Converting the HDSS Dataset from Long to Wide Format

Author: Tathagata Bhattacharjee

Email: tathagata.bhattacharjee@lshtm.ac.uk

DATASETS

This section describes the sources of the data used in the study, such as Health and Demographic Surveillance Systems (HDSS) and HIV clinics. It provides an overview of the data collection frameworks and their relevance to the study.

Population Data

Population data is integral to health research which includes a wide range of information collected from diverse sources to monitor and analyze trends within specific populations. This data typically would include demographic variables such as age, sex, and socioeconomic status, as well as health-related metrics like disease incidence, mortality rates, and migration patterns. Collecting and analyzing population data is essential for understanding the dynamics of demographics, health, and disease within a community, evaluating the impact of public health interventions, and informing policy decisions . The Health and Demographic Surveillance System (HDSS) is a notable example of a framework designed to provide comprehensive longitudinal data, offering valuable insights for research and model development in population health.

Health and Demographic Surveillance System (HDSS)

The Health and Demographic Surveillance System (HDSS) is a data collection framework designed to monitor and analyze population health and demographic trends over time. HDSS systematically tracks a defined population cohort, capturing a broad spectrum of data, including births, deaths, and migration, among other variables, through periodic surveys and continuous monitoring. This system is particularly valuable for longitudinal studies, as it provides crucial insights into changes and trends within a population over extended periods.

HDSS aims to provide a detailed account of demographic changes within a specific geographic area or community. It enables researchers to observe and analyze patterns related to mortality, fertility, migration, and morbidity, contributing to a deeper understanding of population dynamics and health outcomes. The data collection within an HDSS typically involves household surveys and individual interviews.

Verbal autopsy is a data collection method employed within the Health and Demographic Surveillance System (HDSS) to ascertain the cause of death. This method involves conducting structured interviews with family members or caregivers of the deceased to collect comprehensive information about symptoms, illnesses, and circumstances leading up to the death. The gathered data is analyzed to determine the likely cause of death. This inferred cause can then be cross-referenced with clinic data for disease condition confirmation and validation, provided that accurate record linkage between the two datasets has been achieved to ensure correct individual mapping.

In the context of generating synthetic datasets, HDSS data serves as a crucial reference point. Synthetic data generated from HDSS including the verbal autopsy datasets can mimic real-world population characteristics and dynamics, enabling researchers to develop and test models under controlled conditions. This integration helps in overcoming limitations related to data privacy and accessibility, while still providing a realistic basis for the development of models improving record linkage techniques.

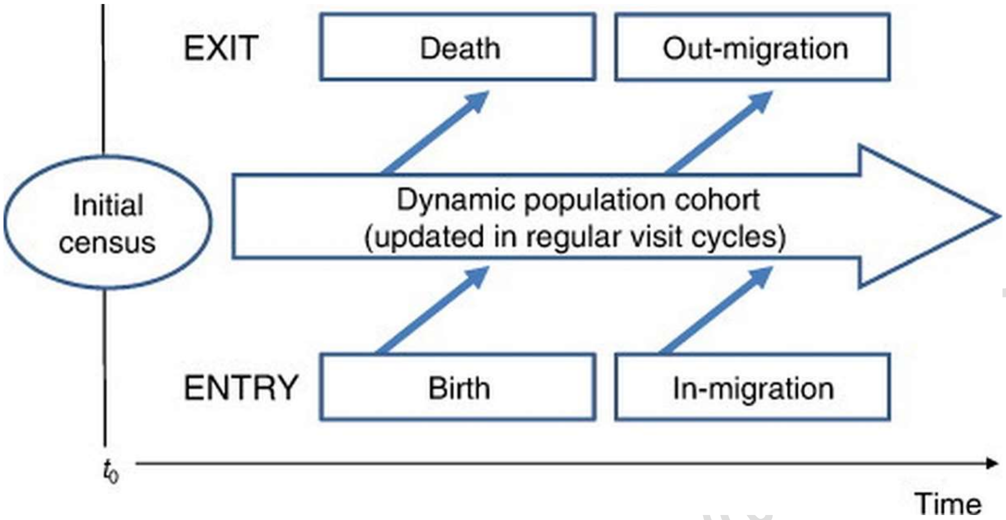

Figure: The concept of a longitudinal Health and Demographic Surveillance System [26]

The Kisesa Health and Demographic Surveillance System

The source datasets used in this study for the generation of synthetic datasets are from the Kisesa Health and Demographic Surveillance system.

The Magu Health and Demographic Surveillance System (Magu HDSS), also known as the Kisesa HDSS, is part of the Kisesa OpenCohort HIV Study in rural North-Western Tanzania, covering the Kisesa and Bukandwe wards in the Magu district. Established in 1994, it monitors pregnancies, births, marriages, migrations, and deaths with regular updates from field workers. The HDSS, serving a population of over 35,000 as of 2014, also conducts sero surveys, and verbal autopsies, and links health data, while addressing HIV stigma and ART access through qualitative studies. It provides crucial insights into population dynamics, fertility, mortality, and migration, supporting the evaluation of HIV/AIDS impacts, national interventions, and district health planning, with data shared via the INDEPTH Network's iSHARE repository.

Clinic Data

Clinic data includes health-related data obtained from medical facilities, such as clinical observations, diagnostic results, treatment logs, demographics of patients, and health outcomes. HIV clinic data specifically includes information on patient demographics (age, sex, ethnicity, socioeconomic status), clinical data (HIV status, viral load, CD4 counts, co-morbid conditions), treatment records (ART regimens, adherence levels, side effects), and health outcomes (symptom progression, opportunistic infections, clinical responses to treatment). Information from the Kisesa open cohort study in Northern Tanzania, extracted from the HIV clinic datasets offers a more comprehensive view of HIV trends. It shows that incidence peaked between the first and second intervals and stayed high thereafter, while prevalence rose from 6.0% in 1994/95 to 8.3% in 2000/01 before stabilizing. Roadside locations had declining incidence rates, especially for women, which resulted in decreased prevalence but increased slightly in remote rural areas. The findings underscore the need for intensified HIV prevention and ART interventions in rural areas and highlight the impact of high mortality rates among the infected and the variable incidence across different regions. The insights into HIV trends were derived from the HIV clinic data, but a more comprehensive analysis could be achieved by integrating the clinic data with the HDSS data. By accurately mapping individuals across these datasets, we could incorporate a broader range of demographic factors and achieve a more nuanced understanding at the micro level. This integration would allow for more detailed research and more insightful outcomes, enhancing the depth of analysis and the relevance of findings by capturing a fuller picture of how HIV impacts different population subgroups and interacting variables.

Point-of-contact Interactive Record Linkage (PIRL) project

the Point-of-contact Interactive Record Linkage (PIRL) software was developed for prospective linking in Kisesa HDSS area in rural Tanzania, where patient identity can be confirmed during clinic visits. Using a probabilistic algorithm based on the Fellegi-Sunter model, PIRL ranks potential matches in under 15 seconds and saves patient notes for future visits. However, this software was not originally designed for prospective record linkage. The linking process relied heavily on feedback from study participants, which means that when data quality issues exist, this approach could effectively facilitate the linkage despite potential inaccuracies in the dataset. The code and the procedure to run the software have been documented and are available in a GitHub repository for public access.

Here, in this study, we are generating the synthetic datasets from the source dataset prepared for the PIRL study.

PREPARING THE WIDE FORMAT LONGITUDINAL HDSS DATASET

To address the issue of migration within the Health and Demographic Surveillance System (HDSS) area, residency episodes are documented using two records per episode in the long or event history format. The first record captures the individual's exit from one location, and the subsequent record documents their entry into a new location. Movements from outside the HDSS area, whether inward or outward, the last recorded location within the HDSS area is utilized. This approach ensures that only internal migrations are captured accurately.

For analytical purposes, the long format, which includes multiple locations for each individual's internal movements, is converted into a wide format. This transformation consolidates all places an individual has resided into a single row. This wide format is particularly useful as it allows for comparisons with the location information recorded at health clinics during the record linkage using machine learning techniques, enhancing the accuracy and efficiency of the linkage.

The dataset for this case study was sourced from the Kisesa Health and Demographic Surveillance System (HDSS) located in Tanzania. The Kisesa HDSS provides comprehensive longitudinal data capturing various demographic, health, and migration-related information within its study area. This rich dataset serves as a valuable resource for understanding population dynamics over extended periods.

Pentaho Data Integration (PDI), also known as Kettle, has been used to convert the long-format to the wide-format. PDI Community Edition, is the open-source data integration tool that facilitates the extraction, transformation, and loading (ETL) of data from various sources into a unified format. It offers a user-friendly, drag-and-drop interface for designing data workflows, making it accessible for both technical and non-technical users. PDI supports a wide range of data sources, including databases, files, and applications, allowing for complex data transformations and integration processes. It also includes capabilities for scheduling, monitoring, and managing ETL processes, ensuring efficient and reliable data integration. PDI is often used in business intelligence, data warehousing, and analytics projects to ensure data consistency and accessibility.

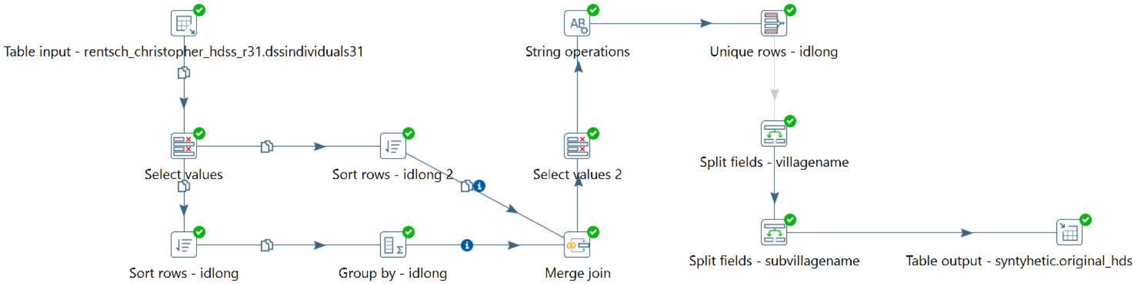
Table input - rentsch_christopher_hdss_r31.dssindividuals31
Select values
Sort rows - idlong
Sort rows - idlong 2
Group by - idlong
Merge join
Select values 2
String operations
Unique rows - idlong
Split fields - villagename
Split fields - subvillagename
Table output - syntyhetic.original_hds

Figure: The PDI Transformation preparing the longitudinal wide format dataset from the source

The steps below outline the Pentaho Data Integration (PDI) transformation process designed to convert HDSS longitudinal data, which includes multiple internal movements, from a long or event history format to a wide format. This transformation captures all residency locations (village names and sub-village names) for each individual within the HDSS area, consolidating them into a single row.

<u>*Step 1: Table input - dssindividuals31*</u> -> The source table was developed as an integral component of the Point-of-contact Interactive Record Linkage (PIRL) project, aimed at linking HDSS and HIV clinic data from the Kisesa study. This innovative project employs a probabilistic approach to initially list names matching participant-provided values at the clinic, followed by confirmation with the participant to accurately link the correct records. The table structure adheres to an event history or long format, where events are sequentially stored across two rows per episode, facilitating comprehensive tracking and analysis of demographic transitions within the dataset.

<u>*Step 2: Select values*</u> -> The unwanted columns from the data stream are removed.

<u>*Step 3: Sort rows*</u> – idlong -> In this step, the data stream is sorted based on the idlong column to facilitate further processing. The idlong column contains a unique identifier for each individual in the dataset, ensuring that all related records for each individual are grouped together sequentially. Sorting the data by idlong is a crucial preparatory step that organizes the dataset, allowing subsequent transformation steps to operate on well-ordered data.

<u>*Step 4: Group by – idlong*</u> -> In this step, the data stream is grouped based on the idlong value to concatenate the village and subvillage names into comma-separated strings. This grouping process consolidates multiple rows for individuals who have migrated within the HDSS area, allowing their various residency locations to be combined into a single column within a row.

<u>*Step 5: Sort rows - idlong 2*</u> - > In this step, a parallel sorting process on the idlong column is initiated. This ensures that both data streams are sorted in the same order based on idlong, which is crucial for the subsequent merging step. The next step involves merging the two sorted data streams, and for this to be successful, both streams must provide rows sorted identically by idlong.

<u>*Step 6: Merge join*</u> -> The Group By step grouped the rows based on the idlong and concatenated the village and subvillage names, producing an output that included only the idlong column along with the concatenated village and subvillage columns. Consequently, it is necessary to reintroduce the other columns that were not part of the Group By output to form a complete dataset. This is achieved by

merging the output of the Group By step with the original data stream. This merge operation results in a comprehensive dataset that includes all original columns along with the newly concatenated village and subvillage fields.

Step 7: Select values 2 -> In this step, unwanted columns are removed, and some columns are renamed to ensure the dataset is both relevant and well-structured for subsequent processing. This step is crucial for cleaning and preparing the data stream, making it easier to work with and interpret.

Step 8: String operations -> In this step, the values of the names and villages columns are converted to Initial Capitals (also known as title case) to ensure uniformity across the dataset. This step enhances the consistency and readability of the data.

Step 9: Unique rows – idlong -> After the merge step, the original rows of the datasets were brought back into the data stream along with the concatenated values, resulting in duplicate entries for individuals (identified by idlong). To ensure that each individual is represented by a single row, the duplicates need to be removed, retaining only one row per individual.

Step 10: Split fields – villagename -> To split the concatenated village names into multiple columns within the same row, the "Split Fields" step is used. This step splits the concatenated strings based on the comma delimiter introduced in the Group By step, resulting in separate columns for each village name the individual has resided.

Step 11: Split fields – subvillagename -> This step splits the concatenated strings based on the comma delimiter introduced in the Group By step, resulting in separate columns for each subvillage name the individual has resided.

Step 12: Table output - synthetic.original_hdss_individuals_wide_r31 -> This step stores the transformed data into a PostgreSQL database table in a wide format, where each individual's residency locations (villages and subvillages) are stored in separate columns within a row. By storing the transformed data in a PostgreSQL table, it is ensured that the dataset is persistent, and structured for efficient analysis and retrieval.