whether two records from two different datasets relate to the same entity



WHY?

1. **Time:** quicker than collecting new data
2. **Cost effectiveness:** makes better use of existing data
3. **Improved data quality:** linkage process may identify quality problems in data like duplicates
4. Opens new **research opportunities**
5. Optimize **the use of admin data**

2. Description of the Project

- Census was conducted in 15th-30th Aug 2022 while Post-Enumeration Survey (PES) was conducted one month after, 15th-30th Sept 2022.

- The purpose of the PES is to measure the accuracy of the Census by independently surveying a sample of the population.

- Data Linkage was used to match Census & PES records using Python programming language.

- This was done at Household level, Enumeration Area level, District Level and Country level.

3. STEPS IN DATA LINKAGE

(i). Data Validation and Cleaning

The process of data validation and cleaning ensures that the inconsistencies in the data are identified before the data is used in the analytics process

(ii). Initial Direct Matching

Begin by matching records based on exact matches

(iii). Fuzzy Matching Algorithms

For records that are still unmatched, focusing on fields with textual data (like names and addresses) that might have variations in spelling or format.

(iv). Probabilistic Record Linkage

Use in the remaining records to estimate the likelihood of matches where the data isn't as clear-cut

(vi). Manual Review

Given the complexity of matching, it's important to incorporate a step for manual review

1. Data Validation and Cleaning

This involves the use methods such as logical consistency checks, and data type verification.

Name	dictrict	HH ID
Muhazi Textiles	54	0116
Kivu%-_%Electronics	54	0119
Nyungwe Suppliers	12	2399
Nyungwe Suppliers	12	2399
Butare Pharma	11	4510
Gisenyi Constructions	11	4520
Kigali Retailers	13	4732AB

Name	dictrict	HH ID
Muhazi Textiles	54	0116
Kivu Electronics	54	0119
Nyungwe Suppliers	12	2399
Butare Pharma	11	4510
Gisenyi Constructions	11	4520
Kigali Retailers	13	4732

2. Initial Direct Matching

Initiate the process of matching data by systematically aligning and matching HHs using a core set of common identifiers.

Identification: Select key variables
(location, names, age, sex, name of the head of household, marital status, Date of birth,)

Alignment: Standardize these variables across datasets for uniformity.

Initial Match: Match records with exact matches across all selected variables.

Result: A preliminary matched dataset of economic units with consistent attributes.

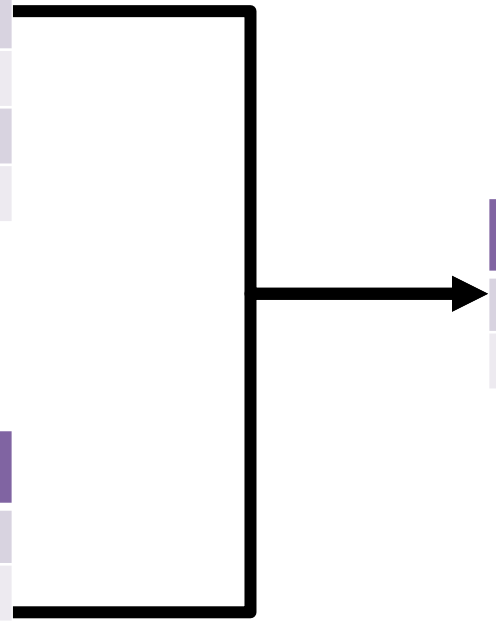
2. Initial Direct Matching

Census data

Name	dictrict	HH ID
Muhazi Textiles	54	0116
Kivu Electronics	54	0119
Nyungwe Suppliers	12	2399
Butare Pharma	11	4510
Gisenyi Constructions	11	4520
Kigali Retailers	13	4732

PES

Name	dictrict	HH ID
Nyungwe Suppliers	12	2399
Butare Pharma	11	4510
Kigali Retailers	13	4732
Gisenyi Crafts	11	4520
Akagera Electronics	12	2399



Matches

Name	dictrict	HH ID
Nyungwe Suppliers	12	2399
Butare Pharma	11	4510

3. Fuzzy matching algorithms

Fuzzy matching algorithms are used to identify matches that are not exact but close enough to be considered a potential match. These algorithms are useful to handle variations in data entry, spelling errors, and linguistic differences.

Soundex: Encodes words based on their pronunciation

Metaphone: More advanced than Soundex, handles a wider range of phonetic variations.

Code	Character
0	a e h i o u w y
1	b p
2	c g j k q
3	d t
4	l
5	m n
6	r
7	f v
8	s x z

4. Probabilistic Record Linkage

Probabilistic record linkage is a method to statistically determine the likelihood that two entries from different data sources refer to the same entity. It's particularly useful when no unique identifier is available.

5. Manual review

Manual or Clerical matching is based on human judgement. A small team of clerical matchers used to resolve hard cases.

4. Outcome and next steps

- **Less Time**: it only took three weeks, compared to 6 months this same process took in 2012,
- **Low cost**: few people (14 people) were part of the process of matching
- **Data Science Skills**: developed a matching module. Built skills will be applied in other linkage projects
- **Reproducibility**: Developed a functional, well structured, reproducible and accurate matching algorithm that can be used elsewhere e.g track the life of establishments over years


Thank You!
alt