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

Report on Different Algorithms

Deliverable 3.3.

Disclaimer: This publication reflects only the author's view and the Research Executive Agency is not responsible for any use that may be made of the information it contains.

Contents

Acronyms used in this document

ELSA: The English Longitudinal Study of Ageing (ELSA)

ESR: Early-Stage Research

GIS: Geographic Information System

IRP: Individual Research Project

HISCAM: Historical CAMSIS (Social Interaction and Stratification Scales)

HISCO: Historical International Standard Classification of Occupations

HRS: Health and Retirement Study

IDS: Intermediate Data Structure

IISG: Internationaal Instituut voor Sociale Geschiedenis

ISTAT: Istituto Italiano di Satistica

KNAW: Koninklijke Nederlandse Akademie van Wetenschappen

SEDD: the Scanian Economic Demographic Database

SHARE: The Survey of Health, Ageing and Retirement in Europe

SLS: Scottish Longitudinal Study

UEDIN: University of Edinburgh

UNIGE: University of Geneva

UNISS: University of Sassari

WP3: Work Package 3

1. Introduction

An algorithm is understood in mathematics and computer science to be a succession of instructions aiming at solving a type of problems or at performing computations, data processing, automated reasoning, or other tasks.

In this deliverable we understand that an algorithm is defined in a limited amount of space and in a previously defined formal language with the goal of the calculation of a function. Beginning from an initial input, it's build of a succession of instructions that carry out a computation that proceeds through a finite number of well-defined successive states after execution, eventually producing "output" and terminating at a final ending state.

It can be summarized as a set of rules defining a sequence of operations that would include all computer programs and eventually stopping with an outcome.

In the course of the ESRs' projects and in the light of the analyses they had to carry out, some of them had to resort to a lot of programming in order to provide solutions and outcome to their research questions. In the following lines we collect the most relevant concepts in the development of algorithms within the programming they developed.

2. ESRs' Developments

University of Lund, Sweden

ESR 3.- Enrico Debiasi.

In order to run statistical analyses on the Scanian Economic Demographic Database (SEDD), which is built on an IDS structure, ESR 3 extracted information in a longitudinal form using the STATA extraction code developed by Luciana Quaranta (2016).

The dataset in its extracted form has been used, analyzed, and further developed throughout the project. Firstly, they developed a common code that can be used by researchers to translate the occupational codes included in SEDD into HISCO codes and consequently into HISCLASS categories. Secondly, they extended the information about date and place of death included in the dataset by connecting SEDD to the Swedish Death Index through probabilistic linking. Eventually, they used the dataset in this updated form for statistical analyses of socioeconomic differences in adult mortality (see Expected Result 3.2 by ESR 3).

In working with data translated from IDS structure they identified strengths, but also some pitfalls that should be kept in mind by users when running the extraction code and when using the occupational codes translator. One of the clear strengths in using such tools is the standardization of final product i.e. of the dataset that is used for statistical investigations of demographic questions. These tools allow for different researchers to have a clean and comparable version of the information contained into the dataset. However, standardization also comes with some warnings as, when building the IDS structure, some decisions have been taken about how to handle certain types of variable and those should be clear to users. For example: the duration of a variable's value can be set to be instant or continuous. In the case of occupations this means that if the variable is set to instant, the information is used only in the point in time in which it was recorded; on the other hand, if the duration is set to continuous the information is used until a new value for the occupation is recorded. While this is a reasonable assumption, it may not be ideal for some research questions and therefore the users should be made aware of such mechanisms.

Radboud University Nijmegen, The Netherlands

ESR 4.- Dolores Sesma.

The substantive topic in this IRP is to find ways to integrate individual residential histories in the study of vulnerability and well-being; to extract migration and household trajectories from micro-data sets; to develop definitions and typologies; to analyze migration trajectories and sequences of households over the life course; and to convert data from datasets on addresses and households to files ready for statistical processing.

The main point concerning this deliverable of WP3 for ESR 4 are algorithms to extract mobility data from IDS converted databases.

In her study about the impact of migrant trajectories over the life course, ESR 4 applied a data-mining based method technique to study migration patterns on the long-run. An Optimal Matching analysis was applied measuring the dissimilarity between two sequences at the minimum total cost of transforming one sequence into other.

ESR 4 followed a state attribute-based costs strategy by using the Gower (dis)similarity algorithm (Gower, 1971). According to Studer and Richard (2016), this distance measure is suitable to determine the pairwise substitution costs for a combination of qualitative and quantitative attributes, allowing to measure the distance between all pairs of attribute vectors which compose the categorical states of migrant trajectories. The method "FEATURES" (Gower distance between state features) was applied to generate substitution cost by means of the function *seqcost* developed in TraMineR package (Gabadinho et al 2011).

The result is a substitution cost matrix with dimensions ns*ns, where ns is the number of seven states in the alphabet of the sequence object. The element (i,j) of the matrix is the cost of substituting state i with state j, which was calculated as the Gower distance between their vectors of *state.features* values. A more detailed explanation is available in TraMineR package. After computing the distance between sequences, a final solution cost was applied to build a typology using cluster analysis.

The study uses a subset of the Historical Sample of the Netherlands, Data Set Life Courses Release 2010.01.

UNISS, Italy

ESR 5.- Vanessa Santos.

In the case of ESR 5, the project focusses on the development of a set of concepts and techniques for managing longitudinal population data, and then perform longitudinal analysis in a long-term perspective. The project intends to analyze, for example, the spread of diseases and mortality rates in single communities, considering the distance between the houses or the kinship relationships and population closeness. The project intends to identify possible similarities between different communities by studying the distribution of mortality in relation to environmental and socioeconomic factors. It also analyses marriages, spatially evaluating propensity for marriage between people from different backgrounds in terms of their demographic, social and economic characteristics. The third aspect concerns the study of migration.

ESR 5's application of algorithms is circumscribed to a specific context within her project. The following paragraphs show a summary of the methodology used in the study they carried out and in what stage the Brooks-Gelman-Rubin algorithm was run.

An ecological study of small area was carried out for the 377 Sardinian municipalities existing in 2015, in the periods 1992-1997, 1998-2003, 2004-2009 and 2010-2015. Individual death entries for the period 1992-2015, broken down by municipality and sex, were used as case source. Municipal populations, broken down by age group (20 quinquennial groups) and sex are obtained for each year. The person years for each period were calculated by adding the population of each year.

Mortality and population data have been derived from ISTAT database.

For the operationalization and creation of the database, they used the 2001 deprivation index created by Caranci et al. as indicator of socioeconomic level. This index classifies municipalities into 5 levels, according to several variables: (i) low level of education, (ii) unemployment, (iii) non-home ownership, (iv) one parent family and (v) overcrowding. This classification is based on the quintiles of the distribution of factor scores, where level 1 municipalities are the richest and level 5 municipalities the less rich.

To calculate the number of expected cases, overall Sardinian specific age group, sex and period mortality rates were multiplied by each municipal person-years for the same age, sex and period pattern. Standardized mortality ratios (SMRs) were calculated as the ratio of observed to expected deaths.

Smoothed municipal relative risks (RRs) with their corresponding 95% credibility intervals, were calculated using the Besag, York and Molliè's conditional autoregressive model. This model fits a Poisson spatial model with two types of random effects, a non-structured effect that considers the municipal heterogeneity, and a structured effect, the spatial term, that considers municipal contiguity. To define area contiguity, we used the adjacent municipal boundaries.

The model takes the following form:

$$O_i \sim Po(E_i \lambda_i)$$

$$\log(\lambda_i) = \alpha + h_i + b_i$$

Where λ_i is the RR in area i, O_i is the number of observed cases, E_i is the number of expected cases, α is the intercept, h_i is the municipal heterogeneity and b_i is the spatial term.

To analyze the effect of deprivation on mortality, Caranci's index (Caranci et al., 2001) was included in the models as a covariate.

The Bayesian estimation of the models was obtained using Markov Chain Monte Carlo (MCMC) simulation methods, through the Gibbs Sampling algorithm via free distribution software WinBUGS. Convergence of the estimators was achieved before 100.000 iterations for three Markov chains, with a burn-in of 10.000 iterations. The convergence was ensured by the Brooks-Gelman-Rubin algorithm and the effective sample size of chains.

The free software R was used to create municipal maps of SMRs, smoothed RR estimates and posterior probabilities that smoothed RR was greater than one (PRPs). To calculate PRPs they used Richardson's criterion, considering PRPs greater than 0.8 as statistically significant.

UEDIN, Scotland

ESR 6.- Gergö Baranyi.

ESR 6's goals involve the understanding of how social and environmental conditions in early life affect subsequent health outcomes. In this case, two expected result have to do with work package 3. The point 6.2 about Data production states that the research data involves linking correct records inter-generationally to pick up information from parents' records through to adding ecological look-up tables (like the use of merging, appending, aggregating data, etc.)

The expected result 6.3 Analysis/algorithms concerns longitudinal modelling of the prepared data, including survival, multilevel and GIS techniques to answer the research questions.

ESR 6's response to the requirement shows that his project does not use any data mining methods or IDS, and the linkage between different sources of information with regards to the SLS (Scotland Longitudinal Study) has to be done by a support officer, as the data are very sensitive. The ESR 6 can only access prepared data (no data mining is allowed for him), as information on mental health medication is very sensitive.

Although this research project does not apply any data mining, it is using information based on natural languages processing algorithm. that NHS (National Health System of the UK) has developed recently, in order to identify certain medication groups. With this new dataset, ESR 6 has eventually rerun some of his analyses, but as stated, not from his own production.

In his projects, ESR 6 has used secondary datasets (SHARE, ELSA, HRS, SLS), where data has been already prepared for use. For these reasons, again, no data mining techniques were required. Data were processed and variables were computed by the data owners; They had no access to raw data.

ESR 6's latest project, however, makes use of administrative data (NHS prescription dataset), where data mining techniques were applied by the data owner (NHS Scotland) before providing them with the dataset. In the prescription dataset, there is a free text field where GPs can put their comments with regards of prescription etc. For their request, NHS eDRIS (Public Health and Intelligence Strategic Business Unit) extracted from these free text entries the prescribed quantity (e.g. 4 times a day 30mg) for all psychotropic medication prescriptions, using natural languages processing methods (like in McTaggart et al., 2018). They used this information to exclude lower dosage prescription from our dataset.

KU Leuven, Belgium

ESR 7.- Sam Jenkinson.

The field and main goal of this IRP is to link large numbers of individual time segments from large samples or populations to those of relevant networks, capturing the situation for particular periods and change over time means accounting for intra- and intergenerational connections.

ESR 7 is engaged with Family and Population Studies (FaPOS) department in KU Leuven, focusing on family research from longitudinal, intra-intergenerational and comparative perspective. Also, a highly valued database on 19th century Antwerp (COR-database) has been created, consisting of intra- and intergenerational demographics, sociological information recorded at micro, meso and macro levels. This database allows one to link three generations of longitudinal life course trajectories on key demographic events. Moreover, additional cross-national IDS datasets from the European Historical Population Samples Network (EHPSN) can be incorporated.

ESR 7's main goals with respect to WP3 are:

1. Concepts and techniques for individual record linkage. Algorithms and relevant syntaxes to construct intergenerational life course trajectories from original or IDS-converted multi-level and multi-source databases.
2. Methodologies and techniques for the longitudinal analysis: harmonization and transformation of historical data sets
3. Methodologies and techniques for the longitudinal analysis: harmonization and transformation of historical data sets

ESR 7 Report.

This appendix describes the process outputs for some of the bullet points explained in the section of ESR 7 project in KU-Leuven where algorithms were applied.

1.*Standardization and evaluation of individual linkages (COR-2010)*

The purpose of this exercise was to examine the quality of the linkage (IDNR, or unique individual identification number) of individual records and attributes across different sources in the COR database.

In order to do so, the authors began by splitting the database into observations from the original historical sources: birth, death, marriage certificate and event (i.e. population register) databases. Then they compared the accuracy of vital information of gender, dates and locations of birth, death and names across IDNR. This gave information concerning the size of any errors in linkage currently within the database.

The second step was to relink the database from the original source tables, ignoring the previous identification numbers given during the prior record linkage process. This allowed the authors to evaluate the process of the original linkage and provide the linkage algorithm for others to use. Here initially they used the methods as in line with the initial linkage 2010 (Van Balen). The authors use a stochastic record linkage method algorithm provided by Sariyar and Borg (Sariyar & Borg, 2010) as part of their R package "record linkage" (2016).

The record linkage process provided by the package uses the Fellegi-Sunter Model (Sariyar & Borg, 2010), as seen below.

$$\omega_{\tilde{\gamma}} = \log(\frac{P(\gamma = \tilde{\gamma} \mid Z = 1)}{P(\gamma = \tilde{\gamma} \mid Z = 0)})$$

Where the linkage relies on the assumption of conditional probabilities regarding comparison patterns. This works on the probability that a random vector γ=(γ1,…,γn) , having the value γ̃=($\tilde{\gamma}$1,…,$\tilde{\gamma}n$) , is conditional on the matching status of Z. Where Z=0 stands for a non-match and Z=1 equals match. In the full Fellegi-Sunter algorithm these are used to compute weights which are used in order to discern matches and non-matches. The weights within the package were computed using an expectation maximum (EM) algorithm in line with Haber (1984) and Contiero et al (2005).

Then common variables are used across observations sources to calculate the likelihood of a record being a match using a string comparison tool. The selected variables here include given and family names, birth location, birth day, birth month and year separately. These are then used to calculate similarities across different records and to create pairs.

The string comparison used here is the Jarrow-Winkler distance string comparison algorithm (Winkler, W.E 1990). This function works by measuring the edit distance between two strings and calculates the minimum numbers of single character transpositions to transform one word into another.

The original Jaro distance calculation can be seen below;

$$Jaro\ dist = \frac{1}{3}\ (\frac{m}{|s_1|} + \frac{m}{|s_2|} + \frac{m-t}{m})$$

Where m= the number of common characters which are within half of the length of the longer string and t the number of transpositions. The Jaro-Winkler distance is the same, but with the following changes

$$JaroWinkler\ dist{=}Jari\ dist{+}l(1{-}Jarodist)$$

Where ι represents the common prefix at the start of a string up to a maximum of 4 characters, a standard approach for these purposes.

In addition to this the process using the procedure of blocking (Sariyar & Borg, 2010) is begun to ensure the strictest criteria for record matches, before sequentially relaxing the criteria in order to report the frequency of matches given at a particular matching stringency. Blocking limits the number of matches by ensuring that one particular column or variable is a 100% match. This uses a phonetic algorithm contained within the package before moving on to assess the individual data contained in other columns using the string comparison tool. This is useful to provide perfect matches but also used to reduce computation time. The process begins with one column which is a unique column based on all data contained in the other columns used for linkage. This column contains a string made of all names, birth dates and birth locations. This will then give us the number of matches which are 100% identical across all comparison fields for birth and death certificates. The authors then slowly remove the number of variables, over three rounds initially here, included in the initial blocking identifier column and report the number of matches and therefore their quality.

The steps are implemented as follows. The first name was split into 5 possible first name variables owing to the likelihood for some to have multiple given names. Initially all 5 of these first names, as well as family name, Birth location in NIS code, birthday, birth month and birth year are included in the blocking criteria. In the second round only the first three given names, surnames, birth places and birth dates are included and in the final round the blocking criteria included only the full birth dates (day, month, year) and two first names and family names.

During this linkage procedure, another algorithm under the shape of a cleaning function is created in R. This involves standardising all variables, just for the purposes of linkage, to lower case and removing any white space before or after characters which may have been mistakenly entered by initial data entry. This caused multiple problems on initial linkages as many observations had stray white space before, after and within names making comparisons difficult. In addition erroneous capitalisation was frequently also a problem.

The exercise was carried out first with both birth and death certificates, representing the easiest to match with only one entry in each record per person, in comparison to marriages and events where each person may have multiple observations. The first step was to remove duplicate rows containing identical information. This process of sub-setting the tables to retain only unique information removed 171 rows from the birth table and 183 from death[1].

Following this the cleaning function was carried out on both tables. This removed white spaces from before or after each string. It also removed blank information and stray punctuation inputted during data entry with no meaning and replaced with "NA" which is standard for the R package "record linkage".

After the first three rounds there were 1,514 pairs created. Of those 79 were new pairs not existing in the prior database, representing 5.2% of all new IDNR. These new pairs have been extracted for analysis to investigate why these were missed in the initial linkage. Many appear identical, indicating that it is the script cleaning stage which has led to new linkages and in others it is where information is missing and replaced with NA (Not Applicable, the code R programming language to the absence of data to prevent from errors produced by NULL -empty- data). However, 1431 cases that has the perfect match are used to create the first step of individual table.

The remaining points in the ESR s project involved the specification of standardized variables for meta data, the structure of the individual attributes on core variables from the initial output; the state of art on production of indiv-indiv data, and, generally speaking, the nature of the records in COR_2010.

Nevertheless, the identification in the sources of members of the families required further use of algorithms. For instance, getting ID mother involved the following steps: first, sorting all individual units by identification house, allowing to identify individuals to select to those who have resided one moment in time in the same location. All individual identifier is indicated in IDNR. Number in the household is given where 1 is the head of the household. Identifying the id number of the mother (=2625) can be done by matching the number in the column number in household, Idnrmo and then to IDNR. The assigned idnr of the mother is created at idmoeder. As it can already be seen, this process of identifying the id of the mother is complex.

How to identify siblings: the variable rellid equals to 13 for household number 22, 23 and 24. Since 13 is brother or sisters, the authors assume that these 3 individuals are sisters/brothers and the mother is id 2625. As this information is interpreted, they assume that they are biologically related but cannot verify. The remaining rellidX variables and corresponding indrX are great mystery as the authors are uncertain which relationship they refer to given the fact that in this location, only number of the household 1, 21-24 and 29 exists and all others in between are not recorded. The authors find the guessing and reinterpreting of the relationship between individuals to be particularly problematic to be used for the creation of indiv-indiv table.

In addition to the above, there are spatial information present in the location table which potentially allows to create the relationship variable when the location of residence is shared. This is the potential information that may be used to create indiv-indiv attribute by replacing the source information of the items noted just above.

- Reference house information concerning

[1] Within the birth and death tables there were some rows for people which were identical within the table rather than between.

- Municipality (refgem)
- Population register (refBR)
- Quarter number (refwk)
- Quarter house number (refwkhsnr)
- Street (refstr)
- House number (refhsnr)

The authors considered “workable” solutions to the aforementioned problem given the circumstances: what methods they can apply in obtaining the indiv-indiv table. At the conceptualization stage, they first acknowledge that there exist familial and non-familial relationship to each individual.

The first familial ones are namely determined by intergenerational linkages mostly referring to mothers, fathers and children, and if identifiable aunts, uncles, cousins. They consider that this information can be identified and matched using relevant sources, primarily through certificates.

The second ones refer to the location where non-familial relationship is established when they share the same address for instance. This can be achieved by making use of location table, population register (if possible) and then also with the relevant certificates. For the latter, the obtained location identifier is present (a1, a10, …) (population register, huissamen2) where these sources allow to match with the relevant individuals in the population register. However, if the same is applied they to only work with raw inputs and not with the acquired variable present source, they will need to match location with individuals through addresses. This means another linkage based on location variable can be performed by identifying location (addresses) present in the certificates and population register.

IISG – KNAW, The Netherlands

ESR 2.- Francisco Anguita

In the course of the research project “Linking the American Census with the HSN”, ESR 11 and ESR 2 developed several scripts using R programming language in order to carry out data linkage between two historical data sets: the Historical Sample of the Netherlands (HSN) and population registers from the American census of 1850 until 1940.

The goal of the project was to try to visualize those records lost form observation in the HSN with the help of the American census of the abovementioned years.

A relevant number of algorithms were developed in order to harmonize, clean and standardize both large sets of records (HSN and Census). Information in HSN was derived from several tables through the correspondent informatic routines (HSN Marriages, HSN Lost, HSN Basic, HSN Survival) as so it was in the case of the very heterogeneous censuses depending on the year.

Methodologically speaking we distinguished two applied approaches on our process of linking. The first one consisted of an algorithm of six general steps, the aim of which was to narrow the margins of the record matching

Initially, we matched only the first four initials of the last name and the first initial of the first name. Then individuals’ birth year should be earlier than the relevant census year, or death year later than the census year. Then we applied the principle ± 2 years between the birth year of a record in HSN and its potential match in the Census. Next, we applied another routine measuring string distances between names (last and first). For it the algorithms of Levanhstein, counting the number of edits a word needs to be converted in a second string suited us. Finally, blocking by sex was the last step.

After this, a set of potential matched paired records (one in the HSN and the other in the Census) were obtained.

To them, another large routine was implemented with the aim of validating this outcome. These algorithms mostly focused on the eventuality of finding relatives of the matches pairs at both, the HSN and the Census.

On the other hand, the second approach developed methods to potentially revert the changes Dutch first and last names underwent upon arrival to the US borders. These changes were mainly "anglizations", and the creation of a "dictionary" that canceled them -achieving the original most likely form of the name- gave us some extra matched records for our outcome.

All the coding was developed with RStudio and R, taking advantage of the multiple functions that the different packages -mostly dplyr- contain. One thing to remark was the necessity to optimize all the routines as all the work involved the managing of a big amount of records (736.000 from the censuses, 85.000, from the HSN). That is why functions working with vectors and matrices from the family "apply" were very welcome as their performing was indescribably faster running all the registers than traditional loops.

ESR 11.- Diogo Paiva

This section presents some of the ideas employed by ESR 11 in the process of georeferencing addresses contained in the Historical Sample of the Netherlands (part of the Expected Result 11.4 - *Building algorithms to implement a GIS based address system for the two databases that are the most suited to start with, result: adding variables with GIS coordinates)*.

In this case, the goal of this project to add geographic coordinates for each historical address present in the current public release of HSN[2] (almost 340,000 observed addresses).

To be able to proceed with the geocoding process, an algorithm was employed to decompose the addresses in their constituent elements (number, street, *wijk* or district and municipality).

Based on the available resources -a dataset with actual postal codes and public spaces in the Netherlands with a corresponding coordinate and a dataset of historical addresses already decomposed into its constituent elements (house number, street, *wijk* and municipality)-, a process of record linkage is defined by ESR 11. The geocodification of the HSN would is achieved by assigning a modern postal code to each historical address. The granularity corresponds to the centroids of the postal codes polygons, which provides an adequate spatial precision.

The record linkage between modern postal codes coordinates and historical addresses are but a final step in a larger workflow of digitization and georeferencing of the historical addresses present in the Dutch population registers (Figure 1). Therefore, the success of linkage is highly dependable on the performance of the previous steps, thus creating the first challenges that we faced. The system of data entry of the HSN relies on a series of checks and revisions after the manual data entry of records that minimizes human transcription error to a minimum. Nonetheless, some mistakes persist and are passed on to database. Moreover, data entry in the HSN follows a logic of literal transcription and minimal interpretation from the person entering the record. Although it helps

[2] HSN Release Version 2010.01.

reduce transcription errors, it mirrors the spelling variety present in the sources. It became evident that a pre-stage of standardization was needed to implement.

As the transcribed addresses were recorded as literal text strings, additional to the spelling variation, the structure of the addresses were also copied in their original format. This created a structural issue when decomposing the addresses into its elements. Since for most of the period that is covered by the sources used in the HSN there was not a single norm followed by all municipalities regarding the addressing system, a variety of addressing formats and systems were recorded. As decomposition of addresses was obtained through an algorithm that could not efficiently take into consideration all the diversity in the recorded addresses, a part of the data was imperfectly arranged.

For example, when the decomposition algorithm considers "A Alexanderplein 21" it can decompose the address as "A" (*wijk*) + "Alexanderplein" (street) + "21" (house number). But when the same address was recorded slightly different as "Alexanderplein 21 A", it becomes very challenging to properly distinguish "21 A" as the structure house number + wijk from just a house number with an extra letter "21A".

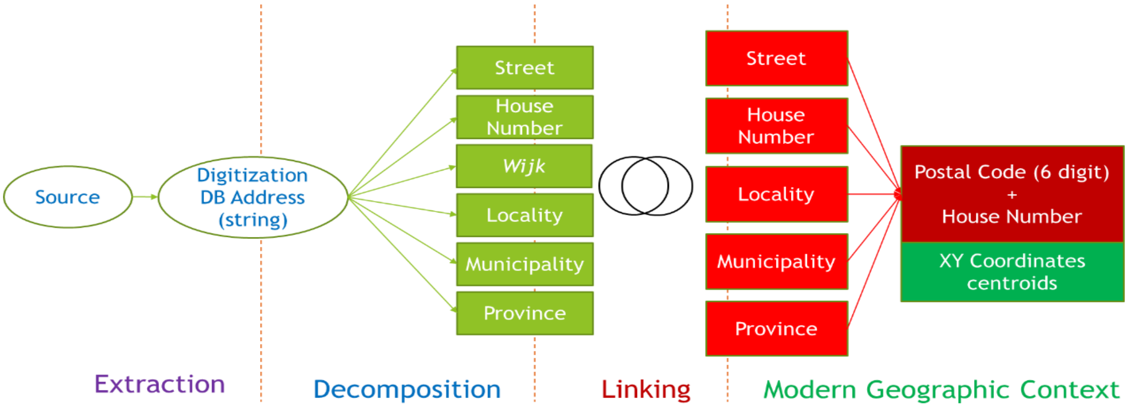

Figure 1 - Methodology of Georeferencing the HSN

Regional and local differences conditioned the organization of administrative space and, consequently, of addressing. In lesser dense municipalities addresses can be composed of only the municipality and a house number, i.e., house numbering system is for the whole municipality, instead of numbering by small sections (like streets or *wijken*).

Besides spelling variation and address system/structure variation which impact data entry and the decomposition algorithm, a third challenge arises from the nature of the sources and the dimension of historical time on administrative processes. The Dutch population registers started mid-19th century consisting on a double page form recording the household members and associated information (including place of residence), with all changes being also recorded. This later evolved to a family card system at the beginning of the 20th century until the 1930s when a personal card was introduced. It was a de-centralized system, with each municipality having the responsibility of bookkeeping. This leads to a degree of diversification of the forms, ways of recording, linguistic differences, personal styles and local specificities that were gradually reduced in the 1900s with introduction of the new systems.

For tackling these challenges, a standardization and normalization process was designed. This is an intermediary step between addresses decomposition and record linkage with the BAG file (modern postal code coordinates). The fields to be standardized are: STRAAT, HUISNR, HUISNRTV, WIJK, WIJKHSNR, WIJKHSTV and DEELGEMENTE. Concerning fields relate to house numbering the main issue was misplaced values, i.e. street names or *wijken* wrongly placed by the decomposition algorithm. Regarding the other fields (streets, *wijken* and *deelgemeenten*), although a fair amount of misplacement was identified, the main issue was the lack of standardization of names. The street names was the most problematic field with a higher degree of entropy and therefore the first efforts were focused on the field STRAAT (Table 1).

Field	*Unique values*	*Unique values (by municipality)*[3]
STRAAT	64,700	86,586
HUISNR	3,331	49,887
HUISNRTV	5,063	10,646
WIJK	3,940	13,236
WIJKHSNR	2,295	45,158
WIJKHSTV	1,085	6,068
DEELGEMEENTE	2,366	3,852
Values to Review	**82,780**	**215,433**

Table 1 - Values for standardize/normalize

The process to deal with street name values considers two main stages: (1) definition and conversion to a standard and (2) identification and typifying the value (Table 2). For the first stage, the standard is the official name (and spelling) present in the BAG file. The second stage verifies if the street can be found today, and if it exists it typifies it based on the suffix (e.g. –straat, -weg, -gracht). A list of possible outcomes is provided in Attachment 1.

STRAAT *(original value)*	*straat_std* *(standardized value)*	*Type* *(identified and typified)*
2e Atjehstr.	Tweede Atjehstraat	straat
Prinsegr	Prinsengracht	gracht
K. Houtstr	Korte Houtstraat	not found

Table 2 - Examples of street standardization process

As this process of standardizing values and typifying them has a large human input, especially at the beginning of the project, a set of rules and decision-making diagram was established early on to prevent deviations (Attachments 2 and 3). Dealing with the whole set of street names took more than 500 hours of work, spread over several months. Without these rules it would be increasingly hard to make the same decisions in similar cases as time (and knowledge of the data) progressed. Because some values in STRAAT are misplaced, a new standardized variable was created for localities: *woonplaats_std*. After all street names were processed, the other fields were standardized following a similar standardizing scheme with a parallel process of reassigning misplaced values (Table 3).

[3] I.e. summing unique values counts per each municipality. This allows duplicates in the case that same values were recorded in different municipalities. For example, the value “dorpstraat” is counted once in *Unique values*, but 76 in *Unique values (by municipality)* as 76 different municipalities recorded a Dorpstraat.

STRAAT	*WIJK*	*DEEELGEM*	*straat_std*	*wijk_std*	*woonplaats_std*
Mijnden A				A	Mijnden
		Stoutenburg			Stoutenburg
		Oud. Aa			Oud Aa
J.v/d Doesstr.			Jacob van der Doesstraat		
D	Kruidenierstraat		Kruidenierstraat	D	

Table 3 - Examples of standardized addresses

Due to the lack of a dictionary or conversion file to assign validated standards, the process was initially fully manual coding using R. Every municipality was coded with the Amsterdam Code that served as criterion to create subset of street names to standardize. After the initial provinces were processed (Drenthe, Flevoland – Urk – and Friesland) a small dictionary was obtained. The following provinces were standardized by a semi-automatic process. A sub-routine was designed to convert original street names into known standards and afterwards verify if the standardized form was present in the BAG file. If this failed, then a manual standardization was necessary, which was aided by a small algorithm that suggested five likely options. The introduction of this subroutine accelerated the process without compromising the rigour of verifying official names.

Once the process of standardization and normalization is concluded, the data is prepared to be linked with BAG file and thus georeferencing the historical addresses present in the HSN. Because different address systems were used along the period of the HSN, a triple system of coordinates: street, locality and municipality.

Street coordinates (s_lon, s_lat) were linked with the HSN addresses through the BAG file. Given the historical nature of the addresses of the HSN, a simple process of record linkage would fail substantially. Therefore, a variable matching criteria was used to improve the efficiency of the linkage. Street coordinates are linked in a four turn process. For each turn, a subset of addresses that failed matching is produced and only those are tried with new criteria. Firstly, addresses from HSN are linked with BAG file using street, house number and municipality. Secondly, street, locality and municipality are the criteria. Thirdly, to deal with historical municipalities that were annexed into others, old municipality as locality, street and current municipality are used. Finally, just street and municipality are connected.

Coordinates for localities (p_lon, p_lat) and municipalities (m_lon, m_lat) were obtained by linking the woonplaats_std with the Huijsmans' file. For those localities that are inexistent in this file, the coordinates were obtained through the georeferencing of historical maps of municipalities compiled by J. Kuyper and locating the missing localities, using ArcGIS Pro. Finally, a composite coordinate (n_lon, n_lat) is created of the most precise location for each addresses. These values are the street coordinates unless these are unknown in which case the locality coordinates are taken and if they are also unknown then the municipality coordinate are used.

Tools

- Combos

  - Goal: Merge (inner join) a dictionary (conversion table) of street standards with un-standardized streets, to try to provide a validated standard to a given original street name
  - Process
    a. Gets list of original street names from one municipality
    b. Fetches the dictionary of all known street original to standard conversions
    c. Defines the most frequent conversion for each original value (sometimes same original name correspond to different standards in different municipalities) and discard less frequent
    d. Merge list of street names to standardize (a) with list of known conversion (c) => for those not successfully matched manual standardization is required
- Found or Not
  - Goal: To determine automatically if the assigned standards exist in the BAG file and typify
  - Process
    a. Gets list of official streets and public spaces from BAG of a given municipality
    b. Checks if each standard is in the list from BAG (a)
    c. For standards found in the list from BAG (a), automatically typifies the name (if is straat, weg, laan, etc.) for the most common types
    d. For incomplete street names it typifies “not found”
- Move on
  - Goal: To automatically typify those standards that were not found in the BAG file
  - Process
    a. Assign type “not found” to those more common suffixes (-straat, -weg, -laan, -steeg, -singel) that do not require further revision.

UNIGE, Switzerland

ESR 15.- Rose van der Linden.

The goal of this ESR ’s research is to test their hypotheses (the importance of life conditions in infancy, the intergenerational transmission of resources, or the accumulation of (dis)advantages across the life course result in a large intra-cohort gradient) within an integrated framework, analysing the roots of social inequality in health.

Concerning the point of analysis/algorithms of Work Package 3, the project involves longitudinal modelling of the data prepared including multilevel modelling to answer the research questions.

The contribution of the Geneva team was mainly limited to the maintenance of TraMineR, a world-wide known tool for data mining and sequence analysis written in R and completely free of access. They supported the use of TraMineR by LONGPOP researchers exploiting databases structured according to the IDS standards. For example, Sam Jenkinson’s (KU Leuven ESR 7) research on risks of divorce was accompanied by Matthias Studer, TraMineR developer and current leader, as was the construction of clusters of residential life trajectories from the birth to 50 among a sample of some 8’000 Dutch residents, extracted from the Historical Sample of the Netherlands -research done by Dolores Sesma (Radboud University ESR 4).

3. References

Alter, G. & K. Mandemakers (2014). **The Intermediate Data Structure (IDS) for Longitudinal Historical Microdata, version 4.** Historical Life Course Studies 1 (2014), 1-26, published on line 26th of May 2014. PI: http://hdl.handle.net/10622/23526343-2014-0001?locatt=view:master

Klancher Merchant, E. & Alter, G. (2017). **IDS Transposer: A Users Guide**. Historical Life Course Studies, 4, 59-96. http://hdl.handle.net/10622/23526343-2017-0004?locatt=view:master

Quaranta, L. (2016**). STATA Programs for Using the Intermediate Data Structure (IDS) to Construct Files for Statistical Analysis**. Historical Life Course Studies, 3, 1-19. http://hdl.handle.net/10622/23526343-2016-0001?locatt=view:master

Schumacher, R., Matthijs, K., & Moreels, S. (2013**). Migration and reproduction in an urbanizing context. Family life courses in 19th century Antwerp and Geneva**. Revue Quetelet/Quetelet Journal, 1, 51–72.

McTaggart, S., Nangle, C., K., Caldwell, J., Alvarez-Madrazo, S., Colhoun, H., & Bennie, M. (2018**). Use of text-mining methods to improve efficiency in the calculation of drug exposure to support pharmacoepidemiology studies**. International Journal of Epidemiology, 1–8, doi: 10.1093/ije/dyx264