MASTER

Marriage Mobility Visualization for Genealogical Data

Xu, Jinfan

Award date:
2020

Technische Universiteit
Eindhoven
University of Technology

Department of Mathematics and Computer Science

Marriage Mobility Visualization for Genealogical Data

Master Thesis

Jinfan Xu

Supervisors:

Dr.ir. Huub van de Wetering
Dr. Michael Burch
Dr. Kees Huizing
TU/e

Christian van der Ven
BHIC

Eindhoven, November 2020

Abstract

The Brabant Historical Information Center (BHIC) collects various types of genealogy data in the province of Noord Brabant. The datasets they collect contain the records of different events in one's life, including birth, marriage, military, prison, residence register and death. And for each kind of event, the event information, such as event date and event place, and the information of people who involve in this event, such as birth date and birth place, are recorded. In this way, the mobility patterns, which refers to the trace of one's life, can be observed. Currently, except for searching and viewing the related records in a textual way, BHIC has no better way to let the users explore the data. In this case, the users are not convenient to match the place names in the datasets to the corresponding geographical locations. Besides, mobility patterns are difficult to be distilled from textual records. Simultaneously, analysis of the multiple attributes of people, such as birth date and gender, can not be provided by searching. Therefore, BHIC is interested in having a web tool for both the public and experts to explore the geographical data in a more interactive and efficient way.

An area that has not been extensively researched by BHIC is visual data exploration. In this project, we choose to visualize the marriage dataset in BHIC datasets to explore the mobility patterns caused by marriage. A visualization-based tool is designed to explore marriage mobility. To keep the user interface of the tool uncluttered, multiple views, including a map, a matrix, a heatmap, and a sidebar, are designed to provide different information to the users. In this way, the users can not only view marriage mobility patterns, but also explore both aggregated and detailed information. Furthermore, the tool provides various interactive functions to enable the users to perform different tasks. For example, the linkage of these views enables the users to explore the comprehensive information for some specific places. In addition, we take the non-expert users into account, so that they can easily interpret these views and explore the data.

Preface

I have learned a lot during the master project. First, I would like to thank my supervisors Dr.Ir. Huub van de Wetering and Dr. Michael Burch from Eindhoven University of Technology, for their patient guidance during the project. In this project, they guided me how to resolve the goals to small tasks and requirements, how to design an interactive tool, and how to logically finish the report step by step. The advice they have given when I met problems were always instructive, and their comments for the improvement of tool and report were very helpful.

I would like to thank Christian van der Ven from BHIC. While communicate with his about the BHIC dataset and the goals of the project, he patiently answer my questions in detail and gave great support for the project.

I would like to thank my friends and families who have accompanied and encouraged me during the project. As the situation of Corona-virus and the lockdown, it is lonely to stay and work at home. The greetings and company from friends really reduced the pressure and ensured my work efficiency. The understanding and support from my parent also encouraged me a lot during the whole master project.

Contents

Chapter 1

Introduction

Genealogy is a discipline that studies families, family history, and their lineages. Genealogical data is essential for studying society, history, culture, economy, and their changes. The use of genealogy information to comprehend the spatial distribution of population and the spatial characteristics of migration, which is a considerable significance to social development study.

In the past, a traditional method to store genealogy data was manually recording on paper. After the computer was invented, the genealogy data began to be recorded digitally. The drawback of digitizing these materials is that it takes a great deal of time and effort. More specific, there are a great many logistical problems that need to be addressed in digitizing records. First, the physical acts of scanning, cropping and color correction all take time. Next, working with fragile materials that need delicate handling takes time.

However, by making a digital copy and placing them online, people are no long to worried about preservation and more people have access to the genealogical information. Besides, with the digitized genealogy, researchers can perform different techniques to make the data more visible and understandable. Visualization is widely used in processing the genealogy data, for it can convey the statistical information of genealogy data in a graphical form by using the human visual ability.

Various visualizations provides an ability to comprehend huge amounts data[26]. The most common visualization for genealogical dataset is family tree, such as a pedigree chart, which shows some information and the structure of ancestors[10]. More recently, additional details of individuals, such as geographic information about their life events, are supported to be visualized. More and more researchers pursue to analyse geographical information. Land records, maps, and even GIS are increasingly used by genealogical investigators[25]. If the geographic information is the main focus, then some different visualization techniques can be considered. Therefore, how to effectively extract and visualize geographical information is a topic deserved to be studied.

The Brabant Historical Information Center (BHIC) is the center for the history and the center for the genealogy of the province of Noord-Brabant [1]. It collects genealogical data of the Netherlands of past several centuries. It manages over 1,500 archives and collections (almost 40 kilometers of paper and parchment). It is a trusted source of historical information not only for individuals but also for organizations. Since the head office located in the Citadel in 's-Hertogenbosch was established, BHIC began to work on digitizing the paper records. More and more archive documents have been scanned and can be viewed as photos via their website. Meanwhile, the main information in the scanned documents are transformed to text in XML format. Now, many of scanned photos of documents and XML files can be viewed and downloaded on the website of BHIC.

The BHIC dataset in XML format contains more than 18 million personal records and covers a period of more than four centuries, the data model consists of five main parts in each dataset:

1 Person: The Person part is used to store the information on the deed about a person.

2 Event: The Event item is used to store the information on the record about an event.

3 Object: The Object part within can be used to store information of a nature other than specific personal or event information.

4 Source: The Source section concerns the source data such as record number, inventory number, etc.

5 Relation: The Relation component concerns the mutual relationships between persons, events and objects, such as Relation Event-Person, Relation Person-Object, and Relation Person-Person.

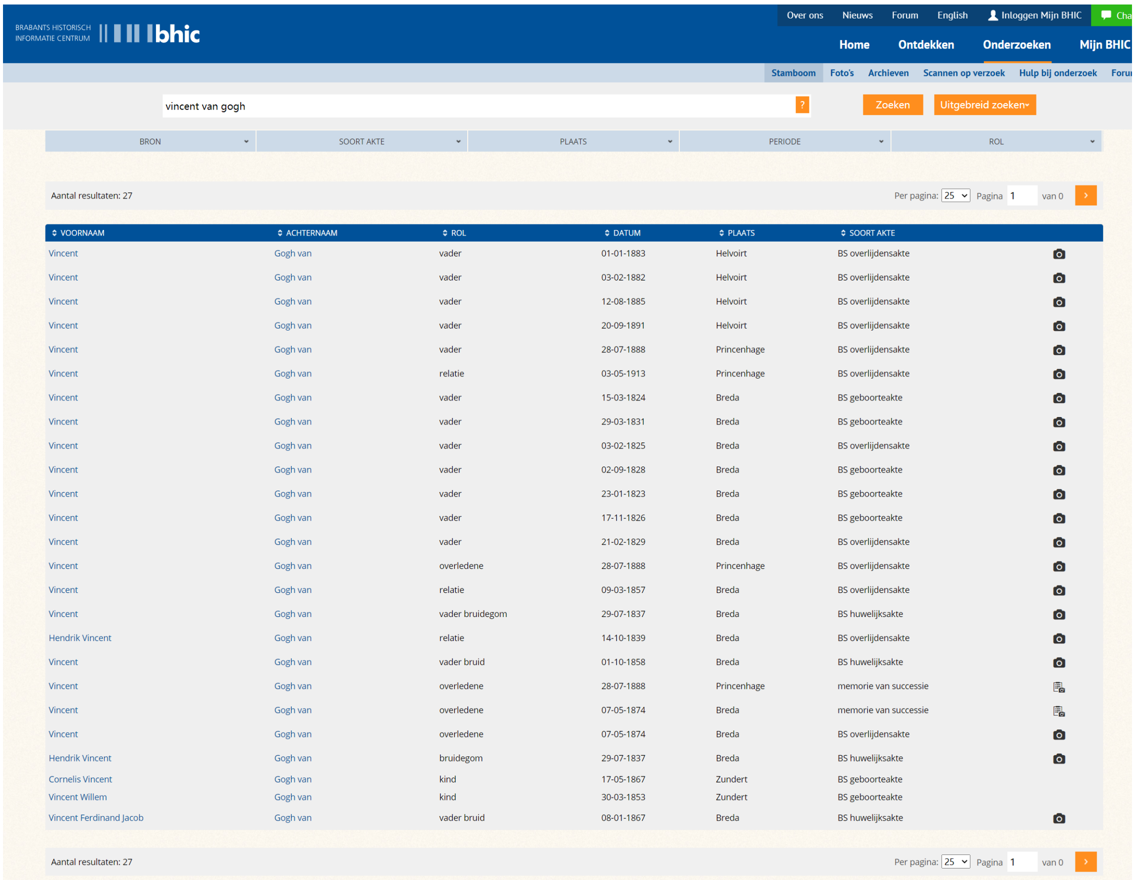
As object part will mainly be filled with additional information and does not have standard data types, we will not analyze the attribute in object. Also, as the source attribute is not of interested, the attribute in source part will also not be analyzed.

There are totally 10 datasets, which describe the event records in one's life. Three datasets are the records that collect from church, 'DTB Trouwen' is wedding certificate, 'DTB Dopen' is baptismal certificate, and 'DTB Begraven' is burial certificate. Three datasets are the civilian status that from the goverment, 'Burgerlijke Stand - Geboorte' is birth certificate, 'Burgerlijke Stand - Huwelijk' is marriage certificate, and 'Burgerlijke Stand - Overlijden' is death certificate. Four datasets records the different event, 'Bevolkingsregister' is population register, 'Memories van successie' is memories of succession, 'Militieregister' is militia register, and 'Gevangenisregister' is prison register. The names of the datasets, their structures, and their attributes in person, event and relation part are shown in Table 1.1:

Dataset Name	**Event**	**Relation**	**Person**
Memories of Succession	Id, type, date	Deceased	id, name, gender
Birth Certificate	Id, type, date	Child	id, name, gender, birth date
		Father and mother	Id, name, gender
Marriage Certificate	Id, type, date, place	Bride and groom	Id, name, birth date, birth place, gender
		Parents' of bride and groom	Id, name, gender
Death Certificate	Id, type, date, place	Relatives of deceased.	Id, name, gender
		Deceased	Id, name, gender
Militia Register	Id, type, duration, place	Militaiman	Id, name, birth date, birth place, gender
Prison Register	Id, type, duration, place	Prisoner	Id, name, birth date, birth place, gender
Wedding Certificate	Id, type, date, place, religion	Bride and groom	Id, name, birth place, gender
Population Register	Id, type, duration, place	Civilian	Id, name, birth date, birth place, gender
Burial Certificate	Id, type, date, place, religion	Father or relatives of deceased	Id, name, gender
		Deceased	Id, name, gender
Baptismal Certificate	Id, type, date, place, religion	Parents and witness	Id, name, gender
		Child	Id, name, gender

Table 1.1: Attributes of BHIC dataset

BHIC currently has no other way of letting people explore their database, other than by searching on specific names, dates, places, and the result is only a list of the related records. As shown in Figure 1.1, if we search the name 'Vincent van Gogh', all the records that related to this name is listed. BHIC experts would like to have an online tool to explore their data interactively. They want to dig into the data with various operations, such as select, filter, click, zoom, resulting in different diagrams, such as pie chart, map, timeline.



vincent van gogh

Aantal resultaten: 27 Per pagina: 25 Pagina 1 van 0

VOORNAAM	ACHTERNAAM	ROL	DATUM	PLAATS	SOORT AKTE	
Vincent	Gogh van	vader	01-01-1883	Helvoirt	BS overlijdensakte	📷
Vincent	Gogh van	vader	03-02-1882	Helvoirt	BS overlijdensakte	📷
Vincent	Gogh van	vader	12-08-1885	Helvoirt	BS overlijdensakte	📷
Vincent	Gogh van	vader	20-09-1891	Helvoirt	BS overlijdensakte	📷
Vincent	Gogh van	vader	28-07-1888	Princenhage	BS overlijdensakte	📷
Vincent	Gogh van	relatie	03-05-1913	Princenhage	BS overlijdensakte	📷
Vincent	Gogh van	vader	15-03-1824	Breda	BS geboorteakte	📷
Vincent	Gogh van	vader	29-03-1831	Breda	BS geboorteakte	📷
Vincent	Gogh van	vader	03-02-1825	Breda	BS overlijdensakte	📷
Vincent	Gogh van	vader	02-09-1828	Breda	BS geboorteakte	📷
Vincent	Gogh van	vader	23-01-1823	Breda	BS geboorteakte	📷
Vincent	Gogh van	vader	17-11-1826	Breda	BS geboorteakte	📷
Vincent	Gogh van	vader	21-02-1829	Breda	BS overlijdensakte	📷
Vincent	Gogh van	overledene	28-07-1888	Princenhage	BS overlijdensakte	📷
Vincent	Gogh van	relatie	09-03-1857	Breda	BS overlijdensakte	📷
Vincent	Gogh van	vader bruidegom	29-07-1837	Breda	BS huwelijksakte	📷
Hendrik Vincent	Gogh van	relatie	14-10-1839	Breda	BS overlijdensakte	📷
Vincent	Gogh van	vader bruid	01-10-1858	Breda	BS huwelijksakte	📷
Vincent	Gogh van	overledene	28-07-1888	Princenhage	memorie van successie	📄
Vincent	Gogh van	overledene	07-05-1874	Breda	memorie van successie	📄
Vincent	Gogh van	overledene	07-05-1874	Breda	BS overlijdensakte	📷
Hendrik Vincent	Gogh van	bruidegom	29-07-1837	Breda	BS huwelijksakte	📷
Cornelis Vincent	Gogh van	kind	17-05-1867	Zundert	BS geboorteakte	
Vincent Willem	Gogh van	kind	30-03-1853	Zundert	BS geboorteakte	
Vincent Ferdinand Jacob	Gogh van	vader bruid	08-01-1867	Breda	BS huwelijksakte	📷

Aantal resultaten: 27 Per pagina: 25 Pagina 1 van 0

Figure 1.1: Example of a search result of BHIC, searching 'Vincent van Gogh' and showing the related records of searched name

1.1 Objectives

In the communication with BHIC experts, they emphasize that they especially hope to focus on exploring geographic information. Mobility in geographical information deserves to be explored. When we explore the BHIC dataset, we found that although there is an attribute that records the person id in every record in every dataset, the same person would have different person ids when recorded in multiple times in the same or different datasets. As a result, if we want to identify the same person among the records, we have to define an interpretation to link them. Although it is theoretically possible to extract mobility information by linking different datasets, we stick to mobility information that is explicitly recorded. Specifically, we concentrate on the datasets that have two location names in their records. The eligible datasets are marriage certificate, militia register, prison register, wedding certificate, and population register.

However, the mobility of caused by militia register and prison register could be temporal nature and would not really change their residential places, so the mobility could be meaningless. In the population register dataset, we do not have the information on the places that the people move from or move to, so it is difficult to dig the population mobility from this dataset. Marriage certificate and wedding certificate both collect the data about marriage event. But they could

have overlapping records, and the overlapping records are difficult to be identified. Thus, instead of combining them together, we will choose one of these to visualize. The marriage certificate dataset has more records and information than the wedding certificate dataset. Thus, we reach an agreement to focus on the marriage certificate dataset to analyze marriage mobility. So the objective of this project is that:
"How to design a tool that can visualize and analyze marriage mobility in an effective and interactive way?"

In order to achieve this objective, the following questions need to be answered.

- What kind of information is recorded in the dataset?
- How to deal with incomplete and erroneous data?
- How to concurrently visualize a large number of marriage movements?
- How to visualize this multi-variate dataset?
- How to design a user-friendly dashboard and interaction model?

1.2 Thesis Structure

The remaining parts of this report is divided into 7 chapters. In Chapter 2, the attributes in the marriage dataset are introduced and the possible questions and requirements are discussed presented. In Chapter 3, we propose some existing visualization techniques for genealogical data and flow data, and discuss their advantages and limits.

In Chapter 4, the detailed information of each attribute and the pre-processing of different attributes are described. In Chapter 5, an overview of the solution and the details of the implementation are presented. Chapter 6 evaluates whether the tool meet the requirements, and shows some interesting results with several use cases. Finally, the conclusion of the project is given, and some possible future works are described in Chapter 7.

Chapter 2

Background

In this chapter, the project background are introduced and the requirements are presented.

2.1 Dataset Description

The dataset is a marriage registration dataset recorded by BHIC in the 19-th and 20-th centuries. The number of the extracted records is 484429. The dataset was accessed from the site of Opendata[2] on March 16th, 2020.

Each record in the dataset contains three kinds of information. The first kind is the information about the marriage event, which includes event id, event type, event date (year, month, day), and event place. The second kind is the information about relations, which records the identity of a person, including bride, groom, and mother and father of bride and groom. The third kind is the information about a person, including person id, person name (first name, prefix last name, last name), birth date (year, month, day), birthplace and gender of groom and bride. Also person id, person name (first name, prefix last name, last name) and gender of father and mother of groom and bride are included in the third kind.

2.1.1 Data Transformation

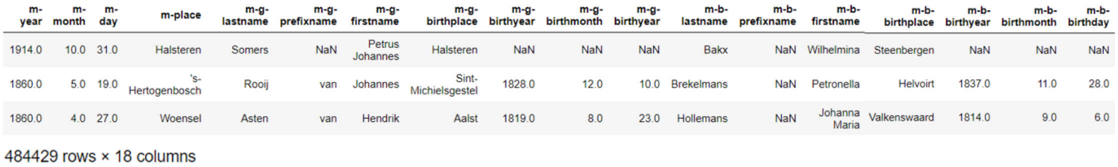
The file downloaded from the site was in XML format. XML is heavily used as a format for document storage and processing, which follows a nested structure. However, the redundancy and verbosity in the syntax of XML will cause higher processing time to access the data when the volume of data is large. Besides, it is difficult to filter or search for some specific records in a nested structure. Hence, to improve the readability and reduce the storage, we transform the XML files to CSV format.

As we would like to focus on the geographical information and the original dataset only offers location information for the marriage registration place and the birthplace of groom and bride, we will not consider the information about father and mother of groom and bride in this project. After the format transformation and data filtering, the file size was significantly reduced from about 133MB to about 83MB, which requires less compiling time and memory. The retained meaningful information in the generated CSV file is listed in the next section for further analysis.

2.1.2 Dataset Attributes

As shown in Figure 2.1, the dataset is cleaned up after the data transformation. There are a total of 484429 rows, each row represents a record that is a marriage movement. The column names correspond to the attributes contained in each marriage movement, and the descriptions of these attributes are as follows.

- Register Date: This attribute describes the register date, including year, month, and day, corresponding to columns named 'm-year', 'm-month', and 'm-day'.
- Register Place: This attribute describes the place that the marriage is registered, corresponding to the column named 'm-place'.
- Name: This attribute contains the first name, prefix last name, and last name of groom and bride. Groom's name corresponds to columns named 'm-g-lastname', 'm-g-prefixname', and 'm-g-firstname'. Bride's name corresponds to columns named 'm-b-lastname', 'm-b-prefixname', and 'm-b-firstname'.
- Birth Date: This attribute describes the birth date of groom and bride, including year, month, and day. Groom's birth date corresponds to columns named 'm-g-birthyear', 'm-g-birthmonth' and 'm-g-birthday'. Bride's birth date corresponds to columns named 'm-b-birthyear', 'm-b-birthmonth' and 'm-b-birthday'.
- Birth Place: This attribute describes the birth place of groom and bride. Groom's birth place corresponds to column named 'm-g-birthplace', and bride's birth place corresponds to column named 'm-b-birthplace'.



m-year	m-month	m-day	m-place	m-g-lastname	m-g-prefixname	m-g-firstname	m-g-birthplace	m-g-birthyear	m-g-birthmonth	m-g-birthyear	m-b-lastname	m-b-prefixname	m-b-firstname	m-b-birthplace	m-b-birthyear	m-b-birthmonth	m-b-birthday
1914.0	10.0	31.0	Halsteren	Somers	NaN	Petrus Johannes	Halsteren	NaN	NaN	NaN	Bakx	NaN	Wilhelmina	Steenbergen	NaN	NaN	NaN
1860.0	5.0	19.0	's-Hertogenbosch	Rooij	van	Johannes	Sint-Michielsgestel	1828.0	12.0	10.0	Brekelmans	NaN	Petronella	Helvoirt	1837.0	11.0	28.0
1860.0	4.0	27.0	Woensel	Asten	van	Hendrik	Aalst	1819.0	8.0	23.0	Hollemans	NaN	Johanna Maria	Valkenswaard	1814.0	9.0	6.0

484429 rows × 18 columns

Figure 2.1: Attributes in marriage dataset after transformation

2.2 Problem Description

To enable the users to comprehensively explore the dataset, multiple questions are presented below. In Section 2.2.2, the project requirements are summarized base on these questions.

2.2.1 Questions

The dataset was provided by BHIC. It collects genealogical data from individuals and aims to provide answers to questions of individuals, organizations, and institutions. BHIC works on making its data accessible to the public. However, currently, the public can only interact with the data in a very limited way, just by a simple search in a textual interface. In this case, a visualization tool may contribute a lot to the exploration of the data. With such a tool the public can explore the data more intuitively and hopefully may draw some interesting conclusions.

The marriage dataset collects the birthplace and marriage register place of bride and groom in each record. If we regard the mobility from their birthplace to marriage register place as marriage mobility, we can explore some specific mobility patterns and analyze the reason behind these phenomena and their impacts on the development of society.

Thus, the goal of this project is to develop a visualization-based web tool to help experts and the public to discover and interpret interesting patterns. Combining the information recorded in the marriage dataset, the following questions should be answered with the web tool.

1 How does marriage mobility change over time?

2 What kinds of patterns or anomalies can be discovered in the marriage mobility? Are they related to any historical event?

3 What is the marriage mobility of specific locations during a specific time period?

4 How are groups of locations with similar mobility patterns related to each other?

5 How does the number of people that married in or were born in specific locations change over time?

6 How does the age distribution of people when they get married change over time?

7 How is the age or birth date distribution influenced by different roles?

8 What is the marriage mobility of a specific family?

2.2.2 Requirement

We transform the questions in Section 2.2.1 to project requirements and motivate them. The project requirements are divided into two parts, functional requirements and non-functional requirements for the visualization tool. For the convenience of expression, we firstly introduce some terminologies. We regard birthplaces and marriage register places as locations. We regard the people that were born in one place and got married in another place as flows. The total number of people married in a location is the inflow of this location, and the total number of people who were born in a location is the outflow of this location. If someone was born and got married in the same place, then he/she was counted towards both inflow and outflow of this location.

Functional Requirements

The interface part defines how marriage mobility is displayed to help users browse the dataset, and the interaction part defines how the users could interact with the web tool.

- **Interface**

  A1 From question 1, the web tool should show the inflow and outflow of each location at different time periods in the same view.

  A2 From question 2, all the inflows and outflows of different locations should be shown at the same time.

  A3 From question 3, while users select some locations, the corresponding flows of selected locations should also be shown.

  A4 From question 4, groups of locations with similar mobility patterns should be identified and clustered, and the users should be enabled to see them.

  A5 From question 5, the tool should be able to show the change of the number of inflows and outflows of selected locations over time.

  A6 As the dataset does not have attribute of age, we use the distribution of birth date to calculate the distribution of age. From question 6 and 7, the tool should be able to show the change of distribution of gender, birth date, and age of selected locations over time.

- **Interaction**

  B1 **Filter** The visualized data in the tool should be able to be filtered according to multiple conditions. First, in some questions, we expect to observe the marriage mobility during a specific time period. Thus, the data should be able to be filtered by marriage time. Second, from question 8, the data should be able to be filtered by name. Moreover, for other attributes in the dataset, the data should also be filtered by these.

  B2 **Basic interactions** The tool should contains the basic interactions for users. For example, zooming, panning and switching multiple map layers in the map.

  B3 **Multiple locations selection** The users should be able to conveniently and quickly select multiple locations to observe the patterns they are interested in.

Non-functional Requirements

Both experts and the general public are the target users. For reasons of usability, some constraints and conditions should be paid attention to in the design and implementation of the tool.

C1 **Authenticity** The marriage dataset is a historical dataset. To keep the the authenticity of the dataset, we should do minimal operations and interpretations to the dataset.

C2 **Clean interface** Visualizing all attributes in the same view will cause redundant information and confusion of the users. Therefore, the attributes should be visualized in different views, and the design of the layout of these views should be clean.

C3 **Real-time response** There are almost half a million records in the dataset, so the performance of the tool should be considered. The loading time to show the initial page should be within seconds, and the interactions should be responded to in real-time.

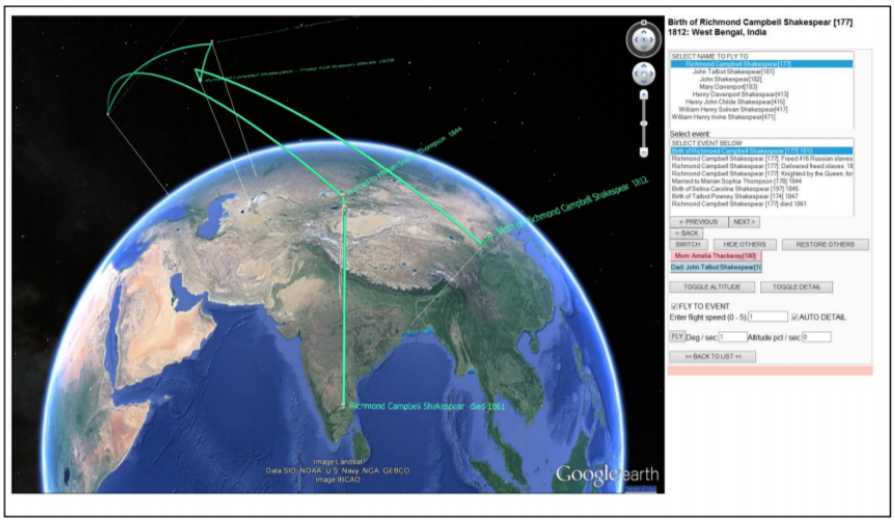
Chapter 3

Related Work

In this chapter, we introduce some visualization methods that might be applied in this project, and analyse their advantages and shortcomings. We start with the visualization techniques in the genealogical analysis, and then explain more specific techniques for origin-destination data.

3.1 Genealogical Data Analyses and Visualizations

Genealogy data can be applied in many fields and visualized in multiple ways. For example, in the field of genetic medicine, Nobre and Gehlenborg [19] introduced a linearization approach for visualizing multivariate trees and tree-like graphs to analyse hereditary diseases. In the field of linguistics, Rohrdantz and Hund [22] presented an extended sunburst visual display to analyse the language genealogy by comparing different features of languages.


Richmond Campbell Shakespear died 1861
Birth of Richmond Campbell Shakespear 1812
Image Landsat
Data SIO, NOAA, U.S. Navy, NGA, GEBCO
Image IBCAO
Google earth
SELECT NAME TO FLY TO
Richmond Campbell Shakespear[177]
John Talbot Shakespear[181]
John Shakespear[182]
Mary Davenport[183]
Henry Davenport Shakespear[413]
Henry John Childe Shakespear[415]
William Henry Sulivan Shakespear[417]
William Henry Irvine Shakespear[471]
Select event:
SELECT EVENT BELOW
Birth of Richmond Campbell Shakespear [177] 1812
Richmond Campbell Shakespear [177]: Freed 416 Russian slaves
Richmond Campbell Shakespear [177]: Delivered freed slaves 18
Richmond Campbell Shakespear [177]: Knighted by the Queen, for
Married to Marian Sophia Thompson [178] 1844
Birth of Selina Caroline Shakespear [197] 1845
Birth of Talbot Powney Shakespear [174] 1847
Richmond Campbell Shakespear [177] died 1861
<- PREVIOUS
NEXT >
<- BACK
SWITCH
HIDE OTHERS
RESTORE OTHERS
Mom: Amelia Thackeray[180]
Dad: John Talbot Shakespear[1
TOGGLE ALTITUDE
TOGGLE DETAIL
FLY
>> BACK TO LIST <<

Birth of Richmond Campbell Shakespear [177]
1812: West Bengal, India

☑ FLY TO EVENT
Enter flight speed (0 - 5): 1 ☑ AUTO DETAIL
Deg / sec: 1 Altitude pct / sec: 0

Figure 3.1: Interactive Genealogy Explorer showing the lifeline
Source: Shakespear.2015 [24]

Since the BHIC datasets are genealogical data about population, we mainly focus on the visualization of genealogy of population. The genealogical data without geographical information is usually visualized by tree. For example, in order to perform the analysis of genealogical graphs of large families, McGuffin and Balakrishnan [18] introduced a dual tree, which is a subgraph formed by the union of two trees. In terms of the genealogical data with geographical information, there are multiple choices for the visualization. For example, using tree view as the main view, Liu and

Dai [17] developed a system called GenealogyVis. They implement a tree view and a stream view in the system to explore the development of families related to structure, population, migration, and other demographic information. As shown in Figure 3.1, using map view as the main view, Shakespear [24] developed a three-dimensional graphic interface to explore the biographical data of individuals and migration of person in the genealogy. The tool could show the lifelines of individual in a global map and the details of event are listed beside the map.

Compare to other methods, the geographical information in the map view can be quickly interpreted by the users. Thus, for the geographical information in our dataset, we determine to use the map view as the main view, and show other details in different views.

3.2 Visualization of Origin-Destination Data

The geographical data in the marriage dataset includes the birth place and the marriage register place of bride and groom. Such data, which contains details of two or more geographic points, is regarded as origin-destination data or flow data. In this section, according to the characteristics of our geographical information, we describe the visualization methods for origin-destination data.

3.2.1 Node-link diagram

The node-link diagram, which represent locations as nodes and movements as edges, is a direct way to visualize the movability. For the visual clutter problem caused by a large number of nodes and edges, Holten [15] proposed the use of edge bundling to remedy this and reveal high-level edge patterns. Yang and Dwyer [28] presented that the use of the third spatial dimension also can resolve visual clutter in complex flow maps.

Figure 3.2 shows a US migration graph with 1715 nodes and 9780 edges. Graph (a) is the unbundled result, (b), (c), and (d) are the results of using different methods to bundle. The advantages and disadvantages among methods in Figure 3.2 (b), (c) and (d) will not be discussed in detail here. In conclusion, edge-bundling improves the readability of the node-link diagram. there still exist some limits in the node-link diagram with edge bundling. However, the bi-directional edges can not be very readable, and the width of edges is not evident after edge-bundling.

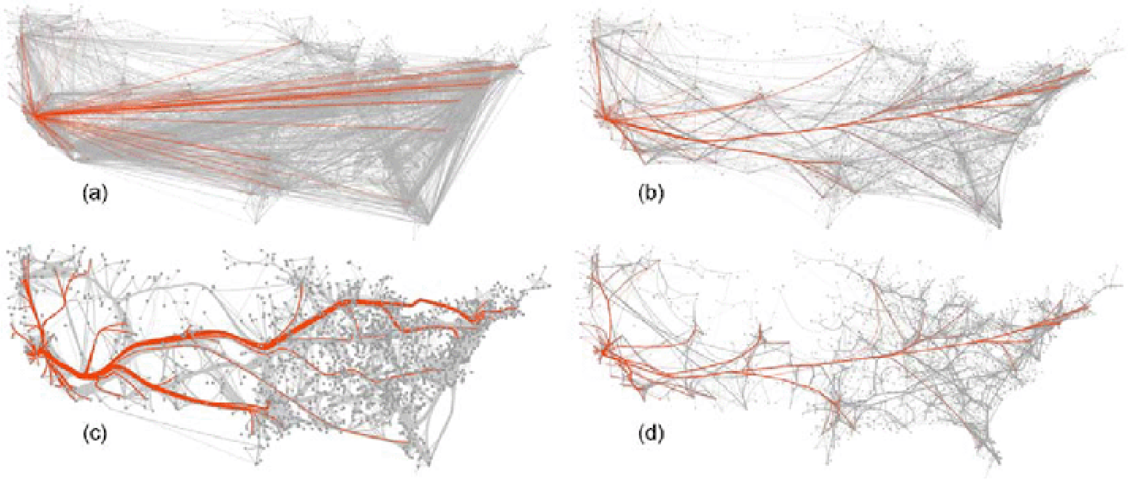


Figure 3.2: US migration graph (1715 nodes, 9780 edges)(a) not bundled and bundled using (b) FDEB with inverse-linear model, (c) GBEB, and (d) FDEB with inverse-quadratic model. The same migration flow is highlighted in each graph.
Source: Holten.2009 [15]

3.2.2 Choropleth map

Another method that can visualize geographical data is a choropleth map, which areas are shaded or patterned in the proportion to a statistic variable that represents an aggregate summary of a geographic characteristic within each area. For the representation of movements, the color can be defined as the total number of movements, and the flows between different area can be shown while clicking specific area. As shown in Figure 3.3Besançon and Cooper [12] compared various improved methods based on the choropleth map. In conclusion, choropleth map can give visible results about how movement is varied over space. However, the precise boundaries of different regions are necessary. Besides, if the values are cluttered, it can be difficult to distinguish between different shades.

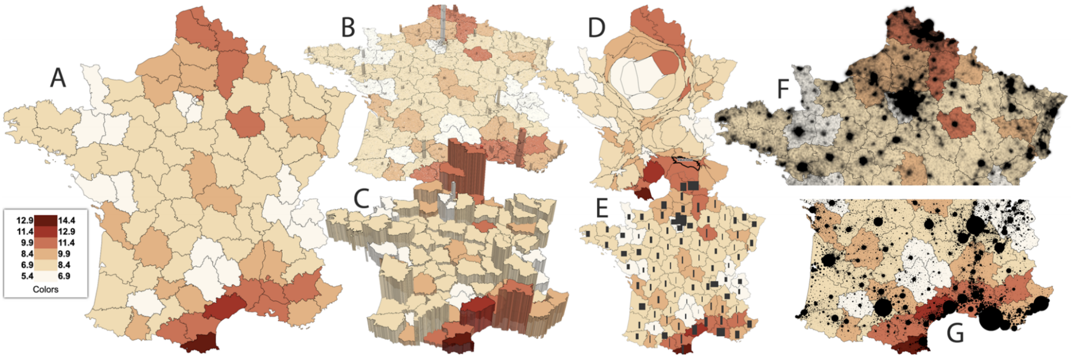


Figure 3.3: Choropleth map (A) augmented with 3D extrusion (C), contiguous cartogram (D), and rectangular glyphs (E) at the same level of granularity and with 3D extrusion (B), Heatmap (F) and dot map (G) at a finer level of granularity.
Source: Besançon and Cooper.2020 [12]

3.2.3 Matrix

As shown in Figure 3.4, columns and rows in the matrix diagram usually represent start points and end points, respectively. Meanwhile, cells represent the value of the corresponding row and column to show whether two nodes are connected. In the movement matrix, boolean values are replaced by numbers to provide more information, and the color of cells represents the number of people move from one location to another.

Ghoniem and Fekete[14] compared the time of finishing various tasks in the node-link diagram and the matrix diagram. Compared to the node-link diagram in the level of 50 nodes and 4000 edges, the matrix diagram is more suitable for large or dense graphs because of its clear layout and superior readability. However, for a large number of nodes and edges, such as hundred of nodes, the matrix diagram is still not readable, and the geographical information and relations between nodes are hidden.

3.2.4 OD Maps

Wood and Dykes[27] proposed an OD(Origins and destinations) map to represent the flows in geographic space. They introduce OD vectors as cells rather than lines, which is similar to the idea of the matrix but reserve the start points' and end points' spatial layout.

As shown in Figure 3.5, OD maps represent all 721,432 US county–county migration vectors. The US map are divided into different cells ,and the cell that a flow belongs to is decided by the coordinate of the center of the origin and destination. The large grid cells represent origin locations, small cells inside represent destination locations, and the color of cells represents the absolute number of movements.

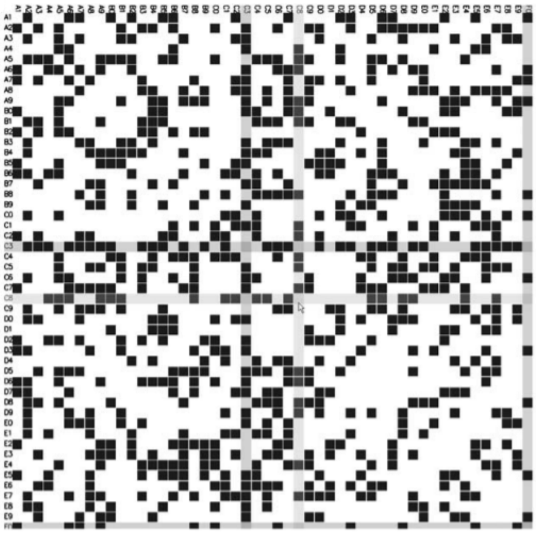


Figure 3.4: An undirected graph(50 nodes and 400 edges) in matrix representation.
Source: Ghoniem and Fekete.2005[14]

OD map is suitable for a large number of movement data(of order $10^5 \sim 10^6$). Similar to the choropleth map, it could be difficult to distinguish between different shades if the values are cluttered. Also, the OD map is suitable for visualizing the movements within a country. As there are many foreign locations in the marriage dataset, OD map can not be used in our visualization.

3.2.5 Circular plot

Nikola and Guy J. [23] introduced a circular plot to visualize and analyze the movement flow intuitively and interactively. Figure 3.6 shows an interactive circular plot of migration flows in different continents, and each chord represents a one-directional flow between two continents. Users can expand to the country level by clicking the different continents. Although the width of the chord clearly reveals the number of movements, the geographical information of different continents and countries cannot be captured. Besides, the form of circle limits the display of the number of locations.

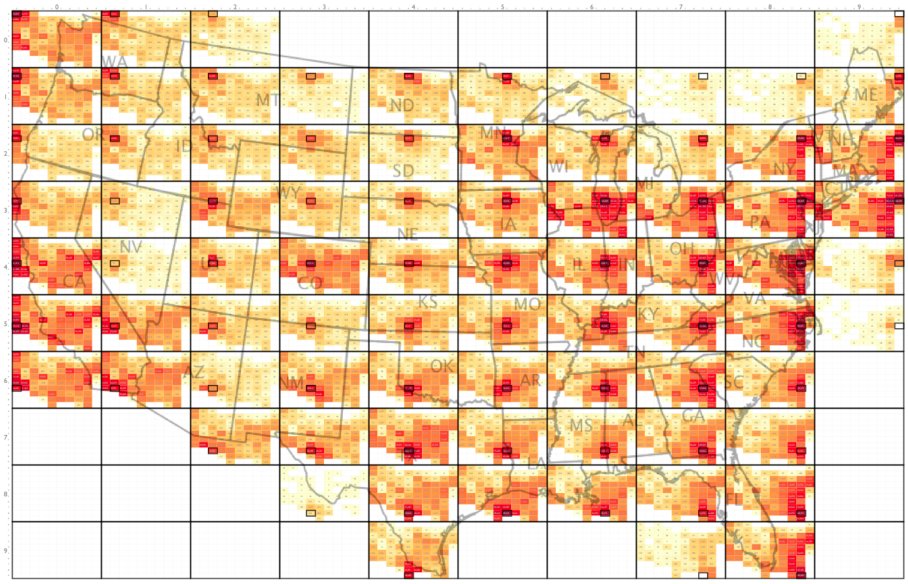


Figure 3.5: OD map that shows US county–county migration vectors
Source: Wood and Dykes.2010[27]

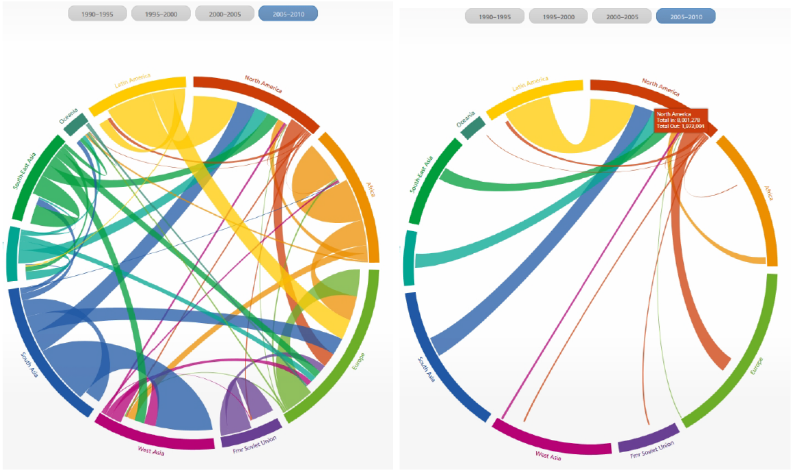


Figure 3.6: Interactive circular plot of migration flows: left - flows between each continents, right - flows between North America and other continents
Source: Nikola and Guy J.2014 [23]

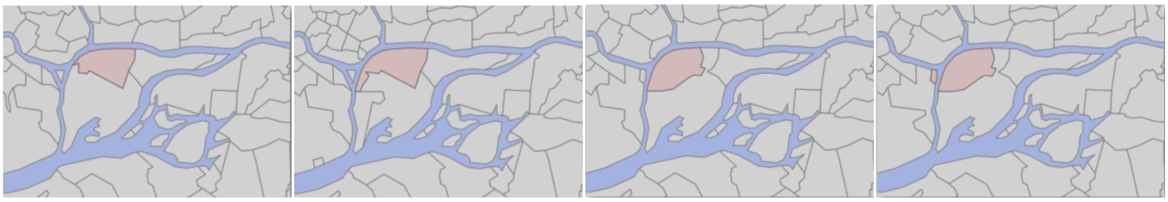
Chapter 4

Data Preprocessing

In Chapter 3, we have chosen to use a map as the main view of the visualization tool. From the related work of flow data, we can conclude that there are two methods, which are borderlines and points, to represent locations in the map. Since the marriage dataset is a historical dataset in 19-th and 20-th centuries, it is a challenge to represent historical locations in the map.

We first analyse the possibility of using borderlines to represent locations. Gemeentegeschiedenis [3] provides the borderlines of all the historical municipalities in The Netherlands from 1812 to 2019. However, if we use the borderlines of municipalities in a specific year to represent all the borderlines in different years in the marriage dataset, the error of historical locations caused by borderlines will be relatively large.

Besides, the borderlines of a specific municipality changed over time. An example of the borderlines of Dordrecht is shown in Figure 4.1. The uncertainty of the borderlines cannot be easily encoded in the map. Moreover, Gemeentegeschiedenis can only provide the borderlines of a specific municipality in a specific year, so the map of The Netherlands can only be constituted by downloading all the individual municipalities. Hence, the borderlines of the municipalities of a specific year all over the Netherlands are not available to access from Gemeentegeschiedenis.



(a) In 1812 (b) In 1841 (c) In 1872 (d) In 1903

Figure 4.1: Borderlines(in red) of Dordrecht in different years

Since it is difficult to use borderlines to represent locations, we consider to represent locations using points. In this case, the centre of the point is usually the coordinate of centre of the borderlines. Thus, the possibility of overlapping coordinates is smaller than the possibility of overlapping areas. Moreover, we use the coordinates of the locations in external datasets, which are the coordinates of a specific year. Hence, the influence of the change of the borderlines is reduced.

As the marriage dataset is a historical dataset inputted from paper records, there are many incomplete and inaccurate data that can cause confusion. Figure 4.2 illustrates a flow chart for the data preprocessing in our project. The incomplete data are caused by missing data, and multiple attributes have missing data. Thus, in the Section 4.1 of this chapter, we first analyse the distribution and the influence of the missing data, and indicate how we handle the missing data. Also, some attributes have inaccurate data. Different from missing data, inaccurate data can be caused by multiple sources, such as wrongly spelled in the original document, wrongly copied from the original document. Based on the principle of representing locations as points, we match the locations with coordinates using some external datasets in Section 4.2. The inaccurate data in the locations are resolved in several steps and the result is given. And according to the principle of keeping authenticity, inaccurate data for other attributes remain untouched.

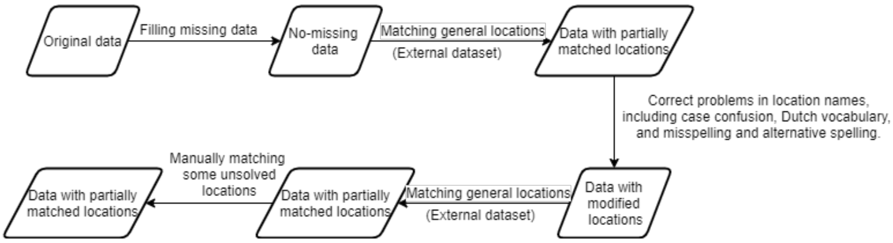

Original data
Filling missing data
No-missing data
Matching general locations
(External dataset)
Data with partially matched locations
Correct problems in location names, including case confusion, Dutch vocabulary, and misspelling and alternative spelling.
Data with modified locations
Matching general locations
(External dataset)
Data with partially matched locations
Manually matching some unsolved locations
Data with partially matched locations

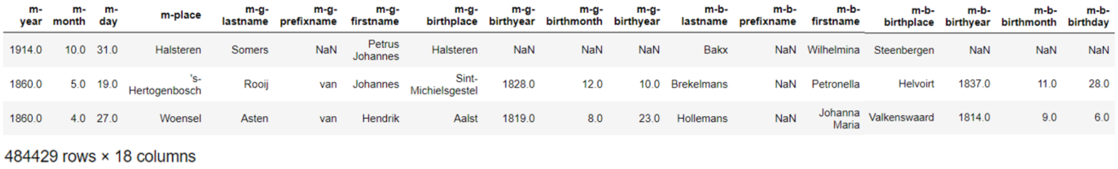
Figure 4.2: Flow diagram of Chapter 4

4.1 Missing data

In this section, the distribution and the influence of missing data are described, and a solution is given.

4.1.1 Distribution and Influence

The attributes of the dataset are shown in Figure 4.3. The columns describe, respectively, the marriage date and location, the name, the birth place and the birth date of bride and groom. The missing data in each attribute is counted. The result reveals that there are only two records whose marriage dates are missing. Thus, the columns that are related to marriage information, including 'm-year', 'm-month', 'm-day', and 'm-place' have approximately no missing data. Also, there is no missing data in the first and last names of bride and groom, including 'm-g-lastname', 'm-g-firstname', 'm-b-lastname', and 'm-b-firstname'.



m-year	m-month	m-day	m-place	m-g-lastname	m-g-prefixname	m-g-firstname	m-g-birthplace	m-g-birthyear	m-g-birthmonth	m-g-birthyear	m-b-lastname	m-b-prefixname	m-b-firstname	m-b-birthplace	m-b-birthyear	m-b-birthmonth	m-b-birthday
1914.0	10.0	31.0	Halsteren	Somers	NaN	Petrus Johannes	Halsteren	NaN	NaN	NaN	Bakx	NaN	Wilhelmina	Steenbergen	NaN	NaN	NaN
1860.0	5.0	19.0	's-Hertogenbosch	Rooij	van	Johannes	Sint-Michielsgestel	1828.0	12.0	10.0	Brekelmans	NaN	Petronella	Helvoirt	1837.0	11.0	28.0
1860.0	4.0	27.0	Woensel	Asten	van	Hendrik	Aalst	1819.0	8.0	23.0	Hollemans	NaN	Johanna Maria	Valkenswaard	1814.0	9.0	6.0

484429 rows × 18 columns

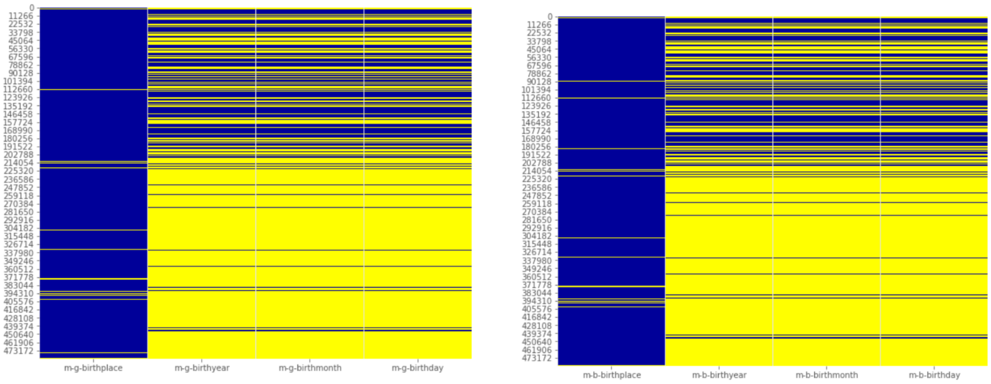
Figure 4.3: Dataset Attribute

The percentages of missing data for the remaining columns are listed as follows.

- m-g-prefixname - 61.0%
- m-g-birthplace - 1.0%
- m-g-birthyear - 72.0%
- m-g-birthmonth - 72.0%
- m-g-birthday - 72.0%
- m-b-prefixname - 61.0%
- m-b-birthplace - 2.0%
- m-b-birthyear - 73.0%
- m-b-birthmonth - 73.0%
- m-b-birthday - 73.0%

Since not every name has a prefix, the missing data in the prefix name is reasonable. From the presentatage of missing data above, we can conclude that 39% of the names apparently have prefixes, and the missing data in the prefix name will have no influence in further analysis.

For other attributes with missing data, we sorted the records according to increasing order according to the birth year, month and day. The distribution of the missing data in the remaining columns is shown in Figure 4.4, where (a) shows the distribution of information about the groom and (b) shows the distribution of information about the bride. The X-axis represents feature names, and the Y-axis represents the index of rows. Yellow color represents the missing data in the rows, and blue color represents non-missing ones. We can conclude that the number of missing birth date is increasing as the marriage date is increasing, and the distribution of missing in the birthplace has no strong pattern.



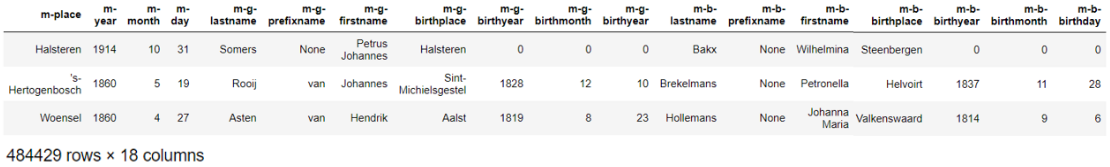
(a) Missing data of groom

(b) Missing data of bride

Figure 4.4: Distribution of missing data, X-axis-feature names, Y-axis-index of rows, yellow-missing data, blue-non-missing data

4.1.2 Solution

In the requirements, we have mentioned that we should keep authenticity as much as possible. Thus, the solution of missing data is that we only replace them with '0' or 'None' in order to mark and filter out the missing data when visualizing attributes with missing values. The missing values in 'm-year', 'm-month', 'm-day', 'm-g-birthyear', 'm-g-birthmonth', 'm-g-birthday', 'm-b-birthyear', 'm-b-birthmonth' and 'm-b-birthday' are filled with '0', and missing values in 'm-g-birthplace' and 'm-g-birthplace' are filled with 'None'. After replacing the empty numerical values as '0', and string empties as 'None', the dataset is shown in Figure 4.5.



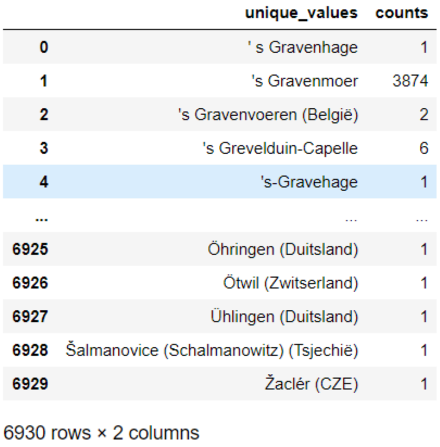
m-place	m-year	m-month	m-day	m-g-lastname	m-g-prefixname	m-g-firstname	m-g-birthplace	m-g-birthyear	m-g-birthmonth	m-g-birthyear	m-b-lastname	m-b-prefixname	m-b-firstname	m-b-birthplace	m-b-birthyear	m-b-birthmonth	m-b-birthday
Halsteren	1914	10	31	Somers	None	Petrus Johannes	Halsteren	0	0	0	Bakx	None	Wilhelmina	Steenbergen	0	0	0
's-Hertogenbosch	1860	5	19	Rooij	van	Johannes	Sint-Michielsgestel	1828	12	10	Brekelmans	None	Petronella	Helvoirt	1837	11	28
Woensel	1860	4	27	Asten	van	Hendrik	Aalst	1819	8	23	Hollemans	None	Johanna Maria	Valkenswaard	1814	9	6

484429 rows × 18 columns

Figure 4.5: Dataset attribute after filling missing data

4.2 Inaccurate Data in Locations

We extract the unique values of the columns that contain location names: register places, birthplaces of bride and groom; they correspond to 'm-place', 'm-g-birthplace', and 'm-b-birthplace' in Figure 4.1. As shown in Figure 4.6, there are 6930 distinct values in locations.



	unique_values	counts
0	' s Gravenhage	1
1	's Gravenmoer	3874
2	's Gravenvoeren (België)	2
3	's Grevelduin-Capelle	6
4	's-Gravehage	1
...	...	...
6925	Öhringen (Duitsland)	1
6926	Ötwil (Zwitserland)	1
6927	Ühlingen (Duitsland)	1
6928	Šalmanovice (Schalmanowitz) (Tsjechië)	1
6929	Žaclér (CZE)	1

6930 rows × 2 columns

Figure 4.6: Distinct values in locations and their frequency

To represent the locations as points in the map, the geographical coordinates of each location are needed. The locations in the marriage dataset include village and municipality names in The Netherlands and other countries. To map locations to their geographical coordinates, we find several publicly available datasets that have location names in The Netherlands and other countries. The location names in the marriage dataset that can be matched to a corresponding location name in such a dataset can be represented as points in the map. These datasets are introduced in section 4.2.1.

The records in the marriage dataset were manually recorded on paper and then digitized. Currently, there are some location names in the marriage dataset that cannot be directly matched with the location names in the external datasets. The first reason is that the unified location names in 19-th and 20-th centuries could be different from the ones of today. Second, different kinds of human errors, such as wrongly copied from the paper documents, could occured. As shown in the flow diagram in Figure 4.2, in Section 4.2.2, we matched part of locations. In Section 4.2.3, we discuss and classify the unsolved problems when matching. Finally, in Section 4.2.4, we divide the inaccurate data into two kinds of locations: the locations with coordinates and locations without coordinates.

4.2.1 External Datasets

In this section, we introduce three datasets used to match coordinates for locations in the marriage dataset, including a village dataset, a municipality dataset, and a country dataset.

Village Dataset

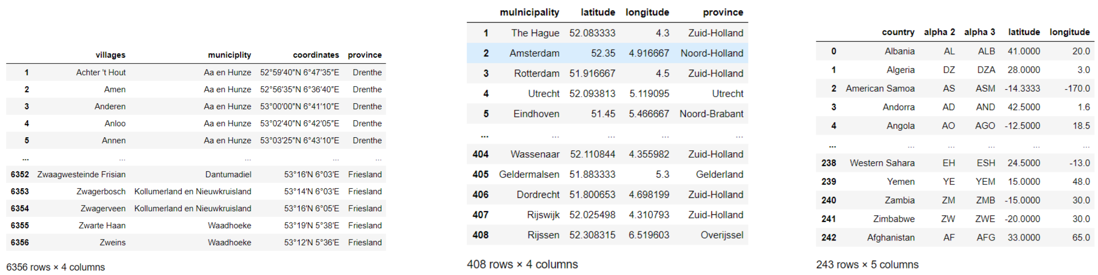
For mapping the locations in the marriage dataset to their geographical coordinates, we introduce a new dataset, as shown in Figure 4.7 (a). The village dataset includes the village names in the Netherlands, the municipality and province that they belong to, and their coordinates. This information is collected from NGA GEOnet Names Server (GNS)[4], and the dataset is provided in 2006 to store the official repository of standard spellings of all geographic names.

Municipality Dataset

For mapping municipality names in the Netherlands in the marriage dataset with their geographical coordinates, we introduce a new dataset, as shown in Figure 4.7 (b). The municipality dataset includes the municipality names in the Netherlands, the province that they belong to, and their coordinates. This dataset has been merged, cleaned, and formatted by simplemaps [5] in 2020.

Country Dataset

As some foreign locations have village names as well as country names in multiple forms, while others only have country names, it is challenging to locate specific coordinates for villages outside The Netherlands. Therefore, we uniformly use the country's coordinates to map such locations. Figure 4.7 (c) shows a country dataset, including the country names, their alpha-2 code, alpha-3 code, and their coordinates. Alpha-2 codes are two-letter country codes defined by ISO standard to represent countries, and alpha-3 codes are three-letter country codes defined to represent countries. There are some records that represent the country names with alpha-2 or alpha-3 code, so these codes are included in the country dataset for matching. The dataset is downloaded from github [6], and it is provided in 2019.



	villages	municiplity	coordinates	province
1	Achter 't Hout	Aa en Hunze	52°59′40″N 6°47′35″E	Drenthe
2	Amen	Aa en Hunze	52°56′35″N 6°36′40″E	Drenthe
3	Anderen	Aa en Hunze	53°00′00″N 6°41′10″E	Drenthe
4	Anloo	Aa en Hunze	53°02′40″N 6°42′05″E	Drenthe
5	Annen	Aa en Hunze	53°03′25″N 6°43′10″E	Drenthe
...	...	...	...	...
6352	Zwaagwesteinde Frisian	Dantumadiel	53°16′N 6°03′E	Friesland
6353	Zwagerbosch	Kollumerland en Nieuwkruisland	53°14′N 6°03′E	Friesland
6354	Zwagerveen	Kollumerland en Nieuwkruisland	53°16′N 6°05′E	Friesland
6355	Zwarte Haan	Waadhoeke	53°19′N 5°38′E	Friesland
6356	Zweins	Waadhoeke	53°12′N 5°36′E	Friesland

6356 rows × 4 columns

(a) Village names and their coordinates in The Netherlands

	mulnicipality	latitude	longitude	province
1	The Hague	52.083333	4.3	Zuid-Holland
2	Amsterdam	52.35	4.916667	Noord-Holland
3	Rotterdam	51.916667	4.5	Zuid-Holland
4	Utrecht	52.093813	5.119095	Utrecht
5	Eindhoven	51.45	5.466667	Noord-Brabant
...	...	...	...	...
404	Wassenaar	52.110844	4.355982	Zuid-Holland
405	Geldermalsen	51.883333	5.3	Gelderland
406	Dordrecht	51.800653	4.698199	Zuid-Holland
407	Rijswijk	52.025498	4.310793	Zuid-Holland
408	Rijssen	52.308315	6.519603	Overijssel

408 rows × 4 columns

(b) Municipality names and their coordinates in The Netherlands

	country	alpha 2	alpha 3	latitude	longitude
0	Albania	AL	ALB	41.0000	20.0
1	Algeria	DZ	DZA	28.0000	3.0
2	American Samoa	AS	ASM	-14.3333	-170.0
3	Andorra	AD	AND	42.5000	1.6
4	Angola	AO	AGO	-12.5000	18.5
...	...	...	...	...	...
238	Western Sahara	EH	ESH	24.5000	-13.0
239	Yemen	YE	YEM	15.0000	48.0
240	Zambia	ZM	ZMB	-15.0000	30.0
241	Zimbabwe	ZW	ZWE	-20.0000	30.0
242	Afghanistan	AF	AFG	33.0000	65.0

243 rows × 5 columns

(c) Country names and their coordinates

Figure 4.7: External datasets

4.2.2 Solved Mapping Problems

In this section, we introduce the directly matched names and solved problems in location names when matching the location names in the marriage dataset with the location names in the external datasets.

Directly matched locations

Among the location names in the marriage dataset, there are location names that contain village names or municipality names in The Netherlands, and alpha-2, alpha-3, or country names. Hence, these names can be directly matched to location names in the external datasets. For such names, we directly extract those names from locations and match them with the coordinates in the external datasets. Table 4.1 shows some example location names in the BHIC dataset and their corresponding names in the external dataset. Aalst and 's Gravenmore are village names in the Netherlands, Aalten and Amsterdam are municipality names in The Netherlands, IT and DE are alpha-2 code of country names, DEU and HUN are alpha-3 code of country names, and China and India are country names.

Table 4.1: General locations

Locations	Corresponding names
’s Gravenmoer	’s Gravenmoer
Aalst NB	Aalst
Aalten	Aalten
Westerkerk (Amsterdam)	Amsterdam
Capannori(IT)	Italy
Dülber (DE)	Germany
Canthop DEU	Germany
Bremberg(HUN)	Hungary
Canten(China)	China
Bombay(India)	India

Case confusion

The use of uppercase and lowercase letters in the location names may cause case confusion. So when we match village names, municipality names, and country names, we ignore the uppercase and lowercase letters. Table 4.2 shows some examples of case confusion and their corresponding names.

Table 4.2: Case confusion

Locations	Corresponding names
loon op Zand	Loon op Zand
oSS	Oss
helmond	Helmond
hEUSDEN	Heusden
leerdam	Leerdam

Dutch vocabulary

For the location names in the marriage dataset, there are many country names that are spelled in Dutch, while in the country dataset, all the country names are spelled in English. To solve this problem, we manually create a list to record their Dutch spellings and English spellings. Table 4.3 shows some examples of country names in Dutch in the marriage dataset and their corresponding country names in English in the manual list. For the full manual dutch vocabulary list, please see Table A.1 in Appendix A.

Table 4.3: Dutch vocabulary

Locations	Corresponding names
Frankrijk	France
België	Belgium
Groot-Brittannië	United Kingdom
Duitsland	Germany
Zweden	Sweden

Misspelling and Alternative Spelling

There are misspellings and alternative spellings in the location names in the marriage dataset. As manually checking all the misspellings and alternative spellings is a huge workload, we only calculate all location names that are one letter different from existing location names in the external datasets and manually filter the possible misspelling and alternative cases. In this case, some misspellings and alternative spellings are identified. Table 4.4 shows some examples of misspellings and alternative spellings, and their corresponding names in the external datasets. For example, 'Roetterdam' is the misspelling of 'Rotterdam', and 'Slabroeck' is an alternative spelling of 'Slabroek'. For the full manual misspelling list, please see Table A.2 in Appendix A.

Table 4.4: Misspelling

Locations	Corresponding names
Rijkwijk	Rijswijk
Roetterdam	Rotterdam
Tuilburg	Tilburg
Slabroeck	Slabroek

4.2.3 Unsolved mapping problems

Although we resolved some problems that can cause mismatching, there are still some location names that can not be matched with the location names in the external datasets. The reasons are summarized as follows.

- Variation in spelling: As shown in Table 4.5, there are some different expressions of village names. For example, in the marriage dataset, there is a location named " s Gravenhage', which corresponds to a village called Den Haag in external datasets, and a location named 'St. Michiels-Gestel', which corresponds to a village called 'Sint-Michielsgestel' in external datasets.

- Irregular abbreviations of country names: As shown in Table 4.6, There are some abbreviations of country names that cannot be matched with any international standardization. For example, 'D' or 'Dld' sometimes refer to Germany, and 'B' refers to Belgium.

- Several villages in one location name: As shown in Table 4.7, the location names that contain several villages' names cannot be matched to a name in the external datasets.

- Historical names: As shown in Table 4.8, there are some location names, which is a village or a municipality in 19-th and 20-th centuries, but does not exist in current village or municipality names in The Netherlands. For example, Maarsseveen was a village that existed from 1815 to 1949, Haskerland was a village until 1984, and Hazerswoude was a municipality until 1991.

- Unclear names: As shown in Table 4.9, some location names are not easily linked to existing location names, such as Bude and Hooven.

Table 4.5: Variation in spelling

Location names
' s Gravenhage
St. Michiels-Gestel
St Michielsgestel

Table 4.6: Irregular abbreviations of country names

Location names
Dietingen (D)
Beeck (Dld)
Loere (B)

Table 4.7: Several villages in one location name

Location names
Eethen, Genderen en Heesbeen
Maren en Kessel
Velthoven en Meervelhoven

Table 4.8: Historical names

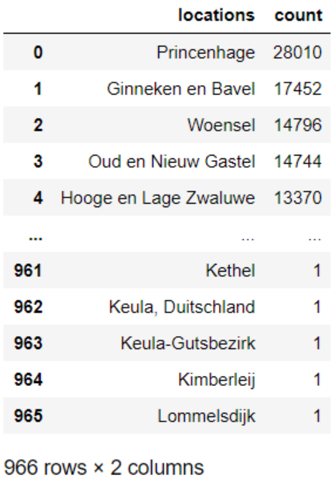
Location names
Maarsseveen
Haskerland
Hazerswoude

Table 4.9: Unclear names

Location names
Bude
Hooven

4.2.4 Result of Mapping

There are initially a total of 6930 distinct values in location names. After multiple preprocessing steps, 5964 locations names can be matched with location names in the external datasets and can be represented as points on the map, while other 966 location names cannot be matched. As shown in Figure 4.8, we calculate the unmatched location names and their frequency, and ordered them by the frequency.



	locations	count
0	Princenhage	28010
1	Ginneken en Bavel	17452
2	Woensel	14796
3	Oud en Nieuw Gastel	14744
4	Hooge en Lage Zwaluwe	13370
...	...	...
961	Kethel	1
962	Keula, Duitschland	1
963	Keula-Gutsbezirk	1
964	Kimberleij	1
965	Lommelsdijk	1

966 rows × 2 columns

Figure 4.8: Unmatched location names

The total frequency of all location names is 1937716, while the total frequency of unmatched location names is 344019, which is 17.75% of the total. As the percentage of unmatched names is relatively high, we manually match 123 unknown location names with high-frequency. Table 4.10 shows some examples of the result of manually matching.

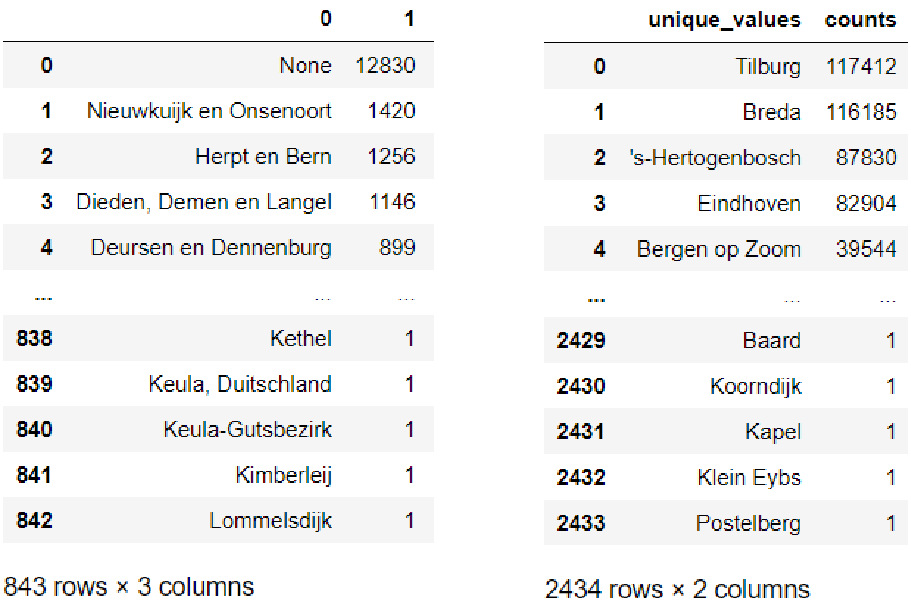
Table 4.10: Manually matching

Locations	Corresponding names
Princenhage	Breda
Ginneken en Bavel	Breda
Woensel	Eindhoven
Oud en Nieuw Gastel	Halderberge

After manually matching, there are still 823 location names that cannot be matched with location names in the external datasets, which are shown in Figure 4.9 (a). The total frequency of unmatched locations is 32625, which is 1.68% compared to the total number of 1937716.

For all the matched location, we use the matched locations names in the external datasets to represent the locations, so that they can be mapped with their coordinates in the visualization. In terms of the unmatched location names, we keep these location names as their original names.

After mapping matched location names and keeping unmatched location names, the unique values in location names reduce to 2434 (see Figure 4.9 (b)).



	0	1
0	None	12830
1	Nieuwkuijk en Onsenoort	1420
2	Herpt en Bern	1256
3	Dieden, Demen en Langel	1146
4	Deursen en Dennenburg	899
...	...	...
838	Kethel	1
839	Keula, Duitschland	1
840	Keula-Gutsbezirk	1
841	Kimberleij	1
842	Lommelsdijk	1

843 rows × 3 columns

(a) Unmatched locations: 0-location name, 1-frequency

	unique_values	counts
0	Tilburg	117412
1	Breda	116185
2	's-Hertogenbosch	87830
3	Eindhoven	82904
4	Bergen op Zoom	39544
...	...	...
2429	Baard	1
2430	Koorndijk	1
2431	Kapel	1
2432	Klein Eybs	1
2433	Postelberg	1

2434 rows × 2 columns

(b) All distinct location names after matching: unique-values-location name, counts-frequency

Figure 4.9: Location names after matching

Chapter 5

Design and Implementation

In this chapter, we first give the overview of the visualization tool, and then the details of the tool are presented. In Section 5.1, the design principles and overview are summarized. From Section 5.2 to 5.5, different views are described. Finally, in Section 5.6, the interactions between different views are designed. All the visualizations are implemented with d3.js library.

5.1 Summary

In this section, we first discuss the design principles. Next, we introduce the overview of the tool.

5.1.1 Design Principles

In order to give a more understandable and usable tool for the users, we aim to design a simple and intuitive visualizations. There are multiple attributes in the dataset, and compared to a single view, multiple linked views enable the user to quickly view a scenario, and put this view to one side and try out another scenario [21]. Thus, the approach of using multiple different individual views simultaneously is adopted. Besides, the users are encouraged to interact with the tool to gain a deeper insight of the hidden information using these multiple views. Hence, multiple ways should be designed for interactions between different views to give the users an insight into the data. Moreover, a filter bar is needed to filter the data that meet multiple conditions and to how the filtered result in all views, so that the users are able to compare the different filtering results by changing the parameters.

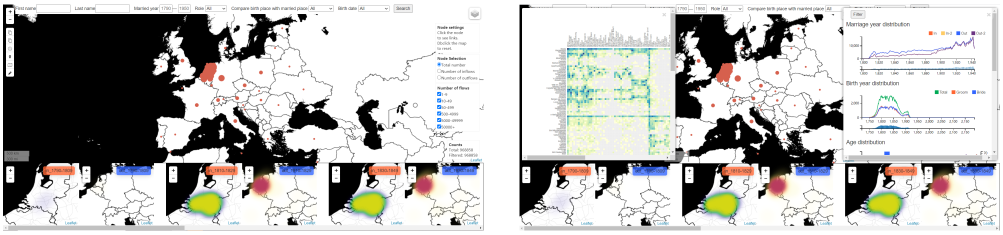
5.1.2 Overview

In Chapter 3, we discuss that the geographical information should be encoded in a map. Hence, a map is designed to demonstrate the locations and the flows. Next, in Chapter 4, we conclude that not all the locations in the marriage dataset can be matched to a coordinate in the map. Two methods are mentioned in Chapter 3 to represent the flow data without a map, which are a matrix and a circular plot. The advantages and disadvantages of two methods are discussed in Chapter 3, which concludes that the circular plot limits the number of locations and the matrix is more suitable for large or dense graphs. Since there are thousands of locations in the marriage dataset, we select the matrix view to represent the locations and flows as well. In the matrix, all the locations can be represented as rows or columns.

The requirements claim that the visualization tool should be able to observe the marriage mobility of different time periods at the same view. Hence, heatmaps are designed to show the trend of marriage mobility. Multiple heatmaps can represent the inflows and outflows of different locations at different time periods at the same time, so that the users are able to compare the differences and interpret the trends.

Except for geographical information, there are many other attributes in the marriage dataset, such as marriage dates, birth dates and roles. In order to analyse the distribution of other attributes, a sidebar is designed to show the detailed information about these attributes. Different kinds of charts are added to show the distribution of the different attributes.

As geographical information is the most important information to be displayed, the views that can show intuitive geographical information, including a map and heatmaps, are shown in the initial page of the tool (see Figure 5.1 (a)). The other two views, including the matrix and the sidebar, are initially hidden and can be expanded in pop up windows by clicking the icons at the top left corner (see Figure 5.1 (b)). Moreover, a filter bar is designed at the top of the tool for the filtering function.



(a) Initial page

(b) The page with all views expanded

Figure 5.1: Overview of the tool

5.2 Matrix

In this section, we first present some general terminologies that are used in the matrix in Section 5.2.1. Then, we describe the primary matrix, its limitations, and solution to the limitations in Section 5.2.2 and 5.2.3. Next, the matrix is reordered, and the final version of the matrix is presented in Section 5.2.4. Finally, implementation details are given.

5.2.1 General Definitions

Definition 5.2.1 Locations

$\mathbb{L}$ is the set of all distinct locations appearing in the dataset after the data preprocessing.

Definition 5.2.2 Source Locations

All the distinct locations in the columns of birth places of groom and bride are regarded as source locations. We define $S = \{s_1, s_2, ..., s_m\}$ as the set of source place names.

Definition 5.2.3 Target Locations

All the distinct locations in the columns of marriage register places are regarded as target locations. We define $T = \{t_1, t_2, ..., t_n\}$ as the set of target place names.

According to the result of Chapter 4, the total number of locations in set $\mathbb{L}$ is 2434, and $m = \pm 2500$, $n = \pm 200$. S and T are subsets of $\mathbb{L}$.

5.2.2 Primary Matrix

In the mobility matrix, each row represents a source location, each column represents a target location, and each cell represents the number of people that were born in a specific source location and married in a specific target location. The larger the value of the cell is, the darker the color of the cell is. The order of the rows depends on the total number of people that were born in these source locations, and the order of the columns depends on the total number of people that married in these target locations. The higher the total number is, the higher the order is.

There are totally $m \times n$ cells, and we define each cell as M_{ij}, where $i \in \{1, 2, ..., m\}$ and $j \in \{1, 2, ..., n\}$.

$M_{ij} = n(s_i, t_j) \ = \ \#people\ that\ were\ born\ in\ source\ location\ s_i\ and\ married\ in\ target\ location\ t_j$

From the data preprocessing result, we conclude that $m = \pm 2500$, $n = \pm 200$, so that the size of the primary matrix is about 2500×200. The first several rows are shown in Figure 5.2.

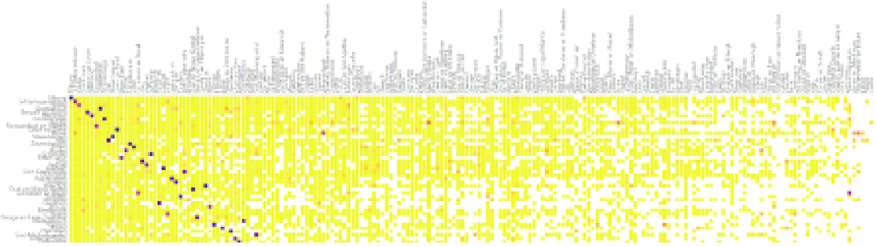


Figure 5.2: Part of the primary matrix

When scrolling to the bottom part of the matrix, the columns names that are on the top part of the matrix are not visible for users because of the large number of rows. Moreover, with the size of 2500×200, the running speed to load the matrix is slow. In order to resolve these two problems, we reduce the number of rows by clustering.

5.2.3 Clustering

In this section, we present a method to decrease the number of rows in the primary matrix.
Definition 5.2.4 Closer locations

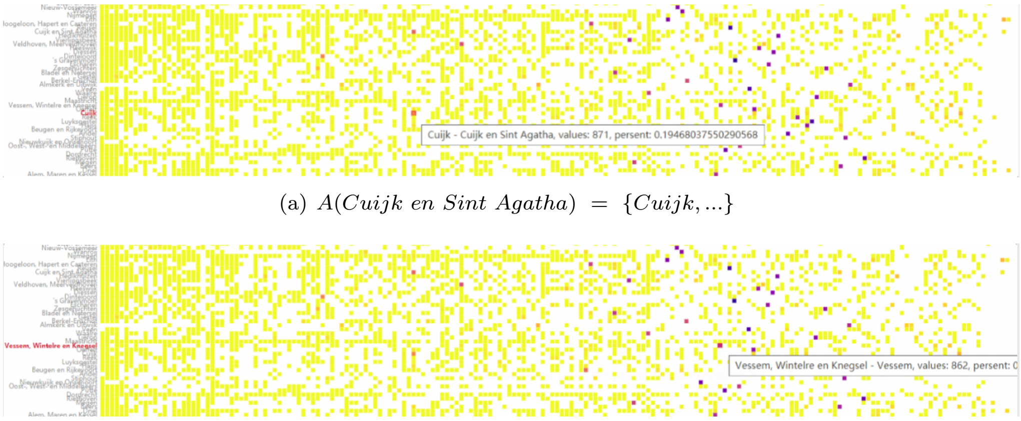
If the number of people that were born in source location s_i and married in target location t_j, which is $n(s_i, t_j)$, is larger than the number of people that were born in source location s_i and married in target location t_k, which is $n(s_i, t_k)$, we define that source location s_i is closer to target location t_j than to target location t_k, where $i \in \{1, 2, ..., m\}, \ j, k \in \{1, 2, ..., n\}$.
Definition 5.2.5 Maximum target location

If source location s_i is closer to target location t_j than to any other target location, then we define that target location t_j is the maximum target location of source location s_i. If a tie happened between two target locations for the same source location, we define that the target location with a larger total number of people that get married in this location is closer to this source location, and it is the maximum target location of this source location.
Definition 5.2.6 Set partitioning

Define A_i as a set of source places that are closer to target place t_i than to any other target place, $i \in \{1, 2, ..., n\}$, then we define the set $A = \{A_1, \ A_2, ..., \ A_n\}$ as the set of new source locations. $\bigcup\limits_{i=1...n} A_i = S, \ i \neq j \Rightarrow A_i \cap A_j = \emptyset$.

According to the definition of the primary matrix, the corresponding column name of the cell with the darkest color in a row is the maximum target location. Hence, from the definition 5.2.6, the row name of this cell is clustered to the set of the column name of this cell. For example, as shown in Figure 5.3 (a), row 'Cuijk' has the darkest color for column 'Cuijk en Sint Agatha'. Therefore, 'Cuijk en Sint Agatha' is the maximum target location of 'Cuijk', and 'Cuijk' is included in new source location 'Cuijk en Sint Agatha': $A(Cuijk\ en\ Sint\ Agatha) \ = \ \{Cuijk, ...\}$. Similarly, in Figure 5.3 (b), $A(Vessem) \ = \ \{'Vessem, Wintelre\ en\ Knegsel', ...\}$.

By clustering, we can observe a new matrix whose number of rows is equal to the number of columns. Figure 5.4 shows the result of matrix after clustering. The clustered matrix has a size of about 200×200. Each row represents a source location set, each column represents a target location, and each cell represents the number of people that were born in a specific source location



Cuijk - Cuijk en Sint Agatha, values: 871, persent: 0.19468037550290568

(a) $A(Cuijk\ en\ Sint\ Agatha)\ =\ \{Cuijk, ...\}$

Vessem, Wintelre en Knegsel - Vessem, values: 862, persent: 0

(b) $A(Vessem)\ =\ \{'Vessem, Wintelre\ en\ Knegsel', ...\}$

Figure 5.3: Examples of clustering

set and married in a specific target location. There are totally $n \times n$ cells, and we define each cell as N_{ij}, where $i, j \in \{1, 2, ..., n\}$.

$$N_{ij} = n(A_i, t_j)\ =\ \#people\ that\ were\ born\ in\ location\ set\ A_i\ and\ married\ in\ target\ location\ t_j$$

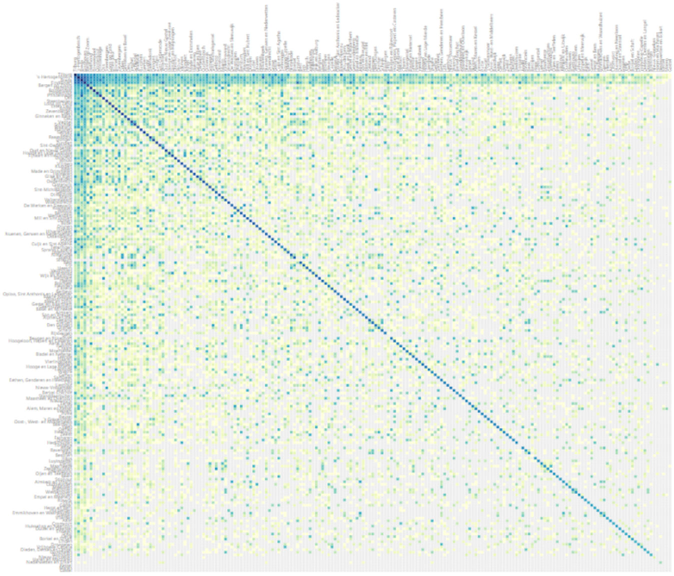


Figure 5.4: Matrix after the clustering

For the clustered matrix, multiple properties can be observed as follows. Since the order of columns of the primary matrix is decided by the total number of people married in target locations, the top rows and columns have a darker color. Besides, as the order of the rows is the same as the order of the columns in the clustered matrix, it is obvious that the values on the diagonal are larger than the other values. Except for these two features, we cannot find any other patterns in the matrix. Since reordering can find a proper order for rows and columns to compose a visual matrix [11], we reorder the rows and columns to explore the patterns .

5.2.4 Reorder

To make the locations with similar mobility patterns closer to each other, we perform a reorder process.

Behrisch and Bach [11] summarized the complexity of different methods of reordering. For a directed and asymmetric matrix the matrix requires two permutations, one π_r for the rows and one π_c for the columns. Since a brute-force approach for a $m \times n$ matrix would actually require $n! \times m!$ comparisons, it is too time-consuming to find the optimal order of the rows and columns.

In our project, a greedy algorithm is implemented to calculate a permutation of rows and columns. When reorder the rows, the first row keep the original order as the first sorted row. Then, the following unsorted rows are ordered one by one. By calculates the similarities that are valued by distances between unsorted row and sorted rows, an unsorted row is inserted after the sorted row with the smallest distance. The reorder of columns are similar to that of rows. In this way, the rows and columns with similarities are adjacent to each other. Although greedy algorithms do not find an optimal solution, it can quickly calculate a "reasonable good" solution.

To measure the distance between individual rows or columns, some functions can be used to calculate the similarity. For two rows or columns x and y:

1 Euclidean Distance: $d(x, y) \ = \ \sqrt{\left(\sum (x_i - y_i)^2\right)}$

2 Cosine Similarity: $T(x, y) \ = \ \dfrac{\sum x_i y_i}{\sqrt{\sum x_i^2}\sqrt{\sum y_i^2}}$

3 Manhattan Distance: $d(x, y) \ = \ \sum \mid x_i - y_i \mid$

4 Hamming Distance: All the non-zero values are counted as 1, and zero values are counted as 0. Hamming distance is the number of positions at which the corresponding symbols are different.

5 Chebyshev Distance: $d(x, y) \ = \ max(\mid x_i - y_i \mid)$

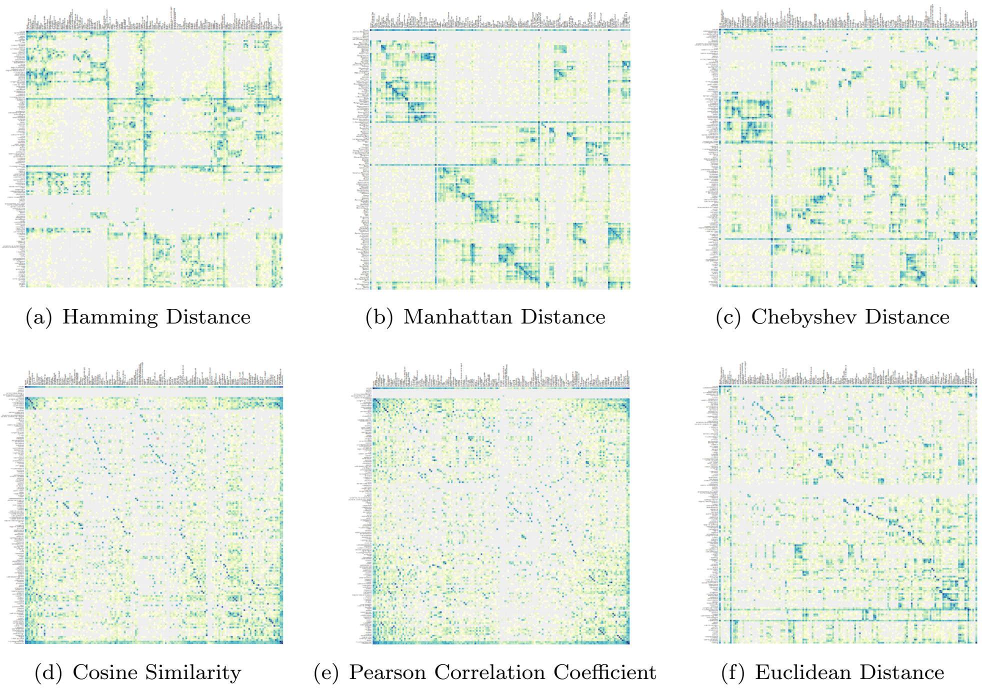
6 Pearson Correlation Coefficient: $p(x, y) \ = \dfrac{n \sum x_i y_i - \sum x_i \sum y_i}{\sqrt{n \sum x_i^2 - \left(\sum x_i\right)^2}\sqrt{n \sum y_i^2 - \left(\sum y_i\right)^2}}$

We attempted all the distance functions above, and the results are shown in Figure 5.5. The last three graphs have no obvious pattern, while the first three graphs have rectangle patterns. These patterns reflect similarities between adjacent rows and columns. As it is difficult to compare the pattern in these three graphs, we select one method to apply in the reordering step, which is Hamming distance. The reordered matrix is shown in Figure 5.5 (a), and adjacent rows or columns after reordering represent the locations that have similar marriage mobility patterns.

5.2.5 Implementation Details

The sequential colour schemes that we select for the matrix is shown in Figure 5.6. This is relatively friendly to common forms of people who are affected by color vision deficiency, and can help the accurate perception of data by as many viewers as possible [20].

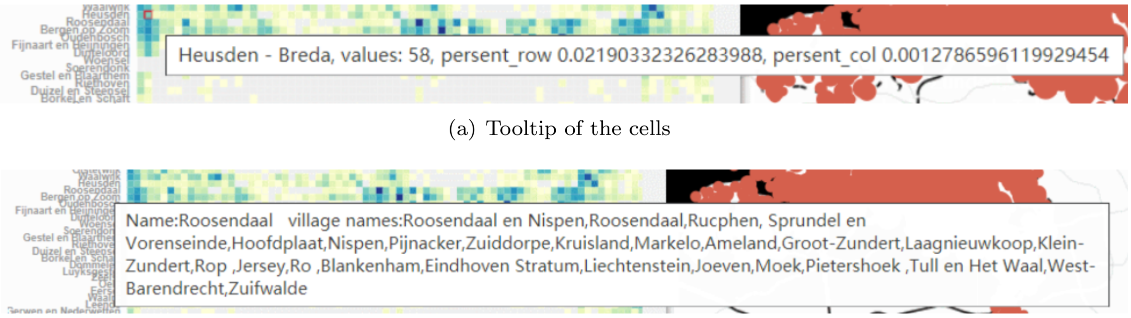
The matrix provides common interactions: when the users hover the mouse over a cell in the matrix, the corresponding row, column, value, the percentage that the value takes in this row, and the percentage that the value takes in this column will be shown in a tooltip (see Figure 5.7 (a)). Besides, in order to provide detailed information, when users hover the mouse over a row, the locations that are clustered in these rows will be shown in another tooltip (see Figure 5.7 (b)).



(a) Hamming Distance
(b) Manhattan Distance
(c) Chebyshev Distance
(d) Cosine Similarity
(e) Pearson Correlation Coefficient
(f) Euclidean Distance

Figure 5.5: Results of reordering

Figure 5.6: Color scheme



Heusden - Breda, values: 58, persent_row 0.02190332326283988, persent_col 0.0012786596119929454

(a) Tooltip of the cells

Name:Roosendaal   village names:Roosendaal en Nispen,Roosendaal,Rucphen, Sprundel en Vorenseinde,Hoofdplaat,Nispen,Pijnacker,Zuiddorpe,Kruisland,Markelo,Ameland,Groot-Zundert,Laagnieuwkoop,Klein-Zundert,Rop ,Jersey,Ro ,Blankenham,Eindhoven Stratum,Liechtenstein,Joeven,Moek,Pietershoek ,Tull en Het Waal,West-Barendrecht,Zuifwalde

(b) Tooltip of the rows

Figure 5.7: Tooltips in the matrix

5.3 Map

The map is placed in the center of the page. Although the map cannot show all the locations in the marriage dataset, it can intuitively display geographical information. In this section, we first present how to visualize the locations and flows in the map. Next, we describe some implementation details in the map. The map is implemented with spatialsankey.js [13].

5.3.1 Location

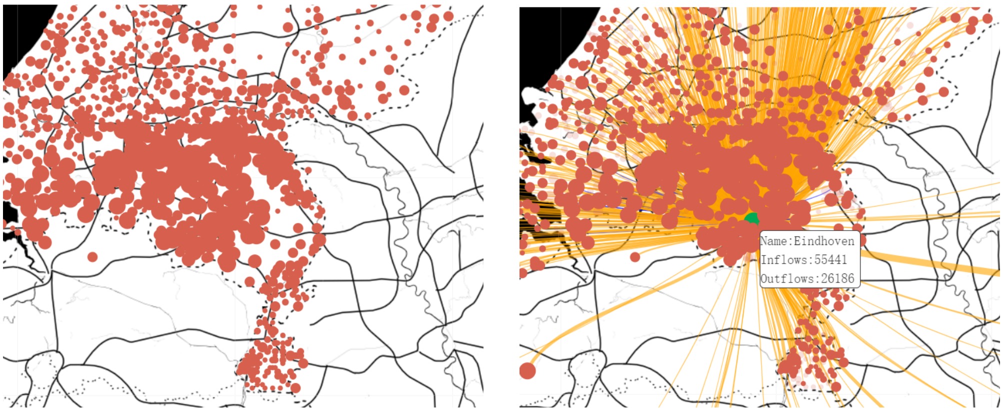
In this section, we first propose a terminology about the locations on the map.

Definition 5.3.1 Locations on the map

We define $\mathbb{L}_{map} = \{l_1, l_2, ..., l_k\}$ as the set of locations $l_i \in \mathbb{L}$ that have geographical coordinates.

According to the result of Chapter 4, $\mathbb{L}_{map}$ is a subset of $\mathbb{L}$, and the total number of locations in set $\mathbb{L}_{map}$ is k =1591.

The method used to visualize the locations have been discussed in Chapter 4, which is to represent locations as points on the map. As shown in Figure 5.8(a), all the locations are visualized as points, and the radius of the points is determined by the total number of inflows and outflows.



(a) Points on the map

(b) Links of select point(green) and tooltip on the map

Figure 5.8: The map view

5.3.2 Flows

We first introduce some terminologies related to flows.

Definition 5.3.2 Flow

For the locations l_i and l_j, we define the number of people that were born in location l_i and married in l_j as a flow $\vec{F}_{ij}$.

According to the definition, $\vec{F}_{ij}$ is a subset of M_{ij}.

Definition 5.3.3 Inflow,outflow

We define the total number of people that were born in a specific location l_i in the map as outflows $\mathbb{F}_i^{out}$, where $\mathbb{F}_i^{out} = \sum_{j=1}^{k} \vec{F}_{ij}$. Similarly, we define the total number of people that married in a specific location l_j in the map as inflows $\mathbb{F}_j^{in}$, where $\mathbb{F}_j^{in} = \sum_{i=1}^{k} \mathbb{I}_{ij}$.

The flows in the matrix are represented as cells from a row to a column, and the flows in the map are represented as curves from the source point to the target point. According to the result of data preprocessing, there are totally 28021 vector flows in the map. If all the flows are shown together, it will be difficult to read and find the pattern for users even if we apply the edge-bundling. Thus, we only show the points and not show the flows in the initial page. In this case, the users can check the vector flows related to the locations that they are interested in by clicking them.

As shown in Figure 5.8 (b), when the user selects a point in the map, the selected point is highlighted, and all the points that are not related to the selected point will become more transparent. All the flows of the selected point are represented as curves in the map. The orange curves represent the flows that regard the selected point as the target point, the blue curves represent the flows that regard the selected point as the source point, and the width of the curves reflects the value of flows. The users can clear the selected points or links by the icons or double clicking. Moreover, the number of the inflow and the outflow of a point in the map will show in a tooltip when the users hover their mouse on the points.

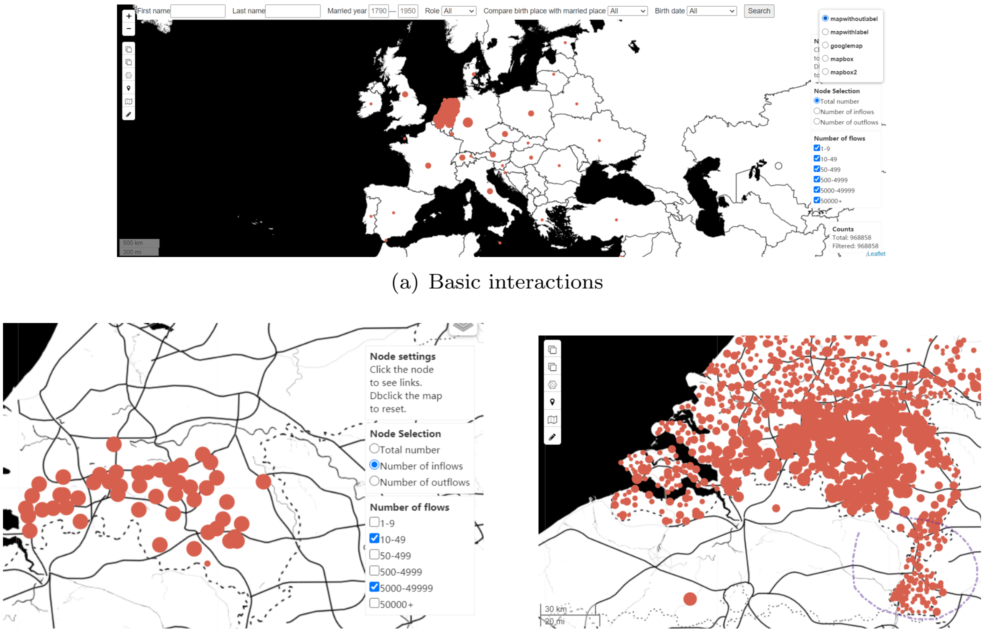
5.3.3 Implementation Details

All the implementation details in the map view are described in this section.

Basic Interaction

There are some normal interactions in the map, including zooming, panning, multiple map layers. As shown in Figure 5.9 (a), the default zooming scale and the center of the map are set to values, which are convenient for users to view the distribution of points all over the world. The users are able to zoom in or zoom out with the zooming buttons and the ratio scale will be changed accordingly. The map also supports panning and dragging.

Figure 5.9 (a) also shows a button on the top right corner of the map that is used to switch five map layers. Multiple map layers are provided to show different information. The default layer, named 'mapwithoutlabel', is the map without label of location names. There are four other layers for users to choose from.



First name Last name Married year 1790 — 1950 Role All Compare birth place with married place All Birth date All Search

● mapwithoutlabel
○ mapwithlabel
○ googlemap
○ mapbox
○ mapbox2

Node Selection
● Total number
○ Number of inflows
○ Number of outflows

Number of flows
☑ 1-9
☑ 10-49
☑ 50-499
☑ 500-4999
☑ 5000-49999
☑ 50000+

Counts
Total: 968858
Filtered: 968858

(a) Basic interactions

Node settings
Click the node to see links.
Dbclick the map to reset.

Node Selection
○ Total number
● Number of inflows
○ Number of outflows

Number of flows
☐ 1-9
☑ 10-49
☐ 50-499
☐ 500-4999
☑ 5000-49999
☐ 50000+

(b) Choice boxes controlling the display of points

(c) Multiple selection

Figure 5.9: Implementation details of the map view

Display of Points

In order to help users to observe the points, two choice boxes are designed on the left hand of the map view to allow showing only a subset of the points. In the first box, the users can choose the attribute used to filter the points. Three options are provided, filtering the points by the number of total flows, filtering the points by the number of inflows and filtering the points by the number of outflows. In the second box, the points are classified in five categories for the chosen attribute in the first box. The default setting is to show all the points, and Figure 5.9 (b) shows an example of selecting 'number of inflow' in the first box, and selecting '10-49' and '5000-49999' in the second box.

Multiple Selection

There is a series of button on the left hand side of the map. We have discussed the function of the first five icons, and the last icon is designed for multiple selection. As shown in Figure 5.9 (c), when the users click this button, they can draw an arbitrary shape to select all the points inside this shape.

5.4 Heatmap

A Heatmap is an intuitive visualization to observe the trend of the distribution of inflows and outflows. The locations are represented as points, and the color corresponds to the number of inflows and outflows. The heatmaps are implemented using leaflet-heat.js [9].

5.4.1 General Idea

Heatmaps display an overview of the marriage dataset. The user can compare marriage mobility during different time periods. The data from 1790 to 1950 were recorded in the marriage dataset. Considering that the number of heatmaps should be large enough to compare the trend of mobility. However, to make sure the heatmaps can be displayed in the corresponding place, we use 20 years as a time period for comparison. Therefore, there are totally eight intervals for the heatmap, respectively, 1790-1809, 1810-1829, 1830-1849, 1850-1869, 1870-1889, 1890-1909, 1910-1929, 1930-1949. For each time interval, there are two heatmaps, one for the number of inflows and the other for the number of outflows. At last, as shown in Figure 5.10, there are 16 heatmaps to display the distribution of inflows and outflows in different time periods.

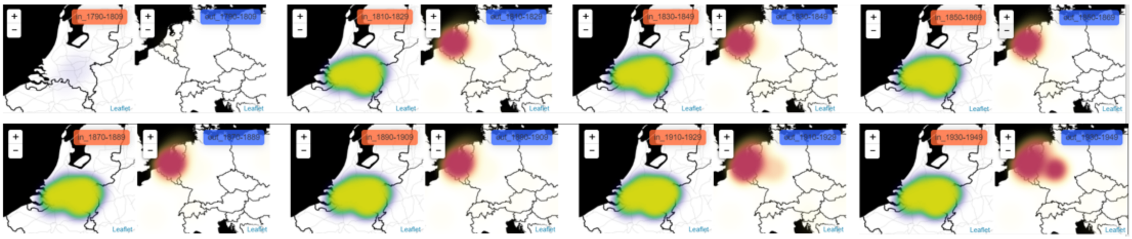


Figure 5.10: Heatmaps: orange label - inflow heatmap, blue label - outflow heatmap, gradient color from blue to green to yellow for inflow heatmaps from yellow to orange to purple for outflow heatmaps.

5.4.2 Implementation Details

To distinguish heatmaps between different time periods, we add a label on the right top corner to mark the time period of the heatmaps and add spacing between adjacent pairs of the heatmap. Moreover, as all inflow locations are within Noord-Brabant province, while all the outflow locations are distributed all over the world, we set different values for the default centres and the zooming scales of inflow and outflow heatmaps. In addition, we customize different gradient colours, separately, from blue to green to yellow for inflow heatmaps, and from yellow to orange to purple for outflow heatmaps. Moreover, the time intervals are fixed in the tool, which will not change when using the filter bar.

5.5 Sidebar

Apart from visualizing the geographical information, other attributes in the marriage dataset should also be shown. For birth date, marriage date, role, related source locations and target locations of the locations that users selected, we design five intuitive charts in the right sidebar to show them. In this section, we present these charts and the interactions between them.

5.5.1 Move year distribution

To analyse the marriage mobility of specific locations, we add a timeline chart to display the change of the number of people that move into or move out from the specific locations by marriage over time. The X-axis represents the marriage year, and the Y-axis represents the number of people. There are four lines in the chart, the orange one represents the number of people move in, the yellow one represent the number of people move in and without birth date when registering the marriage, the blue one represents the number of people move out, and the purple one represent the number of people move out and without birth date when registering the marriage. The default chart shows the result of all the locations in the marriage dataset. Hence, as shown in Figure 5.11 (a), the orange line and blue line are overlapped, and the yellow line and purple line are overlapped. Below the chart, there is a slider that is used for selection of a specific time period. Moreover, some common interactions can be performed in this chart. When the users hover their mouse over a point in the chart, a tooltip will displays the specific value. When the users hover their mouse over the legend on the top right corner, the corresponding line will be highlighted.

5.5.2 Birth year distribution

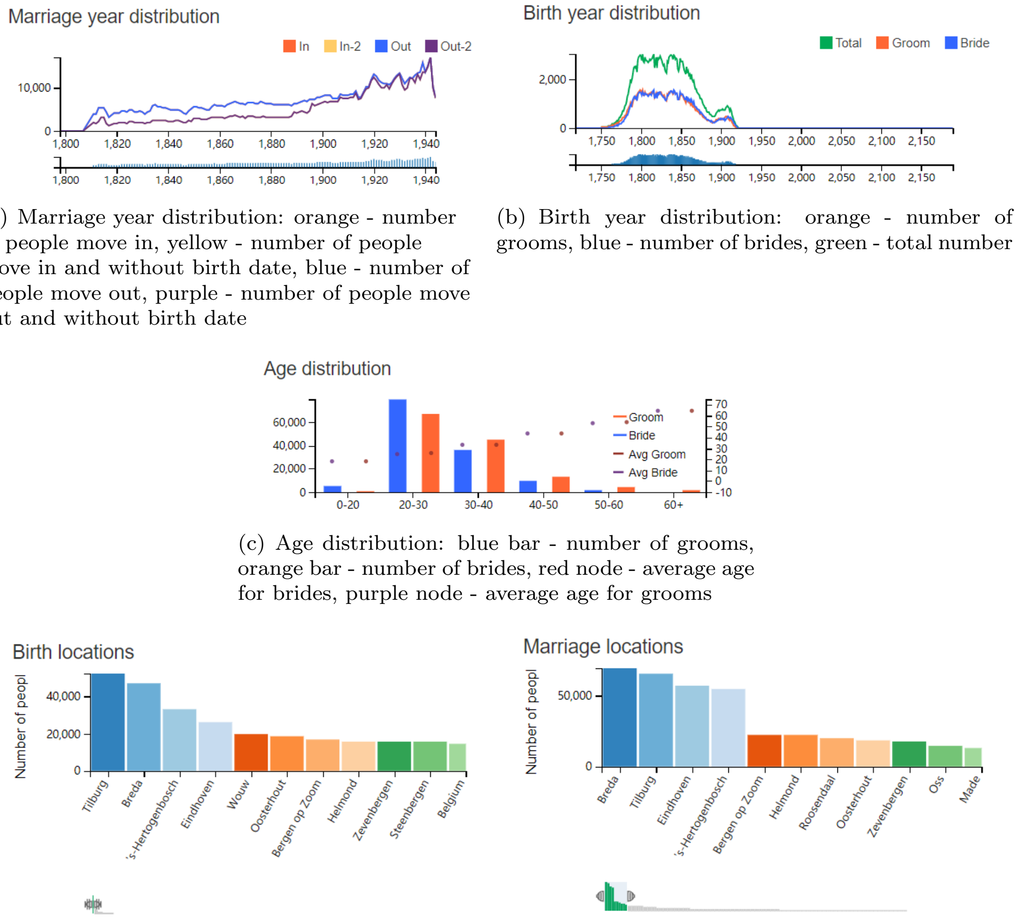
To analyse the birth year of people that move into or move out from specific locations by marriage, we add a timeline chart to represent the number of people that were born in different years in the marriage dataset. The X-axis represents the birth year, and the Y-axis represents the number of people. There are three lines in the chart, the orange one represents the number of grooms, the blue one represents the number of brides, and the green one represents the total number. As shown in Figure 5.11 (b), the default chart shows all the data in the marriage dataset. Below the chart, there is a slider that is used for selection of a time period, and the users can select by dragging the bar. Moreover, same as the first chart, it has common interactions that enable to show tooltips and highlight by hovering the mouse.

5.5.3 Age distribution

Since we have the marriage date of all data, and birth date of part of the data, it is possible for us to calculate ages of part of the data. There are some alternative methods for the age distribution, such as scatter plot, box plot, line chart and bar chart. Consider the response time, the scatter plot and box plot is not considerable. To analyse the age distribution that related to specific locations, a bar chart is a better choice to show the difference between the number of brides and grooms got married in each age range. Also, to compare difference between the average ages of bride and groom in each range, the average ages of them are calculated. The X-axis represents different ranges of age. As the legal age for marriage in the Netherlands changed to 18 since 1965 [7], we choose 0-20 as the first range, then we set the range for each 10 year, and the last range is for the people that married when they are older than 60. The Y-axis on the left side represents the number of people. The blue bars represent the number of grooms within a range, and the orange bars represent the number of brides within a range. The Y-axis on the right side represents the average age of each range, which corresponds to the red nodes for brides and purple nodes for grooms (see Figure 5.11 (c)).

5.5.4 Birth and marriage locations

We use one bar chart to show source locations of selected location, and one for target locations of selected location. If the users select multiple locations, then one bar charts shows the distributions over all locations of where the people moving to the selected locations come from, and vice versa for the other bar chart. As the relevant locations can be many and the space may not be sufficient, the top ten locations are displayed by default. If the users expect to check other locations, a slider is designed below each bar chart to change the displayed locations(see Figure 5.11 (d) and (e)).



(a) Marriage year distribution: orange - number of people move in, yellow - number of people move in and without birth date, blue - number of people move out, purple - number of people move out and without birth date

(b) Birth year distribution: orange - number of grooms, blue - number of brides, green - total number

(c) Age distribution: blue bar - number of grooms, orange bar - number of brides, red node - average age for brides, purple node - average age for grooms

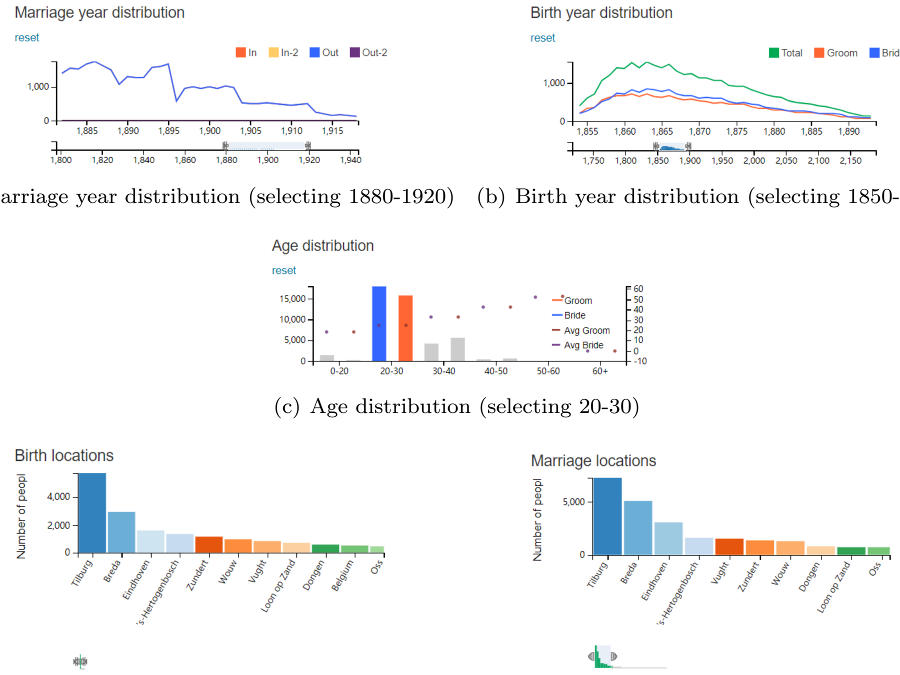
(d) Birth locations (top 10)

(e) Marriage locations (top 10)

Figure 5.11: Five kinds of charts in the sidebar

5.5.5 Interactions between different charts

The sidebar defaults to show the data of selected locations, and the filter functions are designed for interactions between different charts in the sidebar. The user can select specifically a marriage date, birth date period by scrolling the slider of the timeline charts, or select a specific age range by clicking the bar chart of age distribution. Then, the other four charts will be filtered by the selected parameters as well. Figure 5.12 shows an example of interactions between different charts. After selecting the marriage year from 1880 to 1920, the birth year from 1850 to 1900, and married between 20 and 30 year old in the first three charts, all the charts show the filtered result. We can observe that as we selected a period for birth year, the yellow and purple line is zero in the marriage year distribution chart, and red and blue line are still overlapped.



(a) Marriage year distribution (selecting 1880-1920)

(b) Birth year distribution (selecting 1850-1900)

(c) Age distribution (selecting 20-30)

(d) Birth locations

(e) Marriage locations

Figure 5.12: Example of interactions in the sidebar: selecting time period in (a) and (b), and selecting range in (c), then the views in five charts have changed.

5.6 Interactions

We have introduced individual views and the interactions in each view of the tool. In this section, the interactions between different views are presented.

5.6.1 Interactions between individual parts

Figure 5.13 demonstrates the structure of interactions. There is a location list that is used to record the selected location names among all the views. More precisely, the users can add location names to this location list by selecting rows or columns in the matrix or selecting points in the map. After that, all the views will be changed according to the location list.

For example, as shown in Figure 5.14, when the users select some rows or columns in the matrix, the selected rows in the matrix are highlighted with red rectangle in each cell, and the corresponding points in the map are highlighted with yellow color. When the users selected some points in the map, the selected points are highlighted with green color and the corresponding row names or column names in the matrix are highlighted with red. With the selected points in the matrix and map, the heatmaps can automatically update. As automatically updating the sidebar is time-consuming, a 'Filter' button is designed on the top of the sidebar, and it only updates if the users click the button.

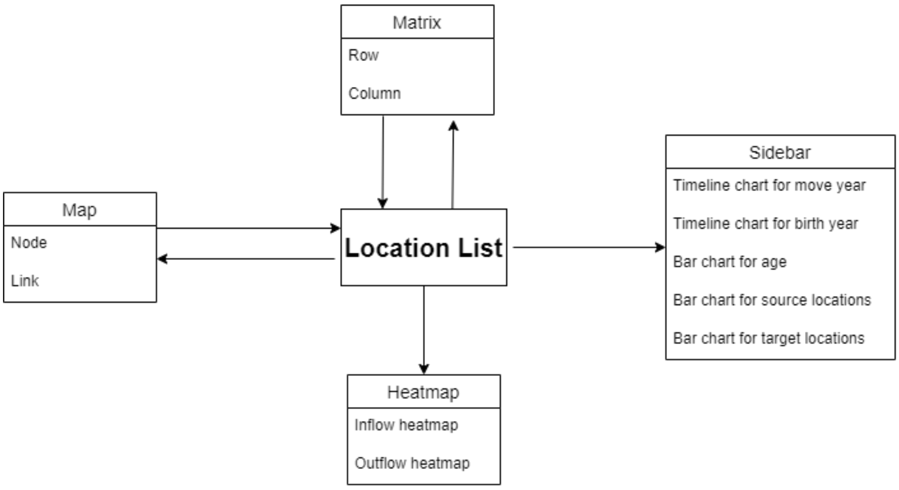

Matrix
Row
Column
Map
Node
Link
Location List
Sidebar
Timeline chart for move year
Timeline chart for birth year
Bar chart for age
Bar chart for source locations
Bar chart for target locations
Heatmap
Inflow heatmap
Outflow heatmap

Figure 5.13: Interactions between individual views

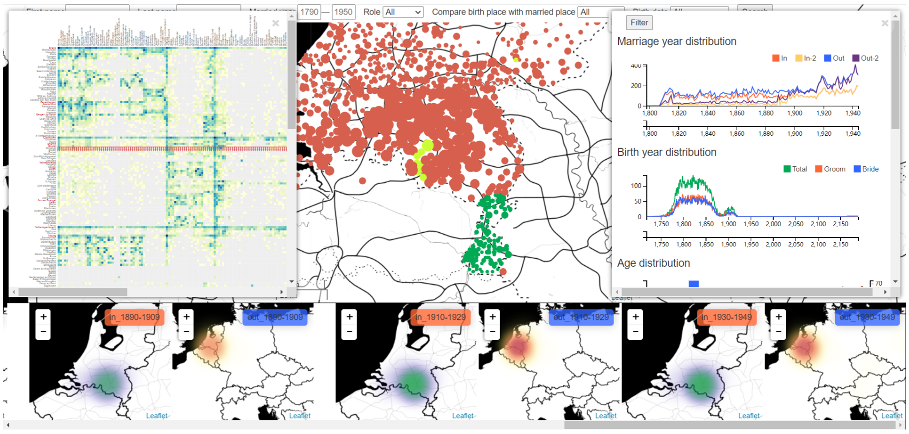


Figure 5.14: Example of interactions between individual views

5.6.2 Filtering

As shown in Figure 5.9, according to the different attributes in the marriage dataset, we provide a filter bar with five kinds of conditions for users to filter the data that they are interested in. The first condition is about names, the users can search for a family or person according to the last name or first name. The second condition is about the marriage year, where the user can input a time period. The third condition is about the role, the default choice is all the people, and the users can select bride or groom here. The fourth condition is about the birth places and marriage places. The default choice is all the data, and the users can choose to check the records that have the same or different marriage places compared to the birth places. The last condition is about the birth date, the default choice is to show all the data; alternatively, the users can select to check only the records with or without missing birth date. Moreover, the result of the number of the records after filtering is shown in the right bottom corner of the tool, as shown in Figure 5.16.

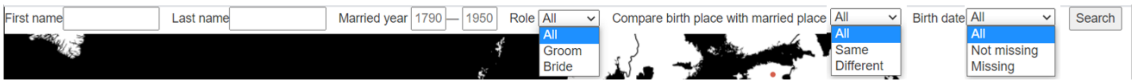
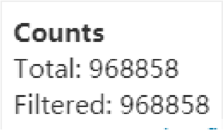


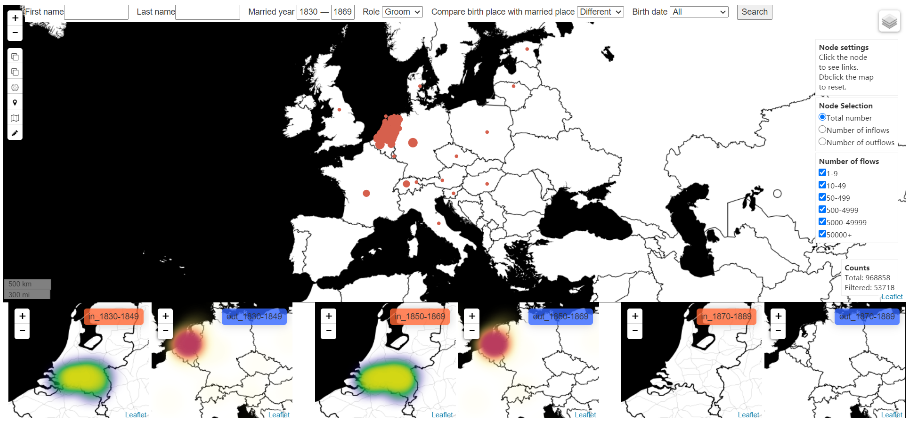
Figure 5.15: Filter bar



Counts
Total: 968858
Filtered: 968858

Figure 5.16: Count box for filtering result

As shown in Figure 5.17, when some conditions are filled or selected in the filter bar, the views are filtered and the number of filtered records are shown. For example, when '1830-1869' is selected for the married year, we can observe that the heatmap of '1870-1889' is empty.



First name
Last name
Married year 1830 — 1869
Role Groom
Compare birth place with married place Different
Birth date All
Search

Node settings
Click the node
to see links.
Dbclick the map
to reset.

Node Selection
● Total number
○ Number of inflows
○ Number of outflows

Number of flows
☑ 1-9
☑ 10-49
☑ 50-499
☑ 500-4999
☑ 5000-49999
☑ 50000+

Counts
Total: 968858
Filtered: 53718

Figure 5.17: An example of filter bar: select '1830-1869' for 'married year', 'groom' for 'role' and 'different' for 'compare birth place with married place'

Chapter 6

Results

The goal of the project is to build a tool to visualize the marriage dataset and enable the users to explore marriage mobility and related details in multiple ways. The structure of the realized tool is shown in Figure 6.1. The users can filter the four views by different conditions with the filter bar, and the four views display different attributes in the marriage dataset. In this chapter, we discuss whether the tool answers the questions and satisfies the requirements in Chapter 2, and present some interesting results that we have obtained using the tool.

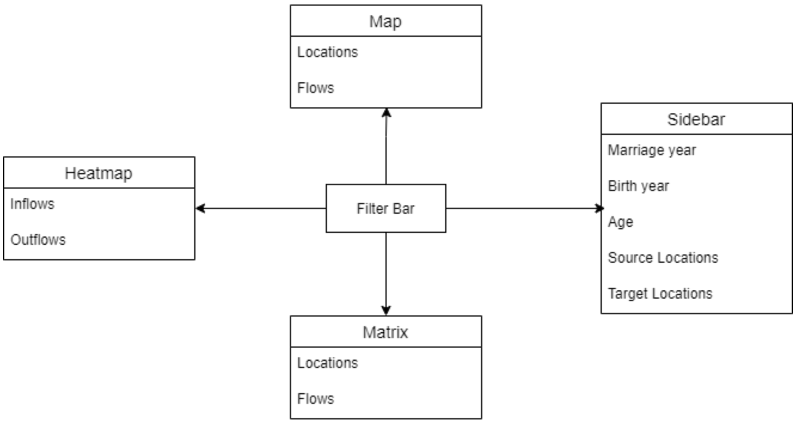


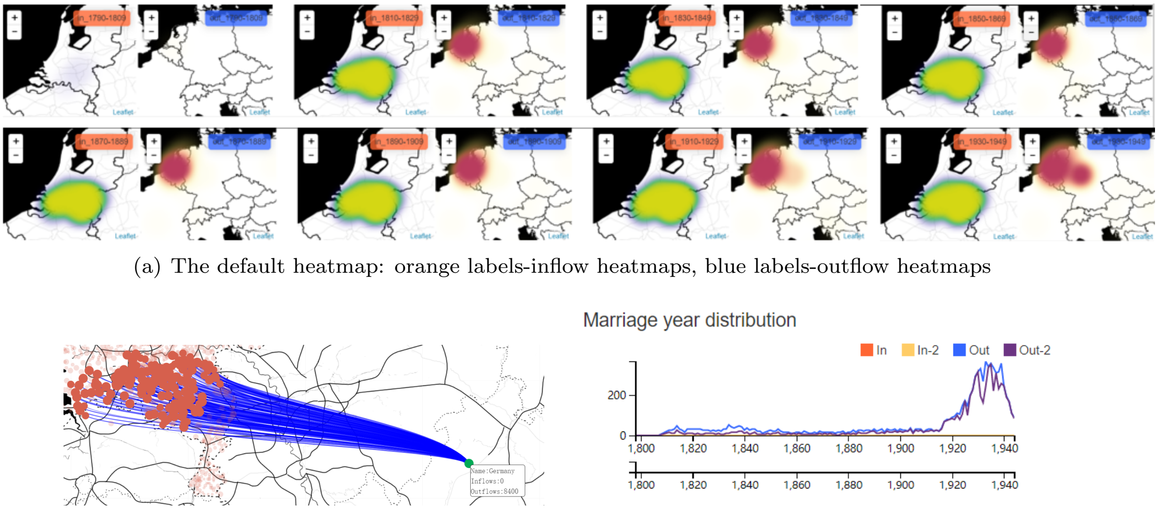
Figure 6.1: Structure of the tool

6.1 Questions and Requirements

In this section, we demonstrate how the tool can be used to answer the questions and satisfy the requirements.

6.1.1 Questions

For question 1, *How does marriage mobility change over time?*, the marriage mobility of different periods can be observed in the heatmaps, the results of the heatmaps are presented in detail in use case 1. Question 2, *What kinds of patterns or anomalies can be discovered in the marriage mobility? Are they related to any historical event?*, is usually answered with the combination of the map, the heatmaps and the sidebar. For example, in the heatmaps in Figure 6.2 (a), we find that more and more people move from Germany to Noord-Brabant by marriage as time goes by. In the map in Figure 6.2 (b) , we observe that there are totally about 8000 people from Germany that married in Noord-Brabant, which is a relatively large number. By checking the marriage year distribution in Figure 6.2 (c), we find that the number of people married in Noord-Brabant and were born in Germany begins to increase since World War I started and keeps in a high value between 1920 and 1940, and rapidly decreased from 1940. As the latest marriage year is 1944, we cannot conclude the reason of the decreasing.



(a) The default heatmap: orange labels-inflow heatmaps, blue labels-outflow heatmaps

(b) The map: green point-selected point, lines-flows related to selected point

(c) Marriage year distribution: blue-move out by marriage, purple-move out by marriage and without birth date.

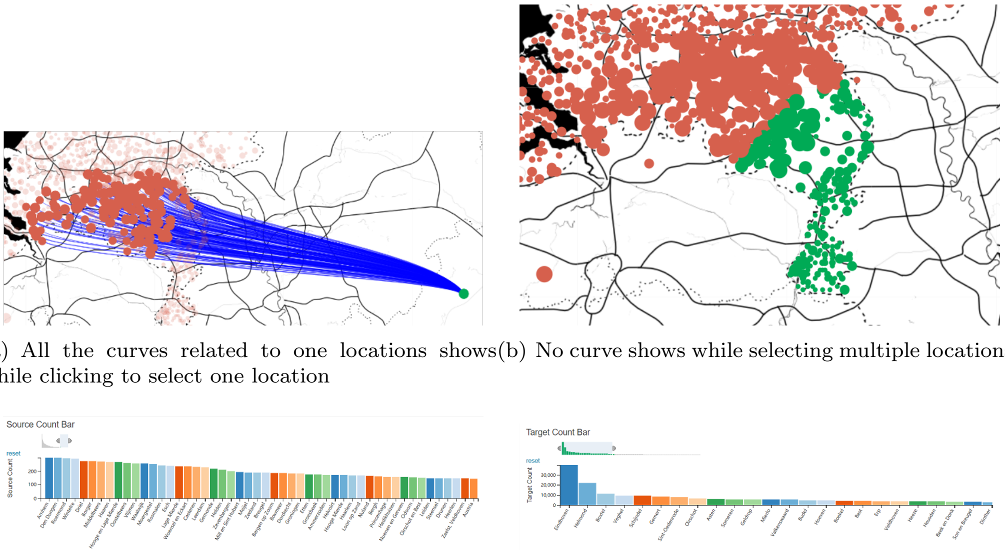
Figure 6.2: Marriage mobility from Germany to Noord-Brabant

For question 3, *What is the marriage mobility of specific locations during a specific time period?*, the users can filter a specific period with the filter bar and then select some locations of interest. As shown in Figure 6.3 (a), if the users select one location, the map can show all the links related to the location. However, if users select multiple locations, showing all curves can be messy. Therefore, as shown in Figure 6.3 (b), the map will not shows the links. In this case, the sidebar shows all the related birth locations and marriage locations information (see Figure 6.3 (c) and (d)).

For question 4, *How are groups of locations with similar mobility patterns related to each other?*, according to the methods that we use in clustering and reordering in the matrix, we can conclude that the villages in the same row and the villages in the adjacent rows have similar mobility patterns. When the users select a row in the matrix, the corresponding points in the map will be highlighted, so that we can observe the geographical distribution of these points. As shown in Figure 6.4 (a), in most cases, the villages in a row are geographically close to each other. Also, as shown in Figure 6.4 (b), in most cases, the adjacent rows in the matrix contain villages that are geographically close to others.

Question 5, *How does the number of people that married in or were born in specific locations change over time?*, Question 6, *How does the age distribution of people when they get married change over time?*, and Question 7, *How is the age or birth date distribution influenced by different roles?*, can be answered with the map and the sidebar. When the users select multiple locations in the map and matrix, Figure 6.5 (a) displays the change of the number of people married over time. Meanwhile, Figure 6.5 (b) displays the change of the number of people that were born in different years, and their age distribution of different roles.

Moreover, as it is not easy to link a family in the marriage dataset, question 8, *What is the married mobility of a specific family?*, cannot be answered accurately. The filter bar provides a searching function of first name or last name, and the users can only observe the data with the same last name. For example, Figure 6.6 shows the result of searching a family name from a well-known name 'Vincent van Gogh'. The sidebar shows the top 10 birth places and marriage



(a) All the curves related to one locations shows while clicking to select one location

(b) No curve shows while selecting multiple locations

(c) Related birth locations of multiple selected locations

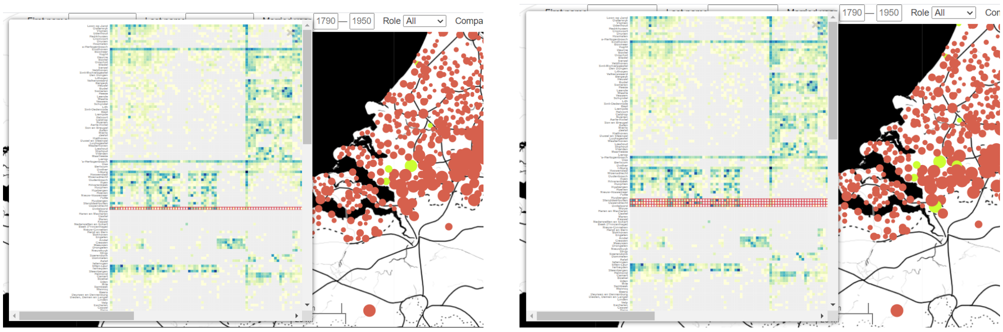
(d) Related marriage locations of multiple selected locations

Figure 6.3: Different ways of showing related locations of selected locations in the tool

places.

6.1.2 Requirements

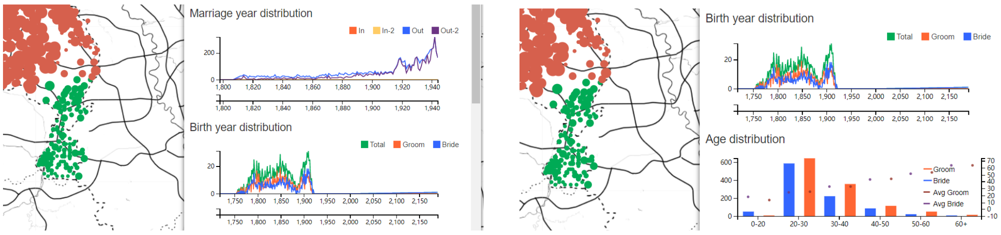
We first consider the functional interface requirements. A1, *the inflow and outflow of each location at different time periods*, is satisfied by the heatmap view. Next, A2, *inflows and outflows of different locations*, is satisfied by the tooltip in the map. Next, the users can check the flows by clicking points, which matches the requirement of A3, *the flows related to particular locations.* Next, the adjacent rows and columns in the matrix are locations that have similar mobility patterns, and the same locations in the matrix and map are linked. Hence, A4, *groups of locations with similar mobility patterns should be identified and viewed*, is satisfied by the interactions between the matrix view and the map view. Next, in the sidebar view, A5, *the change of the number of inflows and outflows of selected locations*, is satisfied by the chart of marriage year distribution. Besides, A6, *birth year distribution and age distribution*, is satisfied by the charts of birth year



(a) Select one row

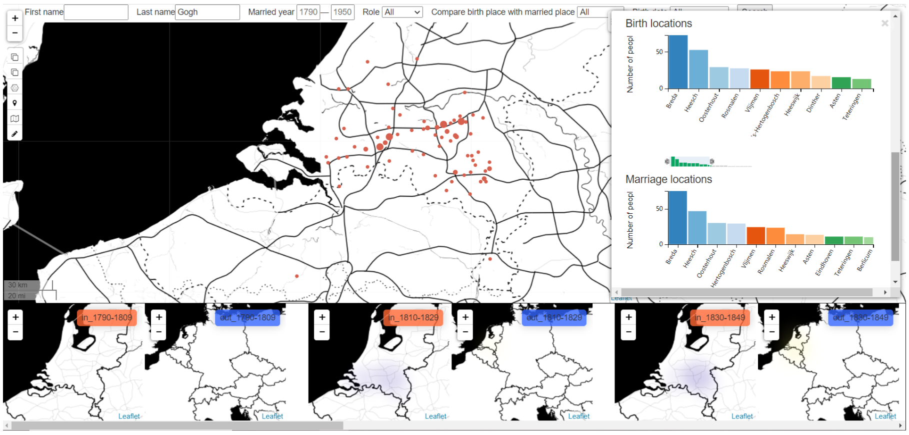
(b) Select several rows

Figure 6.4: Corresponding points(highlighted with yellow) in the map while selecting rows(highlighted with red) in the matrix



(a) Marriage year distribution (b) Birth year and age distribution

Figure 6.5: Filtered sidebar with several selected locations(green points)



First name
Last name Gogh
Married year 1790 — 1950
Role All
Compare birth place with married place All

Figure 6.6: Result of searching 'Gogh' in filter bar

distribution and age distribution. Moreover, the last two charts separately show the distribution of other attributes, including the birth locations and the marriage locations.

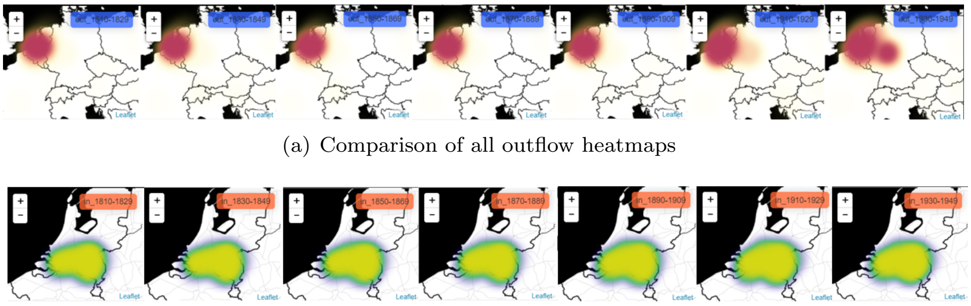
Second, we consider the functional interaction requirements. Filter functions(B1) is satisfied by the filter bar. Basic interactions(B2) are implemented in the multiple views. For example, the zooming and panning are implemented in the map, the heatmaps, and the matrix. Besides, the function of multiple map layers is implemented in the map. Multiple locations selection(B3) is satisfied in the map view, and there is a button to perform the selection of multiple locations. Last, we consider the non-functional requirements. In the data preprocessing and the implementation part, we keep considering the requirement of keeping faithful to the authenticity of the dataset(C1). Considering the clean interface(C2), multiple simple graphs are designed to enhance the comprehension for the users. For real-time response(C3), the loading time of the initial page is about 30 seconds and the interactions in the tool all can be finished within seconds.

6.2 Use Cases

In this section, we describe some interesting results obtained from some use cases.

6.2.1 Use case 1 - Increasing migration due to marriage

In this use case, we present some general conclusions that can be observed from the tool. First, from the heatmap view, when comparing all the inflows heatmaps and all the outflows heatmaps, we can summarize the trend of the distribution of inflows and outflows. As shown in Figure 6.7 (a), the colours in the north part of The Netherlands, and in the neighbouring countries Germany and Belgium, are darker and darker as time goes by. As the color legends of heatmaps are not unified, we can only conclude that the percentage of the number of people move out from these areas by marriage is increasing. Besides, as shown in Figure 6.7 (b), we cannot find any obvious trend inflow heatmaps.



(a) Comparison of all outflow heatmaps

(b) Comparison of all inflow heatmaps

Figure 6.7: Default heatmaps with labels to show the time periods

Next, we observe the charts in the sidebar in the default view, which correspond to the general distribution of different attributes in the whole dataset. As shown in Figure 6.8 (a), the number of people move by marriage each year shows an overall upward trend, and the number of people with missing birth date is also increasing. As shown in Figure 6.8 (b), the distribution of the birth year of bride and groom are generally similar. The range of the X-axis illustrates that there exists a birth year that is larger than 2150, which represents a possible error record.

Figure 6.8 (c) shows that the number of brides under 30 years old is larger than the number of grooms, while the groups of over 30 years old give the opposite conclusion. Besides, in all the groups, the average ages of grooms are larger than the average ages of brides. For example, the average age of brides from 20 to 30 is 24.76, and that of grooms is 25.46. In conclusion, compared to grooms, brides tend to get married at an earlier age. Figure 6.8 (d) and (e) show the top 10 birth locations and marriage locations in the marriage dataset, and we observe that the top 3 birth locations are Tilburg, Breda, and 's-Hertogenbosch and the top 3 marriage locations are Breda, Tilburg and Eindhoven.

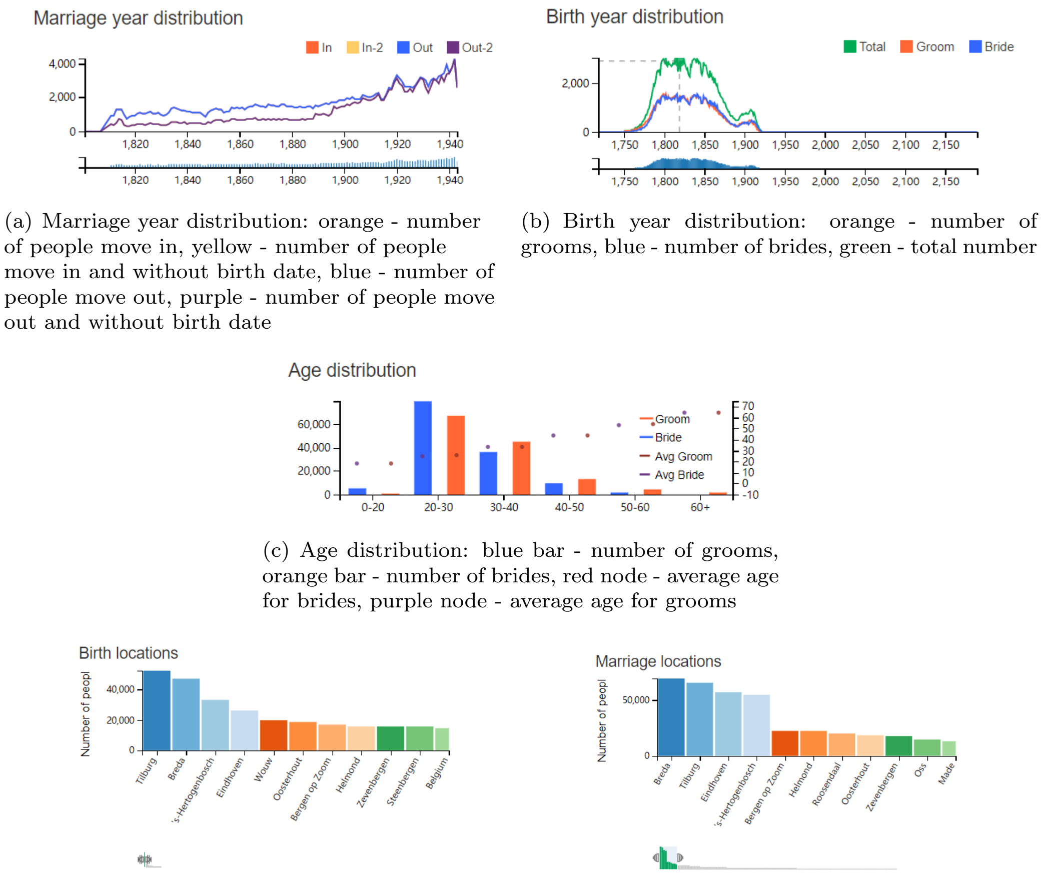
6.2.2 Use case 2 - Marriage mobility of foreign countries

In this use case, we present some detailed distributions of marriage mobility that are related to foreign countries.

First, there are two foreign countries where the total marriage number is more than 5000, which are Germany and Belgium. Since we have discussed Germany in the first section, we only observe the distribution of marriage mobility of Belgium. After selecting Belgium in the map, the detailed information in the sidebar is shown in Figure 6.9. The number of people maintained a stable level until 1914, and rapidly increased and then fell back during 1914 and 1920, and then showed stability again with a higher number than before. From the birth year and age distribution, we observe that there are more grooms tended to get married in The Netherlands, and birth years are mostly concentrated from 1790 to 1820. Moreover, the top 5 locations that they tend to marry are Breda, Tilburg, Eindhoven, Baarle-Nassau and Roosendaal.

As shown in Figure 6.10 (a), we select a time period around the World War I in the marriage year distribution. As shown in Figure 6.10 (b), the top 5 locations that the people from Belgium tended to marry are Uden, Tilburg, Eindhoven, Breda and Roosendaal. By checking other sources, we find that during the World War I, refugee camps were erected in Uden, Baarle Nassau, Tilburg, Bergen op Zoom, and Roosendaal. Thus, the marriage mobility of Belgium is possibly influenced by the World War I.

Second, we discuss the marriage mobility of some other countries. As shown in Figure 6.11 (a), we select all the countries where the total marriage number is more than 50, including Switzerland, France, Indonesia, Austria, Poland, United Kingdom, Czech Republic and Italy. The marriage



(a) Marriage year distribution: orange - number of people move in, yellow - number of people move in and without birth date, blue - number of people move out, purple - number of people move out and without birth date

(b) Birth year distribution: orange - number of grooms, blue - number of brides, green - total number

(c) Age distribution: blue bar - number of grooms, orange bar - number of brides, red node - average age for brides, purple node - average age for grooms

(d) Birth locations (top 10)

(e) Marriage locations (top 10)

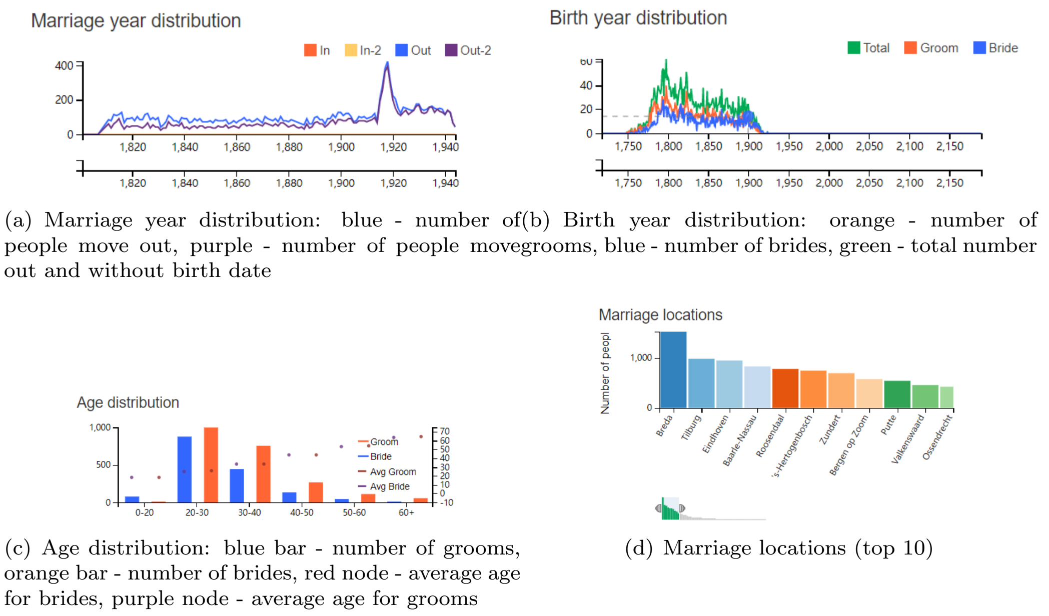
Figure 6.8: Charts in the sidebar in default status

mobility of these countries is illustrated in the Figures 6.11(b)-(e). The number of people was relatively high between 1830 and 1840, as well as between 1920 and 1940. More foreign grooms tended to get married in The Netherlands than brides, and most of these grooms were born during 1780 and 1810. The top 3 locations that they tended to marry are 's-Hertogenbosch, Eindhoven and Breda. As the Industrial Revolution spread to the entire European in the 19-th century, we assume that he marriage mobility of these countries is influenced by the Industrial Revolution and the World War I.

6.2.3 Use case 3 - Marriage mobility of locations in the Noord-Brabant

In this use case, we focus on marriage mobility inside The Netherlands. To eliminate the influence of local marriage, we filter all the records that have different birth and marriage locations. Next, as shown in Figure 6.12 (a), we select all the locations where the number of inflows is more than 5000 in the Noord-Brabant province, and the marriage mobility of these locations is illustrated in other five figures. The number of people in Figure 6.12 (b) kept at a low level in the 19-th century, and increased quickly in the 20-th century. And the number of people move in is larger that the people move out, which represent that many people tend to get married in these locations.

From the birth year and age distribution, there are more grooms tended to get married than brides in these locations in 19-th and 20-th centries. From the birth locations, the number of people married in these locations and did not have clear birth locations is relatively high, and the number of people that were born in Wouw and Breda and married in these locations is also high. Moreover,



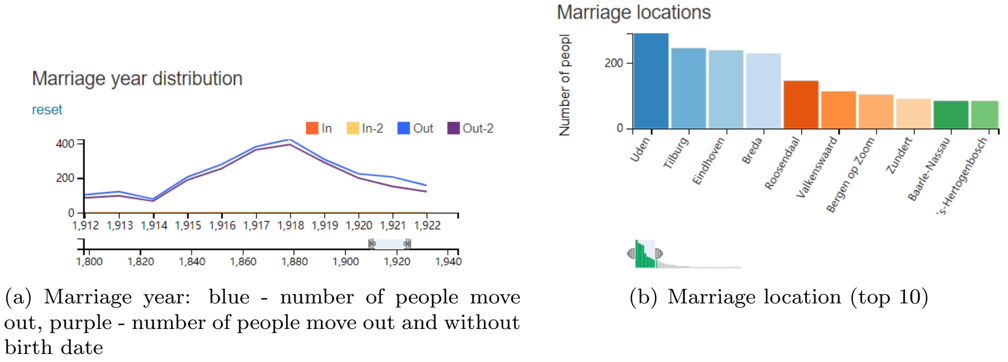
(a) Marriage year distribution: blue - number of people move out, purple - number of people move out and without birth date

(b) Birth year distribution: orange - number of grooms, blue - number of brides, green - total number

(c) Age distribution: blue bar - number of grooms, orange bar - number of brides, red node - average age for brides, purple node - average age for grooms

(d) Marriage locations (top 10)

Figure 6.9: Marriage mobility from Belgium to Noord-Brabant

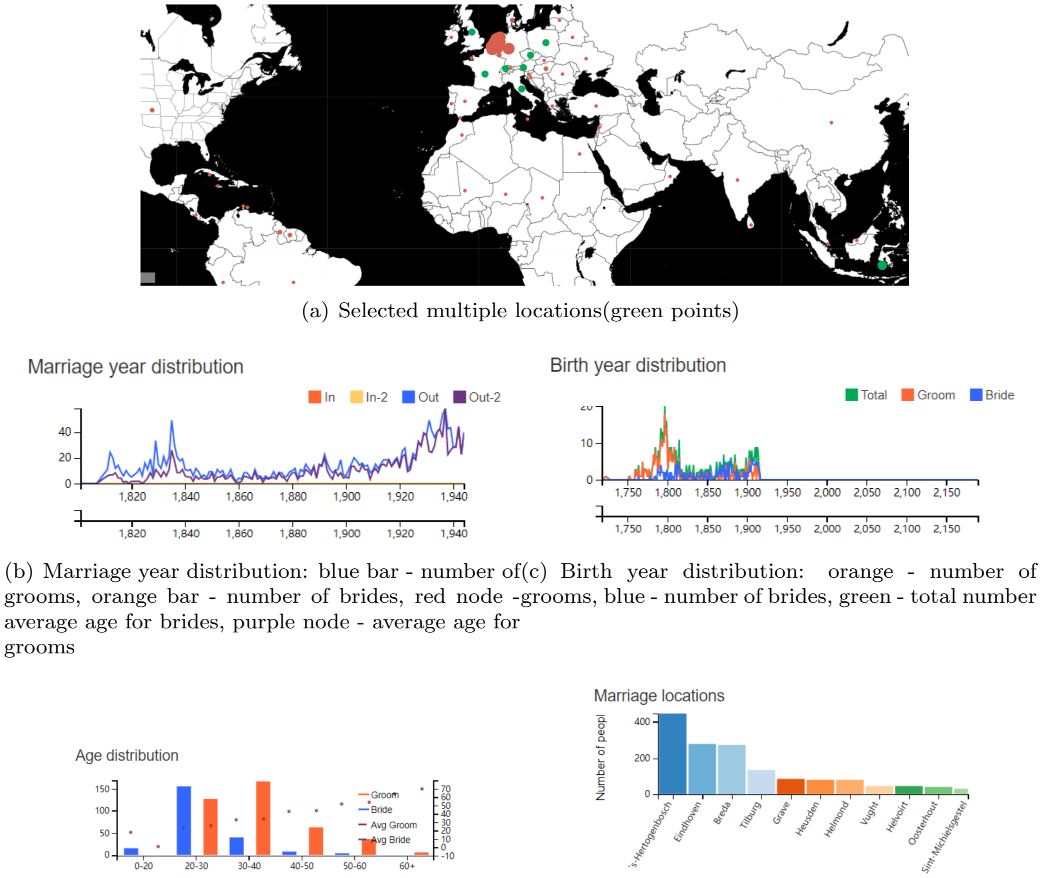


(a) Marriage year: blue - number of people move out, purple - number of people move out and without birth date

(b) Marriage location (top 10)

Figure 6.10: Marriage mobility from Belgium to Noord-Brabant during 1912 and 1922

the top 3 locations that they tended to marry are Eindhoven, Breda and 's-Hertogenbosch.



(a) Selected multiple locations(green points)

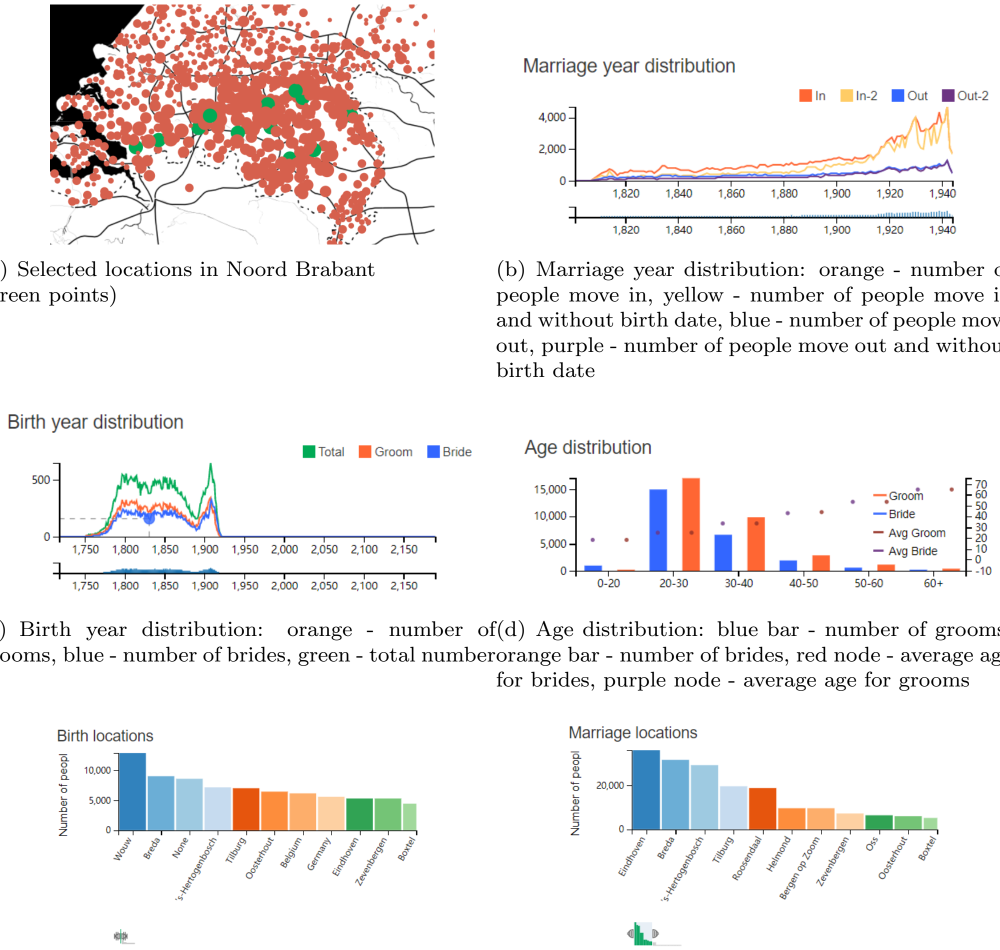
(b) Marriage year distribution: blue bar - number of grooms, orange bar - number of brides, red node - average age for brides, purple node - average age for grooms

(c) Birth year distribution: orange - number of grooms, blue - number of brides, green - total number

(d) Age distribution: blue bar - number of grooms, orange bar - number of brides, red node - average age for brides, purple node - average age for grooms

(e) Marriage locations (top 10)

Figure 6.11: Marriage mobility of some foreign countries where the total marriage number is more than 50, including Switzerland, France, Indonesia, Austria, Poland, United Kingdom, Czech Republic and Italy.



(a) Selected locations in Noord Brabant (green points)

(b) Marriage year distribution: orange - number of people move in, yellow - number of people move in and without birth date, blue - number of people move out, purple - number of people move out and without birth date

(c) Birth year distribution: orange - number of grooms, blue - number of brides, green - total number

(d) Age distribution: blue bar - number of grooms, orange bar - number of brides, red node - average age for brides, purple node - average age for grooms

(e) Birth locations (top 10)

(f) Marriage locations (top 10)

Figure 6.12: Marriage mobility of some locations where the number of inflows is more than 5000 in the Noord-Brabant province, including Boxmeer, Helmond, Eindhoven, Zevenbergen, Oss, Boxtel, 's-Hertogenbosch, Oosterhout, Tilburg, Bergen op Zoom, Roosendaal and Breda.

Chapter 7

Conclusion and Future Work

7.1 Conclusion

With some existing BHIC datasets, we select to visualize the marriage mobility in the marriage dataset. Current techniques, which are simply searching and showing records in text, are not sufficient to provide a satisfactory insight. Therefore, a visualization tool for analysis of marriage mobility patterns is designed and implemented in this project. The tool contains four views and a filter bar to filter the data in all views. Each view shows different attributes and multiple interactions are provided for users to explore the dataset intuitively.

Reviewing the multiple objectives and following questions that we mentioned in Chapter 1, we find that they are answered in the subsequent chapters. For the first question, *What kind of information is recorded in the dataset?*, the attributes of the marriage dataset are introduced. Based on these attributes, combined with the requirements from BHIC experts, multiple questions and corresponding requirements are proposed. For the second question, *How to deal with incomplete and erroneous data?*, the detailed information of each attribute is analysed and we find that there is incomplete and erroneous data. Based on the principle of keeping authenticity, we select to fill the incomplete data with 0 values. Besides, with the help of external datasets, we have matched part of locations to their coordinates. Then, the matched ones are designed to be presented in the matrix and the map, and the unmatched ones are presented only in the matrix.

For the third question, *How to concurrently visualize a large number of marriage movements?*, we select the matrix and map as the visualization method to concurrently show a large number of marriage movements. To make the views more visible, the locations in the matrix with similar marriage mobility patterns are clustered and reordered in adjacent rows. Besides, to avoid cluttered lines in the map, movements on the map are initially hidden and only shown when users click locations. For the fourth question, *How to visualize this multi-variate dataset?*, multiple views are designed to visualize this multi-dimensional dataset. Four views are implemented to show different attributes. The matrix view shows all the locations and flows between each pair of locations, the map view shows part of locations and flows between selected locations, the heatmap view shows the change of inflow numbers and outflow numbers in different periods, and the sidebar view shows different charts for other attributes. In this way, all the attributes are statistically analysed in at least one view. For the last question, *How to design a user-friendly dashboard and interaction model?*, to improve the usability of the tool, we try to design all the views and graphs more intuitively and easy to operate. Also, we add a lot of interaction inside each view and between views to show more information and connect the information in different views. Finally, response times for the interactions are considered in the design to ensure good performance of the tool.

From the current result, the tool can answer all the questions and satisfy all the requirements we propose. It is possible to interactively explore some interesting marriage mobility patterns with the tool for users. In conclusion, a marriage mobility analysis tool is designed and implemented in this project, and some possible ideas are still open for the tool.

7.2 Future Work

Due to the time limitation and the scale of the project, there are some ideas that can be feasible for the extension of the project. In this chapter, we introduce some promising ideas for the future work.

7.2.1 Linkage of Datasets

Except for the marriage records, BHIC datasets also contain many other important events in life, such as birth and death. However, as the person ids inside and between datasets are unrelated, it is impossible to link different records between these datasets. If we can link the records between different datasets, it is possible to draw the path of one's entire life. For example, the start point is the place that someone was born, and the next points are the places of important events, such as marriage or military service, until the end point of death. Besides, there is information about the parents of bride and groom in the marriage dataset, however, they are not used in our project. If we can apply some algorithms to link the same person inside the marriage dataset, we can obtain mobility pattern for a family.

We also find that the information in the XML files that we can obtain from the open dataset is not complete in the sense that the original documents contain more information. For example, in the population register dataset, the original documents have the information of where a person moved from and where a person moved to, while the XML files do not have that information. If this kind of information can be added, the mobility pattern of a person will be more detailed and precise. Also, if more attributes are taken into account, such as the military records, the tool may find more complete mobility patterns, instead of only mobility patterns based on marriages.

Moreover, if we extract more attributes from the XML file, the analysis can be more complete. For example, in the marriage dataset, if we also extract the ages, they can be the supplement for the missing data in birth dates.

7.2.2 Visualization of the tool

For the different views in the tool, we discuss limitations and possible methods to lift them.

Representation of locations

As the division of the municipalities in The Netherlands were dramatically changed during 19-th and 20-th centuries, and the borders of the municipalities of certain years are not easily available, we choose to use points to represent the locations. However, a point is less clear and accurate for users than a border. Thus, if the borders of municipalities of different times are available, maps of different times can be added as multiple map layers so that the information in the map will be more precise. Moreover, with the borders to represent the locations and using choropleth map, as we discussed in Chapter 3, the flows between different locations can be represented by the shade of colors when clicking, which is also more intuitive than the curves in the map.

Alternative cluster and reorder methods

For the cluster part in the matrix view, there are some alternative methods. First, the locations can be clustered by the geographical divisions of The Netherlands. For example, a COROP region (Coordination Commission Regional Research Programme) [8], which is a division of The Netherlands for statistical purpose proposed by CBS (Central Agency for Statistics), is widely used. However, as historical location names are used in the BHIC dataset, it is a lot of work to match the historical names to other divisions of areas in The Netherlands. Second, we use the maximum value as the element to determine the clustered result for now, while the locations can be clustered by different algorithms. For example, some library, such as scikit-learn, provides clustering algorithms, such as K-means [16]. However, with a more complex algorithm, the calculation time might be problematic.

Moreover, for the reorder part in the matrix view, we select to apply Hamming distance among three methods that result in rectangular patterns. However, it is better to provide a choice box for users to switch the methods, so that they can choose the most appropriate method.

Other details

Due to the limitations of the chosen libraries and the running speed, some details are not implemented in the tool, but can be considered to be added in the future. First, the color scales in the heatmaps are currently independent. However, a unified color scale can make the heatmap more understandable. And a legend in the corner of each heatmaps can be added to illustrate the colors. Second, a synchronized zooming function between different heatmaps are convenient for users to check the mobility. This function can be helpful for the comparison between the heatmaps of different times.

Bibliography

[1] https://www.bhic.nl/over-ons/het-bhic. Accessed: 2020-03. 1

[2] https://opendata.picturae.com/organization/bhic. Accessed: 2020-03. 5

[3] https://www.gemeentegeschiedenis.nl/. Accessed: 2020-06. 15

[4] https://geonames.nga.mil/gns/html/. Accessed: 2020-06. 18

[5] https://simplemaps.com/data/world-cities. Accessed: 2020-06. 18

[6] https://gist.github.com/tadast/8827699. Accessed: 2020-06. 19

[7] https://treaties.un.org/pages/ViewDetails.aspx?src=TREATY&mtdsg_no=XVI-3&chapter=16&lang=en. Accessed: 2020-08. 34

[8] https://web.archive.org/web/20160304001125/http://www.cbs.nl/NR/rdonlyres/5243D99C-0F3F-43C1-BC38-ADEC2FDD9DEB/0/2012indelingcoropkaart.pdf. Accessed: 2020-09. 50

[9] Vladimir Agafonkin. Leaflet.heat. https://github.com/Leaflet/Leaflet.heat, 2014. 33

[10] Robert Ball. Visualizing genealogy through a family-centric perspective. *Information Visualization*, 16(1):74–89, 2017. 1

[11] Michael Behrisch, Benjamin Bach, Nathalie Henry Riche, Tobias Schreck, and Jean-Daniel Fekete. Matrix reordering methods for table and network visualization. *Comput. Graph. Forum*, 35(3):693–716, 2016. 28, 29

[12] Lonni Besançon, Matthew Cooper, Anders Ynnerman, and Frédéric Vernier. An evaluation of visualization methods for population statistics based on choropleth maps. *CoRR*, abs/2005.00324, 2020. 11

[13] Mike Flaxman Daniel Wiesmann. Space sankey. https://github.com/geodesign/spatialsankey, 2014. 31

[14] Mohammad Ghoniem, Jean-Daniel Fekete, and Philippe Castagliola. On the readability of graphs using node-link and matrix-based representations: a controlled experiment and statistical analysis. *Information Visualization*, 4(2):114–135, 2005. 11, 12

[15] Danny Holten and Jarke J. van Wijk. Force-directed edge bundling for graph visualization. *Comput. Graph. Forum*, 28(3):983–990, 2009. 10

[16] Tapas Kanungo, David M. Mount, Nathan S. Netanyahu, Christine D. Piatko, Ruth Silverman, and Angela Y. Wu. An efficient k-means clustering algorithm: Analysis and implementation. *IEEE Trans. Pattern Anal. Mach. Intell.*, 24(7):881–892, 2002. 51

[17] Yuhua Liu, Sicheng Dai, Changbo Wang, Zhiguang Zhou, and Huamin Qu. Genealogyvis: A system for visual analysis of multidimensional genealogical data. *IEEE Trans. Hum. Mach. Syst.*, 47(6):873–885, 2017. 10

[18] Michael J. McGuffin and Ravin Balakrishnan. Interactive visualization of genealogical graphs. In John T. Stasko and Matthew O. Ward, editors, *IEEE Symposium on Information Visualization (InfoVis 2005), 23-25 October 2005, Minneapolis, MN, USA*, pages 16–23. IEEE Computer Society, 2005. 9

[19] Carolina Nobre, Nils Gehlenborg, Hilary Coon, and Alexander Lex. Lineage: Visualizing multivariate clinical data in genealogy graphs. *IEEE Trans. Vis. Comput. Graph.*, 25(3):1543–1558, 2019. 9

[20] Jamie R. Nuñez, Christopher R. Anderton, and Ryan S. Renslow. Optimizing colormaps with consideration for color vision deficiency to enable accurate interpretation of scientific data. *PLOS ONE*, 13(7):1–14, 08 2018. 29

[21] Jonathan C. Roberts. Chapter 8 - exploratory visualization with multiple linked views. In Jason Dykes, Alan M. MacEachren, and Menno-Jan Kraak, editors, *Exploring Geovisualization*, International Cartographic Association, pages 159 – 180. Elsevier, Oxford, 2005. 25

[22] Christian Rohrdantz, Michael Hund, Thomas Mayer, Bernhard Wälchli, and Daniel A. Keim. The world's languages explorer: Visual analysis of language features in genealogical and areal contexts. *Comput. Graph. Forum*, 31(3):935–944, 2012. 9

[23] Nikola Sander, Guy J. Abel, Ramon Bauer, and Johannes Schmidt. Visualising migration flow data with circular plots. Vienna Institute of Demography Working Papers 2/2014, Vienna, 2014. 12, 13

[24] Daniel Shakespear. Interactive genealogy explorer: Visualization of migration of ancestors and relatives. In Serge Ter Braake, Antske Fokkens, Ronald Sluijter, Thierry Declerck, and Eveline Wandl-Vogt, editors, *Proceedings of the First Conference on Biographical Data in a Digital World 2015, Amsterdam, The Netherlands, April 9, 2015*, volume 1399 of *CEUR Workshop Proceedings*, pages 94–100. CEUR-WS.org, 2015. 9, 10

[25] D.J. Timothy and J.K. Guelke. *Geography and Genealogy: Locating Personal Pasts*. Heritage, culture, and identity. Ashgate, 2008. 1

[26] Colin Ware. *Information Visualization: Perception for Design: Second Edition*. 04 2004. 1

[27] Jo Wood, Jason Dykes, and Aidan Slingsby. Visualisation of origins, destinations and flows with od maps. *The Cartographic Journal*, 47(2):117–129, 2010. 11, 13

[28] Yalong Yang, Tim Dwyer, Bernhard Jenny, Kim Marriott, Maxime Cordeil, and Haohui Chen. Origin-destination flow maps in immersive environments. *IEEE Trans. Vis. Comput. Graph.*, 25(1):693–703, 2019. 10

Appendix A

Tables in data pre-processing

Table A.1: Dutch spellings for country names in marriage dataset and their corresponding names in external datasets

Locations	Corresponding names	Locations	Corresponding names
VS	United States	Libië	Libya
Verenigde Staten	United States	Luxemburg	Luxembourg
Amerika	United States	Maleisië	Malaysia
Argentinië	Argentina	Nederlandse Antillen	Netherlands Antilles
Aus	Australia	Noorwegen	Norway
Oostenrijk	Austria	Italie	Italy
Bel	Belgium	Italië	Italy
Belg	Belgium	Java	Indonesia
België	Belgium	Nederlands-Indië	Indonesia
Brazilië	Brazil	Ned-Indië	Indonesia
Groot-Brittannië	United Kingdom	Indonesië	Indonesia
UK	United Kingdom	Indonesie	Indonesia
Engeland	United Kingdom	NOI	Indonesia
Chili	Chile	Indië	Indonesia
Kroatië	Croatia	Ind	Indonesia
Praag	Czech Republic	Ierland	Ireland
CSK	Czech Republic	Polen	Poland
CZK	Czech Republic	Roemenië	Romania
Tsjechië	Czech Republic	Rusland	Russian Federation
Denemarken	Denmark	Schotland	United Kingdom
Egypte	Egypt	Spanje	Spain
Estland	Estonia	Slovenië	Slovenia
Frankrijk	France	Slowakije	Slovakia
Duitsland	Germany	Zuid-Afrika	South Africa
Pruissen	Germany	Zweden	Sweden
Pruisen	Germany	Zwitersland	Switzerland
Hongarije	Hungary	Zwitserland	Switzerland
Litouwen	Lithuania	Oekraïne	Ukraine
Libië	Libya		

Table A.2: Misspelling and alternative spelling location names in marriage dataset and their corresponding names in external datasets

Locations	Corresponding names	Locations	Corresponding names
Aalter ?	Aalten	Reudel	Reusel
Aardt	Aerdt	Rijkwijk	Rijswijk
Aarlanderveer	Aarlanderveen	Roetterdam	Rotterdam
Alkimaar	Alkmaar	Roitterdam	Rotterdam
Beezel	Beesel	Rotteram	Rotterdam
Berchum	Berghum	Sint-Annaland	Sint Annaland
Bergum	Berghum	Sint-Philipsland	Sint Philipsland
Berkal c.a.	Berkel	Slabroeck	Slabroek
Bocholt	Bocholtz	Sluiswijk	Sluipwijk
Borsele	Borssele	Spaarnewoude	Spaarnwoude
Brada	Breda	Spaubroek	Spanbroek
BrĂm	Brem	Tuilburg	Tilburg
Delfy	Delft	Tull	Tuil
Dougjum	Dongjum	Vlaardinger Ambacht	Vlaardingen
Gestel	Gastel	Vreeswijk	Vreewijk
Hoofsplaat	Hoofdplaat	Waaqlwijk	Waalwijk
Hooggezand	Hoogezand	Wehe	Wehl
Huiisseling	Huisseling	Westeremder	Westeremden
Kestern	Kesteren	Wilvega	Wolvega
Maasluis	Maassluis	Windeweer	Winneweer
Meir	Meer	Wow	Wouw
Middeharnis	Middelharnis	Zijbe	Zijpe
Oostehout	Oosterhout	Zuid-Schalkwijk	Zuidschalkwijk
Oostwedde	Onstwedde	Zwartesluis	Zwartsluis
Ostwedde	Onstwedde	oostr	Oost