A step towards the book of life - Bachelor Thesis

Wouter Joosse
4158407

Contents

Chapter 1

Introduction

In modern age computer science, there seems to be an abundance of data. Companies, for example, try to create customer profiles in order to predict the needs of their customers and governments want to get an overview of the people that live in their countries. Mostly this data is structured in *data warehouses*, where each piece of data is stored in a structured way.

Analysis of data often requires that different data sources should be joined, so that the analysis could be improved. However, often problems arise when trying to combine data from different data sources. Data is not saved according to a universal standard. Because of this, each source could save the same information in a different format, making it harder to link this data. For example, dates could be written according to different formats or persons could be mentioned with their full name or with only their initials. The problem of combining records from different data sources and finding data that are about hte same entity is called *Record Linkage* and has been researched extensively in the last few decades. This problem does not only apply to records of people, but to all sorts of entities.

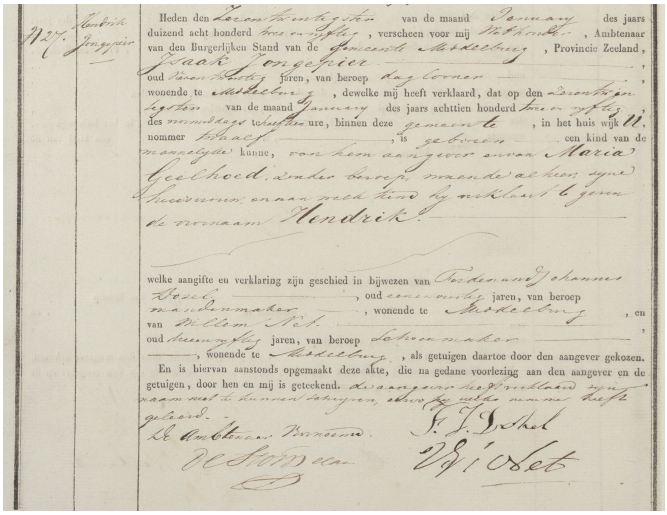
The term *record linkage* was introduced by Dunn [1], who wanted to assemble a 'book of life' for each person which would describe the person's interaction with health and social security systems, and which would also contain the birth-, death- and marriage-certificates of that person. One of the first mathematical models for automatic record linkage was proposed by Fellegi and Sunter in 1969 [2], ...

Throughout history, governments held census to identify civilians and keep records about them, for example in order register taxations and the people that are allowed to vote. So did the government of The Netherlands in the 18^{th} and 19^{th} century. During that time, The Netherlands was occupied by France. The French introduced the 'Burgerlijke stand', who was responsible for recording births, marriages and deaths of the people of The Netherlands. The *genlias* project has started in the last decade of the 20^{th} century with the aim to collect all the certificates that where produced into a database. This database is now available via `http://www.wiewaswie.nl`.

The dataset gives us great insight into the population of the Netherlands in the 19^{th} century. One could, for instance, do research into the migration patterns of that time or do research into the social characteristics of the population. However, unlike more modern registration practices, the citizens of the Netherlands where not provided with unique identification numbers, which makes it harder to do research. Another complicating factor is that it was not uncommon for officials to write different names for the same person on different registrations. For instance, when people married, the persons involved would be registered with the official names. However, when a person passed away, it could happen that the neighbors had to report this event to the civil registration. These neighbors didn't always know the full official name of the person deceased. This is where record linkage could be applied, in order to provide the dataset with unique identifiers per person.

The subject of this bachelor thesis are the problems that can arise when trying to build life courses out of this historical data. In [3] an method was proposed in order to link registration certificates of these life events from the *genlias* project. The (distinct) persons in this database have not been identified by a unique identification number. Therefore, matching of records is done by looking at the names that are provided in the certificates. A link between certificates is made when the names on the certificates correspond to each other (in some degree). This was done by creating an index-tree for all the records, based on

Figure 1.1: An example of a birth-certificate. Source: Zeeuws Archief, http://www.archieven.nl



the occurrence of letters in the names of the persons. This approach groups the names of the persons that are similar to each other together into the same leaf-nodes of the tree, given a certain levenshtein-distance.

My aim is to assess the quality of the linkages made by this approach and construct life courses for the people in the province of Zeeland. I will then try to improve the linkage method by looking at the certificates of events that have been missed (i.e. if there is no death certificate matched, it could be the case that this certificate is not present, or that the method wrongfully not matched the certificate), and by looking at life courses that are very tightly coupled (i.e. someone has married three times which might could be correct, but this is not very likely). I will try to find rules that could improve the linkage procedure when trying to link historical name data.

Chapter 2

Theoretical Background

Chapter 3

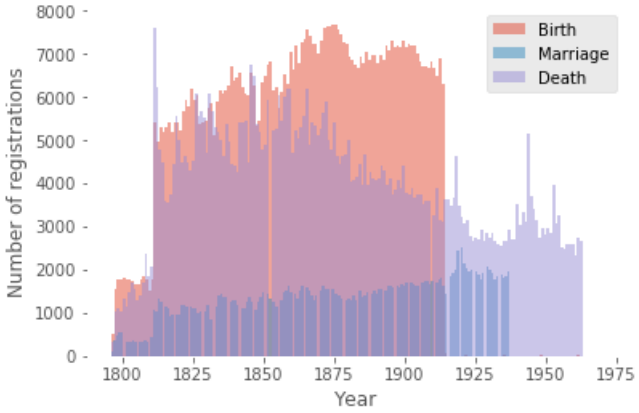
Data

3.1 Description of the records

The dataset provided by the LINKS project is the set of all historical certificates from the province of Zeeland. The dataset contains 1.558.205 distinct registrations, which are split out into 698.285 birth-certificates, 193.921 marriage-certificates (this type also includes some divorce-certificates) and 665.999 death-certificates. Due to privacy regulations, birth certificates are available until 1913, marriage registrations are available until 1938 and death registrations are available until 1963.

The original document consists of an handwritten pre-printed card on which the official filled in the specific details of the event. In earlier versions, the entire document was handwritten. An example of a birth-certificate can be seen in figure 1.1

Figure 3.1: The number of registrations per year, split by certificate-type. Birth-certificates are available until 1913, marriage-certificates until 1938 and death-certificates until 1963



The digitalized certificates are stored across three different tables:

1. Persons, which contains data concerning the people mentioned on a certificate,
2. Registrations, containing meta-data (such as date and the location id),
3. Locations, containing a mapping from location id to names of locations

The Persons table contains records for each person mentioned in a certificate. The number of persons mentioned on a certificate depends on the type of certificate. On birth-certificates, there are three persons mentioned: ego, the mother and the father (although, this is not always the case). On marriage-certificates, there are six persons mentioned: the bride and groom, and the mothers and fathers of the bride and groom. On death-certificates, at least the deceased and his/her parents are mentioned. If the deceased had a partner, in the general case, this parter is also mentioned, but this is not always the case. For each person, the role is also specified, (e.g. bride, groom, father of the bride, mother of the groom,. etc).

Other information that is also recorded in the database include the place of the event and the date. We also added the latitude and longitude of the places to the dataset.

In table 3.1, you can find an example of how a single birth-certificate is stored. Not all possible values are displayed. The same pattern is also valid for both marriage- and death-certificates, but then the date is stored in mar_day, mar_month, mar_year and their respective death-counterparts. By combining the records on the id of the registration, you can create a single record of the entire certificate.

3.2 Population characteritics

Hier komt een beschrijving van de bevolkingsopbouw en een verdeling van de leeftijden waarop events plaatsvinden.

3.3 Preprocessing

The first step is to extract the data that is needed to match the certificates. In order to match the marriage certificates of a couple with the marriage certificate where that couple acts as one of pairs of parent, the certificate id, first names and family-name of the ego and ega and the first names and family-name of their parents where extracted from the dataset. In the case that a person has multiple first names, not all the first-names were used, but only the first first-name. Also, any prefixes to the family-names were dropped.

In the Netherlands, women keep their maiden name when they marry, so in total there are four different last names. This increases the unique combinations of persons and is helpful when matching certificates.

Table 3.1: An overview of the field in a birth certificate

id_person	id_registration	registration_maintype	firstnames	prefix	familyname	sex	civil_status	birth_day	birth_month	birth_year
15405	5136	1	maria cornelia	van	oorsel	f	1	17	9	1864
15406	5136	1	willem hendrik	van	oorsel	m	3			
15407	5136	1	neeltje johanna		christiaanse	f	2			

Chapter 4

Method

Chapter 5

Results

Chapter 6

Discussion

Chapter 7

Conclusion

List of Figures

List of Tables

Bibliography

[1] Halbert L Dunn. "Record linkage". In: *American Journal of Public Health and the Nations Health* 36.12 (1946), pp. 1412–1416.

[2] Ivan P Fellegi and Alan B Sunter. "A theory for record linkage". In: *Journal of the American Statistical Association* 64.328 (1969), pp. 1183–1210.

[3] Marijn Schraagen. "Aspects of Record Linkage". English. PhD thesis. Leiden: Universiteit Leiden, Nov. 11, 2014.