

APPLICATION OF DATA LINKAGE AT NATIONAL INSTITUTE OF STATISTICS OF RWANDA (Post Enumeration Survey)

UWIMBABAZI Denyse

Data Scientist, NISR

Email: denyse.uwimbabazi@statistics.gov.rw



www.statistics.gov.rw



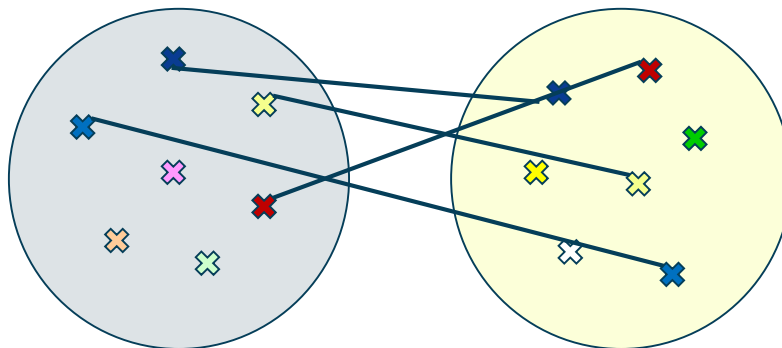
info@statistics.gov.rw






+250 788 383103

WHAT IS DATA LINKAGE?

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



WHY DATA LINKAGE?

-  **Time efficiency:** quicker than collecting new data
-  **Cost effectiveness:** makes better use of existing data without need of new data collection
-  **Improved data quality:** linkage process may identify quality problems in data like duplicates

STAGES OF DATA LINKAGE

01

Pre-linkage

Data pre-processing
such as Cleaning and
editing data, parsing,
choosing matching
variables

02

Linkage

Bringing datasets together
to identify records that
belong to the same person

03

Post-linkage

Examining matching
process, estimating error
rates, carrying out
analysis

POST ENUMERATION SURVEY (PES)

WHAT IS POST ENUMERATION SURVEY?

- PES is a survey carried out immediately after census where a large and enough sample size is drawn to measure the coverage (under/over coverage) of Census.
- It can be used to verify accuracy of responses for some key census questions.
- Data Linkage was used for PES 2022 to measure the coverage of the 5th Rwanda Population and Housing Census (5th RPHC).

WHAT IS POST ENUMERATION SURVEY? (Cont'd)

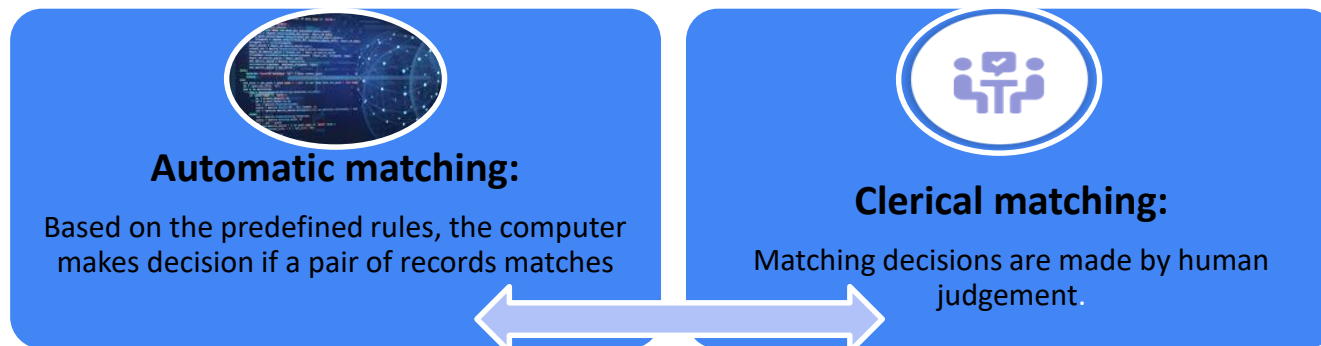
- Data collection for 5th RPHC started on August 15th, 2022 and for PES on September 15th 2022
- They were both conducted for a 15 days period preceded by Households listing.
- Tablets were used for Data Collection. This made the task of data linkage (matching of PES to Census data) faster than in 2012.

HOW DATA LINKAGE WAS USED IN PES 2022

- Data linkage was done using a matching algorithm developed by NISR with the support of ONS-UK
- The algorithm was built to find matching records and/or similarities from Census and PES data in all sampled areas at different stages.
- Standardized Levenshtein and Soundex edit were used to account for errors in names

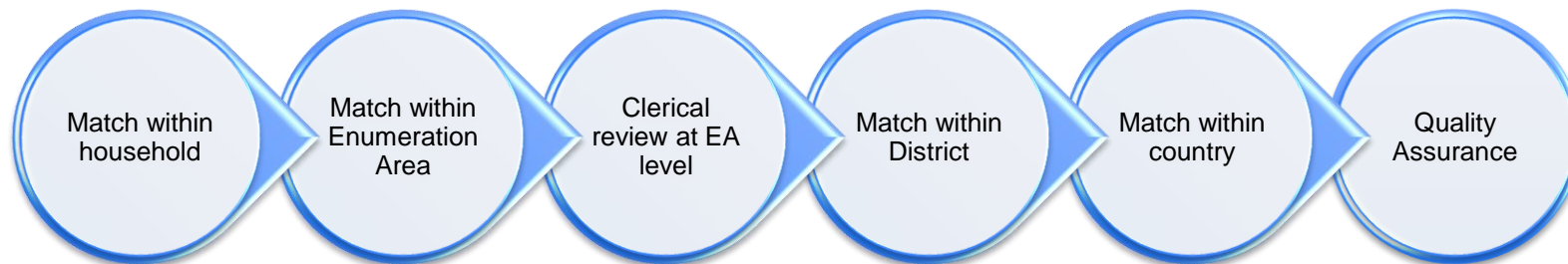
TYPES OF MATCHING

Two types of matching were used:



Small team of clerical matchers used to resolve hard cases, conflicting cases with Clerical Resolution Online Widget (CROW) system and to assure quality of the automatic matching.

STAGES OF MATCHING



- Example of match key: First name, last name, birth month and year, sex and location (HH_ID, EA_ID,...)
- CROW example:

Clerical Matching

Tools:

Current Record

	Datasource	First_Name	Middle_Name	Surname	Birth_Month	Birth_Year	Sex	District	Sector	Cellule
cen:		NADEGE		HAKIZA		1923.0	F	BUGESER	GASHOR	KAGOMASI
pes:		NADEGE		HAKIZIMANA	6.0	1924.0	F	BUGESER	GASHOR	KAGOMASI

Comment:

9 / 10 Records

QUALITY ASSURANCE

- The quality of the matching was checked by computing **Precision** and **Recall**
- These targets should be high compared to typical linkage quality targets (above 99.75%), as any linkage error directly impacts the quality of the census estimates.
- Results:
 - Net Coverage rate : 98.7%
 - Precision: 99.94%
 - Recall: 99.98%

KEY ACHIEVEMENTS



Less Time:

It only took three weeks, Most of matches were made automatically and this minimized the manual work (Previous PES took almost 6 months)



Low cost:

Few people were involved in the process of matching



Reproducibility:

Developed a functional, well structured, reproducible and accurate algorithm that can be used elsewhere



Enhanced Data Science Skills:

Developed a matching module



For more information on Rwanda PES 2022, visit: <https://statistics.gov.rw/publication/rphc5-thematic-report-post-enumeration-survey>