Toolkit for Data Linkage: Guidance on Interoperability and Linking Administrative Data with Other Data Types

UN Big Data Regional Hub for Africa

29/April 2025

Dr. Mathias KUEPIE
Independent consultant,
Demographic and socio-economic studies;
Population and geospatial data modelling
tel/whatsap: +33649398058
Emails:
kuepie@yahoo.com
kuepiemathias@gmail.com

INTRODUCTION
The Importance of reliable Data

- Essential for measuring progress towards the 17 SDGs and 169 targets
- Gender-disaggregated data highlights disparities
- Identifying and addressing gaps in marginalized sub-populations
- UNSD, UNECA, and UN Women working on mobilizing survey and administrative data

Limitations of Current Data Collection

- Surveys and censuses: periodicity and sample sizes
- Some variable are very difficult to collect through survey or census (ex:income) or collected with errors(profession, fertility, mortility etc.)
- Administrative not primary designed for statistical purpose
- Administrative data coverage rate some time partial

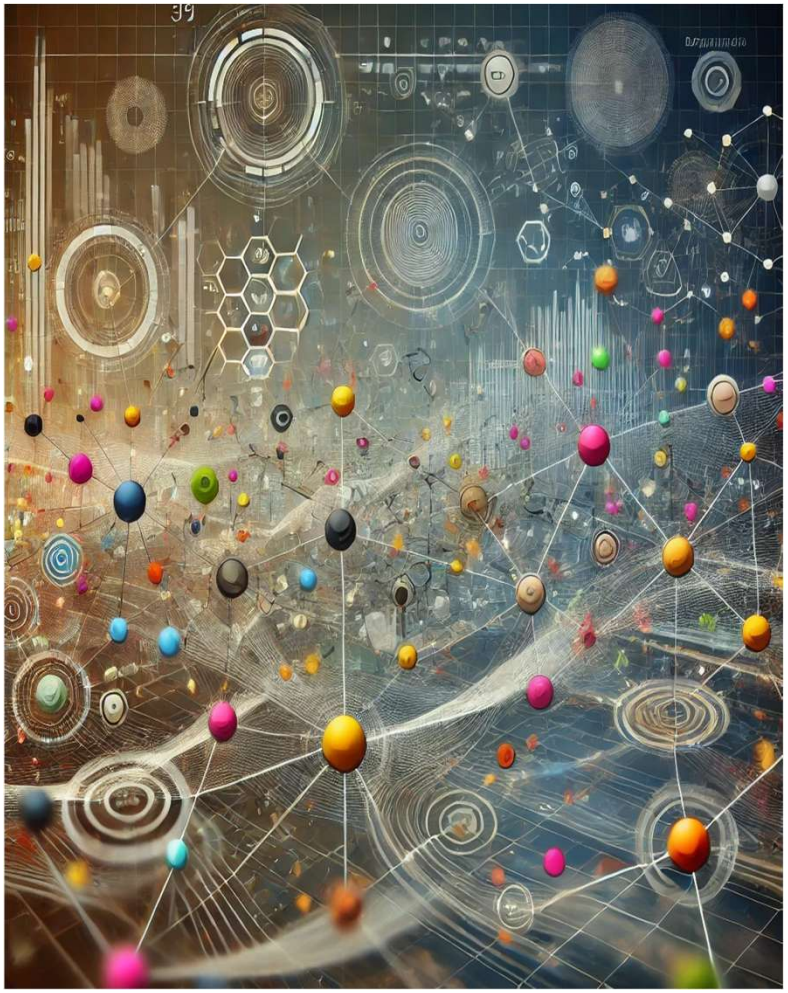
BUT

administrative data is permanently updated

- Collect some events at their occurrence(mortality, fertility) or more accurately(income)

Data linkage and integration

- There is a need of Integrating administrative and survey data.
- It’s is an ongoing challenge, particularly in Sub-Saharan Africa.
- The integration of these sources is crucial for effective SDG monitoring. UN-Women, UNECA, and UNSD have launched a project aimed at guiding data interoperability.



OBJECTIVE OF THE PROJECT

1. Analayze the conditions of making Administrative data available for statistical purpose

2. Individual record linkage of administrative data and survey data

   Deterministic linkage

   Probabilistic linkage

3. Geospatial individual to service linkage

4. Data integration by combining administrative and survey data to compute and monitor indicators

5. Linking survey data with aggregated administrative data

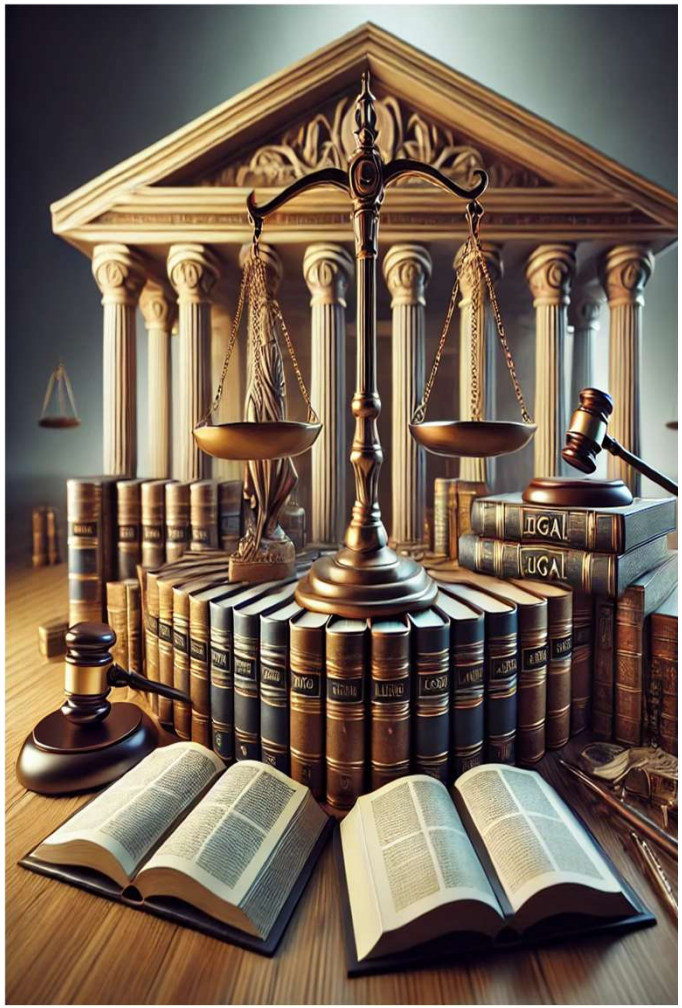
Analyze the conditions of making Administrative data available for statistical purpose

Legal Framework and Agreements

NSOs are custodians of surveys and census data, while administrative data is generated by various sectoral ministries (health, education, etc.).

National statistical laws define the roles and responsibilities for sharing data.

Memoranda of Understanding (MOUs) can formalize admin. data-sharing processes, with a focus on sharing microdata.



Key Considerations for Data Sharing

Data Security: Protect against breaches and misuse.

Anonymization: Remove personal identifiers after data linkage.

Access Control: Restrict access to authorized personnel.

Data Integrity: Ensure data quality through metadata and comparisons with surveys or censuses.

Example: Statistics Norway uses 'fake' unique IDs for data protection.

Data Coverage and Integrity

Coverage Issues: Depends on public service availability; reporting bias can affect data reliability.

Metadata: Essential for transparency about data collection and limitations.

Data Quality Tools: UNSD repository and Stanford’s handbook provide resources for improving data quality and integration with surveys.

Example: DHSI2 health data system integrates detailed health data while ensuring privacy and security.

WHAT IS MISSING IN MOST COUNTRIES

A Real availability of administrative data transform into a statistical database, hosted at the NSOs

Individual record linkage of administrative data and survey data

Objectives of Individual record linkage of administrative data and survey data

The objective of the individual matching of administrative and survey data can be very diverse. For example:

a) Add new variables to the survey data: (e.g.income and revenue)

b) Correct or impute existing variables in the survey (complete date of birth (day /month/year), revenue)

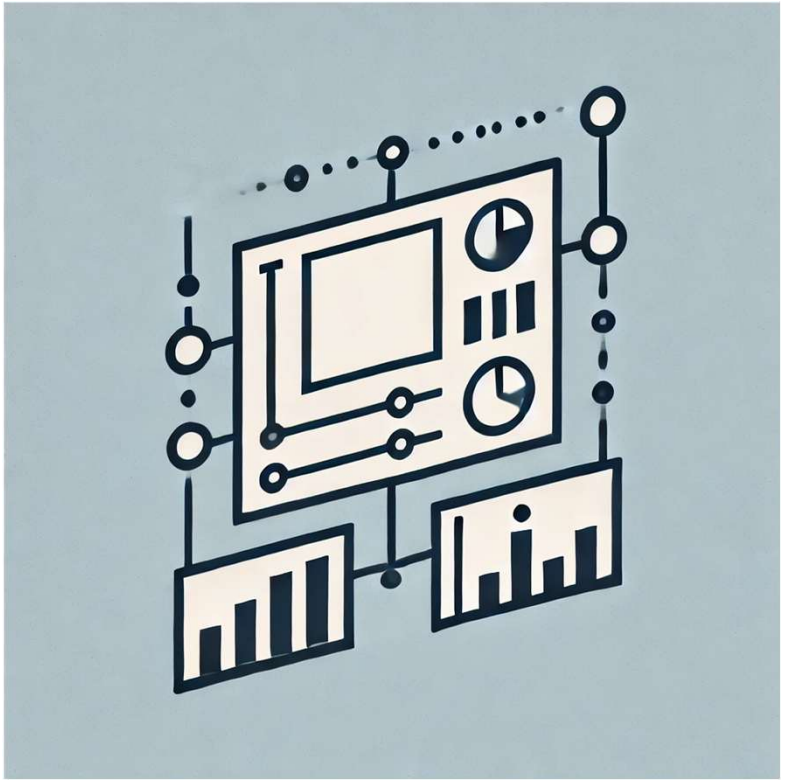
c) Including new records in the survey: generally during population census, some newborns can be omitted, especially if they died few months after birth. Using birth and death registers can help recover this information

d) Check the completeness of the Census

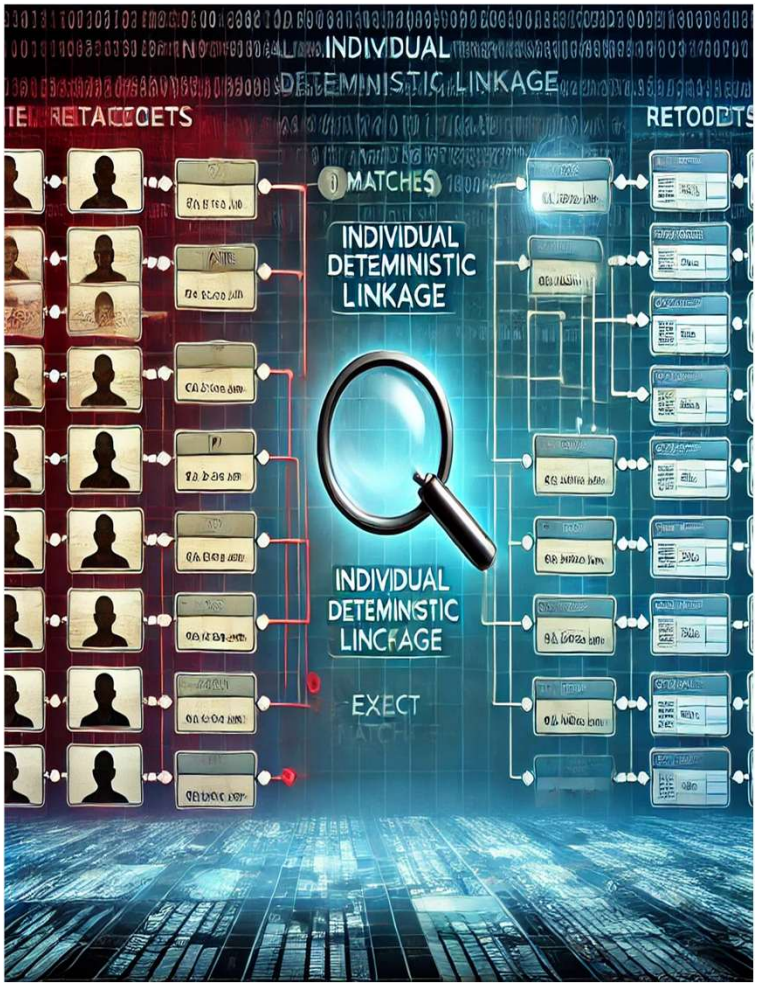
Statistical Pre-conditions to consider linkage

1- The units from the two databases belongs to the same universe

2- there are identifiers variables, that in the ideal scenario, would help pairing data on a one to one basis



Individual deterministic linkage



The example we used in the toolkit

Simulated admin. Data of 150.000 subjects with the following var: District, Full Name, Place of birth, Date of birth, Place of residence, Level of educ, occupation

We then sample **a survey data** of 10%

The Following errors were introduced randomly in the Survey Data for the purpose of this probabilistic matching exercise:

1- altered 10% of the survey on the place of residence by lowering the first letter

2- altered 10% of the survey on the name by lowering the first letter

3- Suppressed 10% of the second name

4- Discarded 10% of month and day of birth

Preparing data for linking

- Description of the content of the two data set
- Dimensions of data sets
- summary of dataset

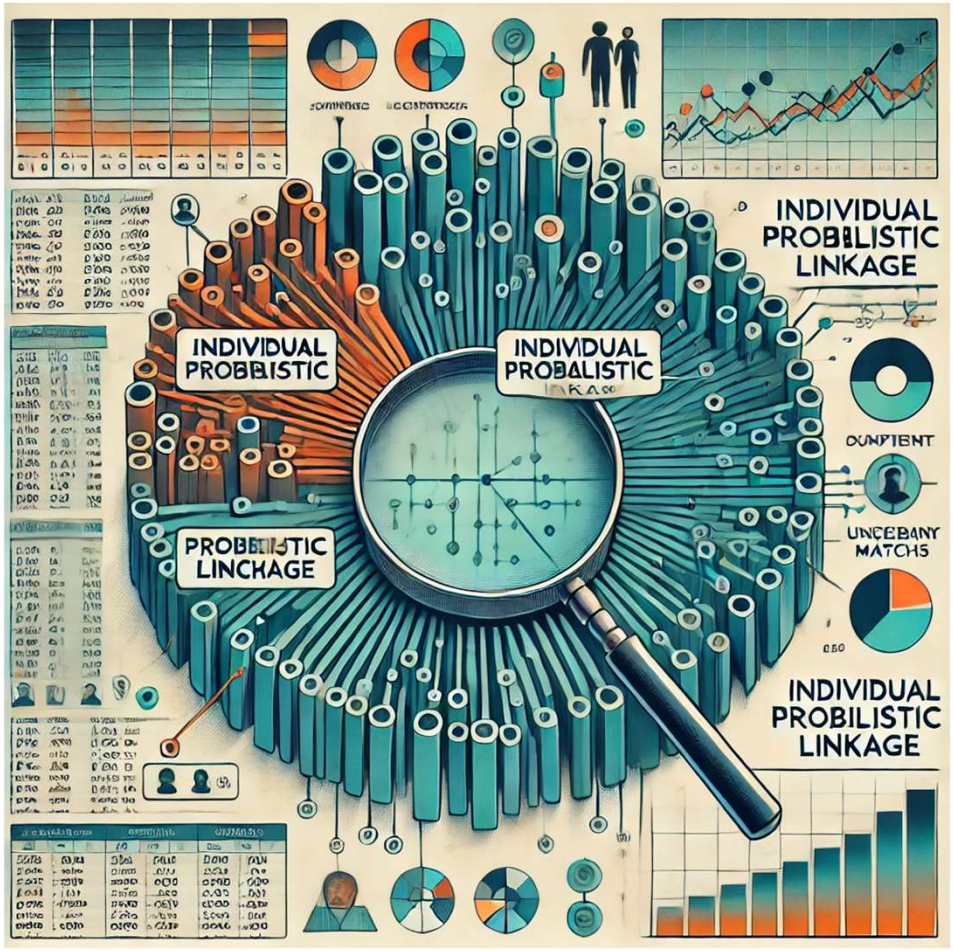
- Preparing data for linking
- The first step is to identify the variables to be used as matching keys in both survey and administrative data
- Make sure that in both datasets each corresponding pairs of variables have the same code names, the same formats (numeric, character, date) and have similar distributions.

- In both datasets: standardize the case of character variables: to lower or uppercase and suppress trailing blank spaces
- In both datasets: make sure that the key matching variables forms a unique identifier when stacked. If it's not the case, duplicates cases should be removed

Implementing deterministic data linkage

- Using left join on r
- Checking merged data
- Filtering out the unmatched cases
- Then continue with probabilistic linkage for the unmatached cased

Probabilistic linkage



Definition of probabilistic linkage

Probabilistic linkage is a technique used to identify and merge records that refer to the same entity across different data sources without a unique identifier(due to errors, some missing data on key variables, etc.). It uses probabilistic models to estimate the likelihood that records from different datasets refer to the same entity.

This method employs statistical models to calculate match probabilities, allowing for a more flexible and accurate matching process, especially when data quality is variable, or identifiers are missing.

Strengths of probabilistic linkage

1. **Flexibility:** Probabilistic linkage can handle variations and errors in data fields, such as typographical errors, misspellings, omission or abbreviation of components of the names, date provided in different formats, etc.. This makes it suitable for real-world data where perfect matches are rare.
2. **Improved Accuracy:** By considering multiple fields and their likelihood of matching, probabilistic linkage often achieves higher accuracy in identifying true matches.
3. **Scalability:** It can be applied to large datasets across different domains, such as healthcare, social sciences, and government databases, where exact matches are not feasible.
4. **Handling Incomplete Data:** Probabilistic linkage can still operate effectively when some data fields are missing or incomplete, using the available information to estimate match probabilities.

Challenges of probabilistic linkage

1. **Complexity:** The implementation of probabilistic linkage requires sophisticated statistical models and algorithms, making it more complex to set up and maintain compared to deterministic methods.
2. **Computational Resources:** Large datasets and the need to compare multiple fields across records can demand significant computational power and time.
3. **Threshold Setting:** Determining the appropriate threshold for match probability can be challenging, as setting it too high might miss true matches (false negatives), while setting it too low might result in incorrect matches (false positives).
4. **Finding error free variables for blocking.** In order to reduce the computational resources, one need to use one or many variables as blocking variables, that means that the matching is performed withing each block. Such variable may not exist. In that case, the computational resources can be high and matching done on servers.

List of some of packages for linkage

Tool	Language	Link	Description
RELAIS	R, Java	https://www.istat.it/en/methods-and-tools/methods-and-it-tools/process/processing-tools/relais	RELAIS (REcord Linkage At IStat) is a toolkit providing a set of techniques for dealing with record linkage projects.
fastLink	R	https://cran.r-project.org/web/packages/fastLink/index.html	Implements a Fellegi-Sunter probabilistic record linkage model that allows for missing data and the inclusion of auxiliary information. This includes functionalities to conduct a merge of two datasets under the Fellegi-Sunter model using the Expectation-Maximization algorithm. In addition, tools for preparing, adjusting, and summarizing data merges are included. The package implements methods described in Enamorado, Fifield, and Imai (2019) "Using a Probabilistic Model to Assist Merging of Large-scale Administrative Records", American Political Science Review and is available at http://imai.fas.harvard.edu/research/linkage.html.
Reclin2	R	https://github.com/djvanderlaan/reclin2	reclin2 is a package for record linkage an deduplication. The focus of reclin2 is on performance, memory and CPU and flexibility. To get the performance reclin2 uses data.table for most of its computations and reclin2 has the ability to spread its computations over multiple CPU cores or machines. In principle record linkage can easily be sped up using parallelization and by using multiple machines using the snow package data can be distributed over multiple machines thereby making use of the memory available on those machines.
RecordLinkage (R)	R	https://cran.r-project.org/web/packages/RecordLinkage/index.html	Provides functions for linking and deduplicating data sets. Methods based on a stochastic approach are implemented as well as classification algorithms from the machine learning domain. For details, see our paper "The RecordLinkage Package: Detecting Errors in Data" Sariyar M / Borg A (2010) <doi:10.32614/RJ-2010-017>.
		https://pypi.org/project/hlink/	hlink(hierarchical record linkage at scale) is a Python package that provides a flexible, configuration-driven solution to probabilistic record linking at scale. It provides

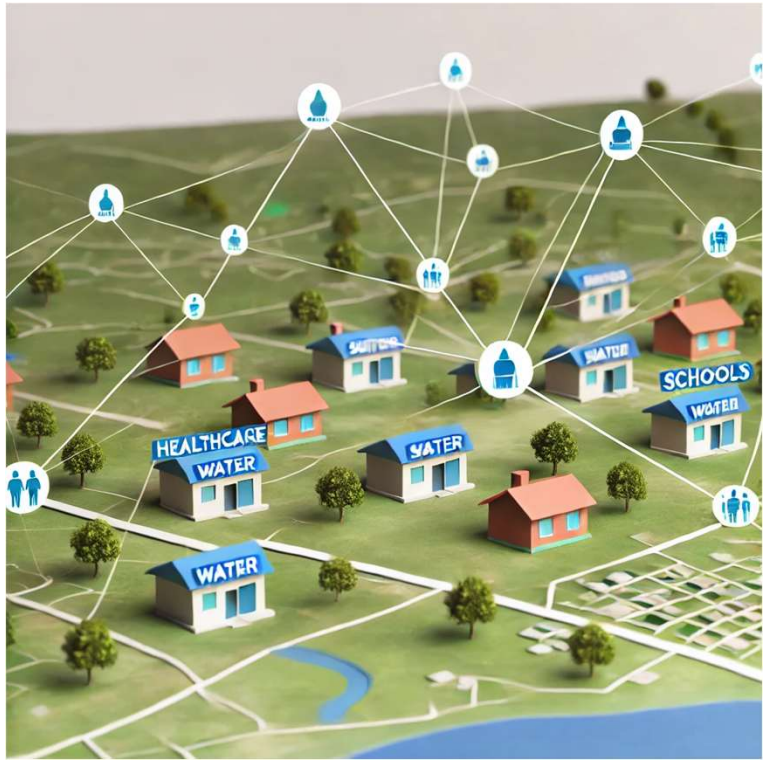
Aplication

Outcomes of the probabilistic linkage

We used reclin2 for the simulated exercise

The probabilistic approache worked in 99% of the cases only 7 cases out of 4183

The Individual/household linkage to services delivery points



Linking Individuals to Service Points

Importance of linking household survey data to service infrastructure: true accessibility analyis

- Geospatial Linking: Calculating distances between households and nearest service points.
- Example in the toolkit using goespatial linkage: Kenya DHS 2008-2009 linkage to SPA 2010 data for health facility proximity.

The use of Data integration by combining administrative and survey data to compute and monitor indicators.

Data Integration for Indicator Computation

Importance : Combining survey and administrative data for enhanced indicator accuracy.

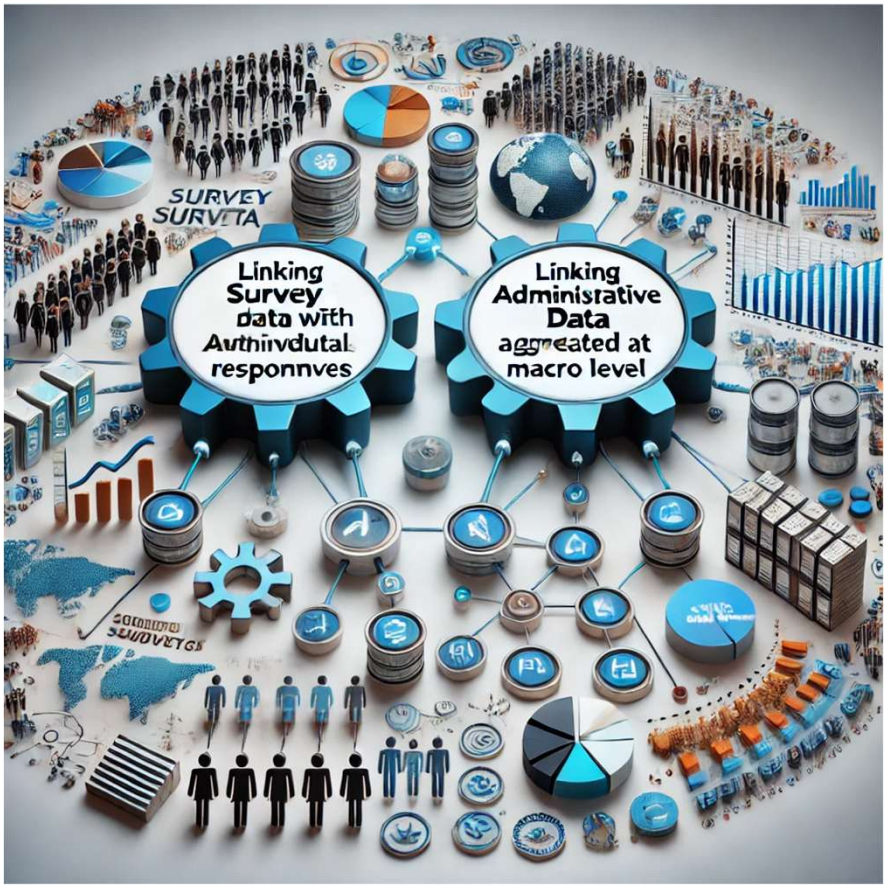
- Methodology: Estimating indicators from surveys, adjusting administrative data, and projecting.
- Example: Adjusting the Modern Contraceptive Rate (CPR) in Kenya using DHS and health data.

The steps

- Step 1: Use the survey data to estimate the indicator of interest (Is) in the population. This estimate is unbiased by definition
- Step 2: Use the administrative data and the population from projection to estimate the same indicator of interest (Ia) in the population. This estimate can be biased if not all the population of interest have access or use the service
- Step 3 Compute an adjustment coefficient k=Is/Ia
- Step 4: Apply this coefficient to adjust the indicator of interest(Iadj =k* Ia) estimated through the administrative data during the non-survey years

This methodology is based on the hypothesis that the bias is constant over time, which can be a strong hypothesis if the service coverage is improving rapidly. But if the annual rate (r) of improvement is known, it can be considered and hence on should use k*(1+r)n instead of k for the adjustment n years after the survey.

Linking survey data with administrative data aggregated at some macro level.



The rationale of the method

- The multiple reasons while on should integrate survey micro-data and different macro quantity
- When performing statistical analysis, one may want to consider some macro (or ecological) variables:

- Production (GDP) of the district
- School infrastructures in the district,
- Health facilities in the district,
- Kilometers of asphalted road in the district,
- etc.

- One way of doing that is to use the individual to service linking that we explained above. But this method supposes that there is a database of infrastructures records. But such data may not exist.
- In that case one can use the individual survey data to macro indicator linkage, provided that the macro indicators are disaggregated to a certain given administrative data

Example in the tool kit:Kenya DHS 2022 and the County gross domestic product (county well)

Conclusion

- Data integration is crucial for effective policy-making and SDG monitoring.
- Leveraging administrative data enhances the accuracy and granularity of indicators.
- Need for anticipating on linkage variables while designing census and survey (ex, in Kenya, Id number was asked during the last census)
- **Make admin data really available for statistical purpose**
- Building robust data infrastructure and promoting data linkage capacity.


ASANTE