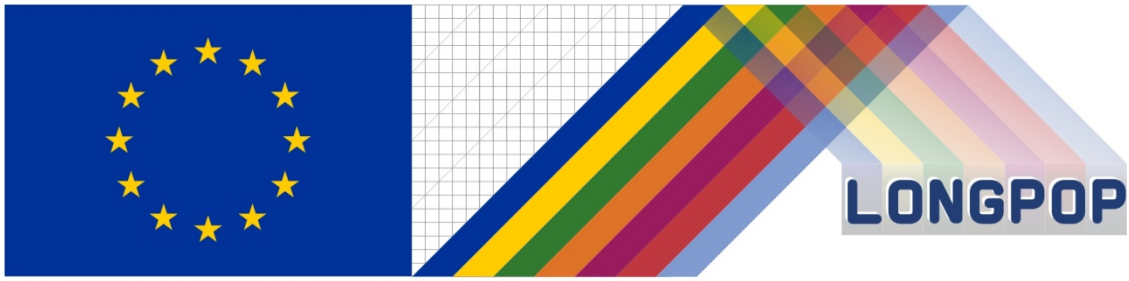

This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 676060

LONGPOP

Methodologies and Data mining techniques for the analysis of Big Data based on Longitudinal Population and Epidemiological Registers

GIS Mobility Tool

Deliverable n. 2.1

Three tools for R: extracting, calculating and visualizing migration data in the IDS-format.

For the LONGPOP-project the Radboud Group for Historical Demography and Family History has developed three tools for the R programming language, the Migration Extraction Tool, the Migration Statistical Tool and the Migration Visualization Tool.

1. ***Migration Extraction Tool***

Creating episode files for statistical analysis to study historical demography is a difficult and time consuming task, especially when using multiple, heterogeneous, longitudinal data sets. Creating analysis files requires data management and programming skills that not all researchers possess. This is especially true for the study of migration since that information is mostly not readily available or explicitly stored in a longitudinal data set.

The Migration Extraction Tool (MET) is developed to help researchers extract data on migration stored in historical longitudinal data sets in the IDS format and create an migration episode file. When historical demographic data sets are stored in the Intermediate Data Structure (IDS), tools developed to create episode files for a particular IDS dataset can be reused for other IDS datasets. This way, the process of comparing datasets is made much more efficient.

The MET consists of 5 parts:

1. Preparation: four steps to execute before the first script can be run.

2. Extract migration data: this results in a table with all migrations from one municipality to another for each Research Person (RP)

3. Build Personal data frame: this results in a table with all relevant invariant data for each RP.

4. Build Chronicle data frame: this results in a table with all relevant variant data for each RP.

5. Create Migration Episode File: the three tables are transformed and combined into an episode table.

The end result is the Migration Episode File (MEF), an episode table with migrations marked as events. The MEF is ready for analysis in R or if exported, in any other current statistical package.

2. *Migration Statistical Tool*

The migration rate indicates what the chance was that a person, often in a particular subgroup, migrated in an observed year. Calculating migration rates can be a difficult task. Specifically the counting of the observation person years can be laborious, especially when distinguishing between different variables and aggregating. The Migrations Statistical Tool (MST) is developed to support complex calculations of the migration rate so that historical demographers can be more efficient, especially when comparing different IDS-datasets with each other.

The Migration Statistical Tool consists of a collection of R scripts that together calculate the migration rates of males and females, for the age cohort 15-24, in two time periods (1850-1900 and 1900-1940) and aggregated to a regional subdivision of the Netherlands. The combination of the sex, age cohort, period in time and aggregation of a regional division implies that the MST has, after minor modifications, potentially a range of building blocks readily available for other research questions and calculations.

The Migration Statistical Tool consists of three parts. Each part contains one script that goes through multiple steps when executed.

1. Migrations: the number of migrations for each group and for each region is calculated
2. Observed person years: the number of observed person years for each group and each region is calculated.
3. Migration rates: the tables created in part 1 and 2 are combined and the migration rates are calculated.

Output

The end result is one table with for each of four groups (for each of the 44 regions) the migration rates.

3. *Migration Visualization Tool*

Although the end result of the MST can be used for analysis, the outcome can perhaps better be analyzed when plotted on a map of the country under research. Mapping data can be a time consuming and laborious task though. The Migration Visualization Tool (MVT) is developed to simplify the use of maps to compare migration rates within a country or between countries. This makes comparisons between different rates for one country or the different rates for different countries much more practical. To map the migration rates the MVT needs a polygon shapefile of a country with a regional subdivision that matches the regional subdivision used in the MST. The polygons that form the regions are be filled with a particular

color that matches a specific class using the migration rates calculated in the MST and a specified shading scheme. The MVT can of course also be used to map other data on polygon shapefiles. With the MVT the outcome of the MST, the table with migration rates, can be effortlessly mapped. The tool is split up in four parts:

1. Install and load R-packages: the necessary packages need to be installed and loaded.
2. Import and join data: the shapefile is imported and joined with the migration data.
3. Classify the data: the migration rates are classified.
4. Create and export the maps: the four maps are created and exported as PNG files.

Output

The end results are four chorepleth maps that show the regional differences between the male and female, 15-24 years old differentiated between 1850-1900 and 1900-1940 (see Figure 1 for an example).

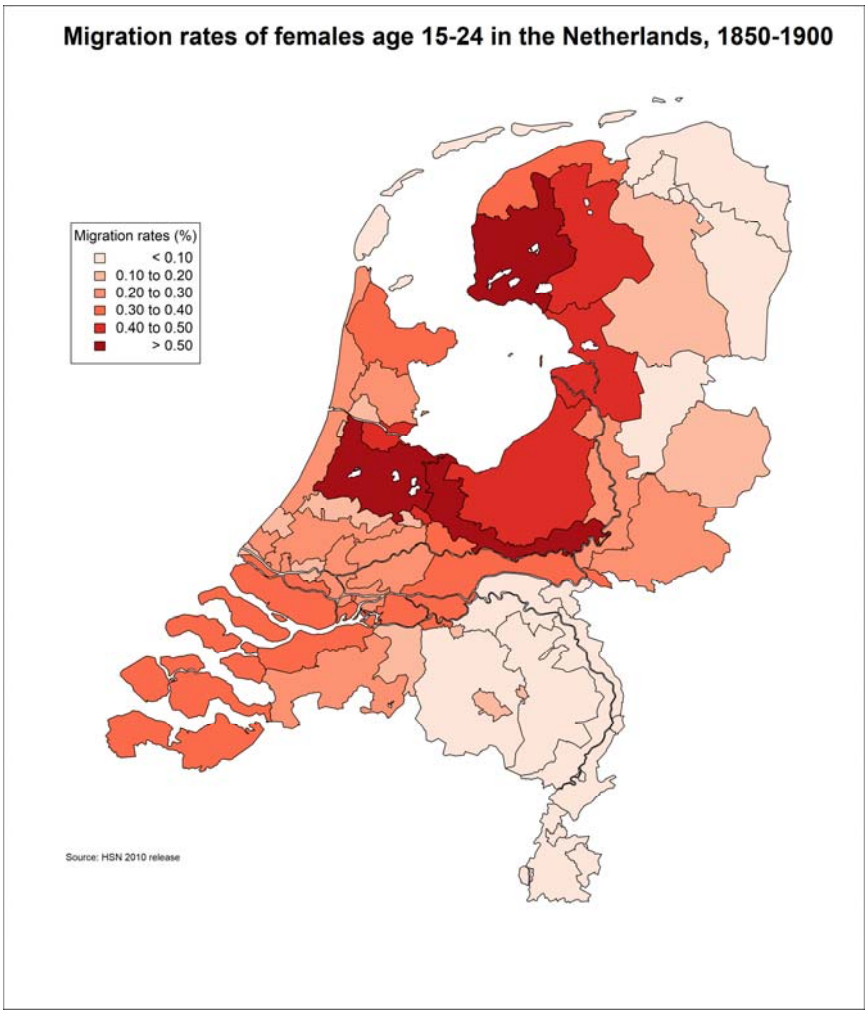

Figure 1: Migration rates for females age 15-24 in the Netherlands, 1850-1900

Remarks

The three tools have been deposited in the EHPS repository:

https://ehps-net.eu/software/migration-statistical-tool

The tools were developed using the IDS version of the HSN. This implies that the R scripts in the MET, MST and MVT have dataset- and country-specific coding in some places, which is inevitable, but the lines of code can be easily changed to suit the specifics of other datasets and/or countries.