Probabilistic record linkage with the Fellegi and Sunter framework

Using probabilistic record linkage to link privacy preserved police and hospital road accident records

Jonathan de Bruin

Master of Science Thesis

TU Delft Delft University of Technology

Probabilistic record linkage with the Fellegi and Sunter framework

Using probabilistic record linkage to link privacy preserved police and hospital road accident records

Master of Science Thesis

For the degree of Master of Science in Applied Mathematics at Delft University of Technology

Jonathan de Bruin

October 28, 2015

Faculty of Electrical Engineering, Mathematics and Computer Science · Delft University of Technology

The work in this thesis was supported by the Stichting Wetenschappelijk Onderzoek Verkeersveiligheid (SWOV). Their cooperation is hereby gratefully acknowledged.

Abstract

Record linkage is the procedure of bringing together information from two or more records that are believed to belong to the same entity. The linking of a pair of records without identifier should be based on attributes both records have in common. The framework described by Fellegi and Sunter [1969] can be used for record linkage. This record linkage framework classifies pairs of records as links, non-links or possible links based on a comparison of the attributes found in both records. In this thesis, the framework is studied and used to link privacy preserved police and hospital road accident records.

The Fellegi and Sunter record linkage framework classifies the pairs of records based on statistical and probabilistic principles. The framework is based on, what they call, linkage rules. The rules map vectors with the comparison of attributes into the probability of being a link, non-link or possible link. The linkage with these linkage rules can be based on random decisions. The mathematical formulation of the framework in the original paper by Fellegi and Sunter contains some weaknesses. This thesis improves the notation of the framework. The framework is also closely related with statistical hypothesis testing. The relation is studied in depth.

A simulation study in this thesis shows that the Fellegi and Sunter framework is effective for linking records. The framework can also be used to estimate the number of incorrect classifications. The simulation study shows that the number of misclassifications can be estimated well. The Fellegi and Sunter framework needs parameters for classification. The parameters depend on the unknown link and non-link distributions and parameters such as the probability that a randomly picked record pair is a link. The record linkage problem is an incomplete data problem where the comparison of attributes is known, but not whether the pair belongs to the same entity or not.

The EM-algorithm is used to estimate the parameters. This iterative algorithm is used to find maximum likelihood estimates of the parameters in this model. The plain EM-algorithm in the context of record linkage can be used, but it has many drawbacks because the maximisation step is not in closed form. The ECM-algorithm simplifies the computation of the maximisation step in the EM-algorithm by undertaking the maximisation conditional on some

of the parameters. These constraints involve that the comparison of attributes are independent given the (unknown) link status. The simulation study shows that the ECM-algorithm is an effective algorithm for parameter estimation. The convergence properties of the algorithm are good.

The Fellegi and Sunter framework is well known for its possibility to use the value of an agreeing comparison of an attribute for additional distinguishing power (to distinguish the set of links from the non-links). For example, a common surname is not very informative for the classification, while a rare surname is. The Fellegi and Sunter framework can use this in the classification. In this document, a new estimation method is presented which can estimate these parameters. The estimation method is based on the EM-algorithm. The results show that this classification and estimation method adds additional distinguishing power to the classification. The convergence properties of this algorithm deserve additional research.

The Fellegi and Sunter framework and the described estimation methods are used to link privacy preserved police with hospital road accident records. All the records represent a road casualty involved in a road accident in the Netherlands. The Stichting Wetenschappelijk Onderzoek Verkeersveiligheid (SWOV) performs a deterministic record linkage on this data. There is no knowledge about the correct links between the files. In this thesis, the framework of Fellegi and Sunter is used to validate their record linkage model. The results in this thesis show similar results for both methods.

The number of links between the police and hospital road accident records shows the same trend over the years for both methods. Overall, the classification with the Fellegi and Sunter framework and parameters estimated with the ECM-algorithm links slightly more records. Each year, the same vector of attribute comparisons divides the links and the non-links. It was found that the number of records with this vector plays an major role in the classification. The parameter estimates are sensitive to the number of record pairs in this class. The solution was found in a preselection of record pairs.

Record linkage can be applied for many purposes, also in road safety research. This thesis shows that record linkage can be used to link news articles found on the Internet with police road accident records. These articles can be used to improve the road safety analysis because they contain additional information like photos and circumstantial information. A manual review of the record linkage showed that there is much space for improvement. Especially on the extraction of information from the news article into a record.

Table of Contents

Preface

During my study, I followed many classes in the field of safety and security. I become more and more interested in safety tradeoffs, countermeasures and risk analysis. Graduating on a subject about safety or security was a logical step. Prof. dr. ir. G. Jongbloed (First Reader) brought me in contact with the Dutch institute for road safety research *Stichting Wetenschappelijk Onderzoek Verkeersveiligheid*. At this institute, road safety researcher and mathematician Dr. F.D. Bijleveld (Third Reader) introduced me in the field of record linkage. He gave me the paper of Fellegi and Sunter [1969] and said: "We need to do something with this." The result is this document. I would like to thank Prof. dr. ir. G. Jongbloed (First Reader) and Dr. F.D. Bijleveld (Third Reader) for their enthusiasm and help.

I would also like to thank Dr.ir. M.B. van Gijzen (Second Reader) for reading my thesis and for being part of my graduation committee, the Stichting Wetenschappelijk Onderzoek Verkeersveiligheid for the graduation possibilities and the road safety researchers for their help. Of the researchers, I would like to point out the help and feedback of drs. N.M. Bos, dr. S. Houwing and dr. H.L. Stipdonk.

My parents and sister deserve their own paragraph. They were indispensable.

Delft, University of Technology — Jonathan de Bruin
October 28, 2015

Chapter 1

Introduction

There is an ongoing process of storing more and more data [Hilbert and López, 2011]. Massive amounts of data are stored by organisations like governments, businesses, hospitals, intelligence agencies and universities. Information systems containing those data are valuable sources of information. There is often a need to integrate and merge data from multiple sources in these information systems [Christen, 2012*a*]. For example, multiple data sources can contain information about identical persons or businesses. Linking these data sources can improve the data quality and enrich the data sources. Linking data about identical persons or businesses (in general: entities) between multiple data sources is called *record linkage* or *data matching*.

Record linkage is relatively straightforward when the records in the data sources contain (unique) entity identifiers like national identification numbers, ISBNs and consumer product codes [Christen, 2012*a*]. If identifiers are not present, the characteristics of the entity may make it possible to link the data sources. For personal information, examples of characteristics are; (sur)name, date of birth, place of birth, sex and hair colour. If most characteristics in the records are identical, then the records probably belong to the same person. There are several mathematical methods to link information between two data sources (semi-)automatically based on the characteristics of the entities. The most-well-known methods link the data based on deterministic and probabilistic principles.

In this thesis, the probabilistic record linkage framework proposed by Fellegi and Sunter [1969] will be studied. Fellegi and Sunter provide a record linkage classification framework based on statistical and probabilistic principles. The framework has some weaknesses in its mathematical formulation. This thesis will pay extra attention on the mathematical formulation of the framework. The framework relies to a number of parameters that depend on the 'true' record linkage of the data sources. In general, these parameters need to be estimated because they are unknown. This thesis will describe several estimation methods for the parameters of interest in the Fellegi and Sunter framework. Also, a new estimation method will be proposed.

Record linkage is used for linking, enriching and cleaning of data. Therefore, record linkage is valuable for many organisations and research institutes. It is successfully used in many applications and research fields. A few of the many examples are; merging and deduplication of national census data [Winkler, 1999], detecting national security and terrorism threats by national intelligence organisations [Gomatam and Larsen, 2004], detecting bank fraud [Christen, 2012*a*], privacy-preserved linking of records for health research [CHeReL, 2015] and merging client information for advertisement proposes [Christen, 2012*a*]. In this thesis, record linkage will be applied on a problem in road safety research. The record linkage framework of Fellegi and Sunter [1969] will be used to perform a privacy preserved record linkage with police and hospital road accident data.

For road safety research, it is important to have data available about road accidents. In nearly all countries, road accidents are registered by the police [Amoros et al., 2011]. The police data about road accidents play an important role in the analysis of road accidents and the signalling of trends. The analyses and signals are used to develop effective countermeasures [Rosman, 2001]. The road accidents recorded by the police are not enough for a good analysis of road safety [IRTAD, 2011]. The medical consequences for the road accident victim are interesting for analysis. Therefore, it is interesting to link police road accident data to hospital data. In practice, linking police and hospital data is not straightforward because the personal information is privacy-sensitive. In most countries where record linkage is used for road safety analysis, the hospital and the police road accident data are made anonymous before they become available for research [IRTAD, 2011]. The anonymous data needs be linked based on the stored characteristics of the road casualty found in the police and hospital data. One can think about characteristics such as date of birth, sex, mode of transport and location of the accident.

The Dutch road safety institute Stichting Wetenschappelijk Onderzoek Verkeersveiligheid (SWOV) is the main institute for road safety research in the Netherlands. Each year, the SWOV performs a record linkage of police and hospital data. This record linkage is used for analysis of road safety. The SWOV makes use of a deterministic record linkage model based on mathematical distance measures [Reurings and Bos, 2009]. In general, if the distance between a police and hospital record is small, then the records probably belong to the same road casualty. In this thesis, supported by the SWOV, Dutch police road accident records are linked with Dutch hospital records. For this linkage operation, the probabilistic record linkage framework of Fellegi and Sunter [1969] is used.

The primary reason for the SWOV to conduct this study is to validate the current model for linking police and hospital road accident records. If the probabilistic record linkage model in this thesis shows reasonable similar linkage results, then the credibility of the currently used distance-based approach increases. This validation is important because the 'true' record linkage is not available. For this thesis, the main question is: Is the probabilistic record linkage framework of Fellegi and Sunter [1969] useful to link police and hospital road accident records? If it is useful, what are the advantages and disadvantages compared with the current method?

This thesis consists of two parts. The first part, Part I, will give a theoretical basis for probabilistic record linkage and the second part, Part II, will be used for simulations and applications. Chapter 2 will give a global overview of the record linkage methodology. The record linkage methodology involves some steps that play an important role in the linkage

process. In Chapter 3, the record linkage framework of Fellegi and Sunter [1969] will be described. A large focus will be on improving the formulation and notation. The final chapters in Part I present estimation methods for the parameters of interest in this model. The Expectation-Maximisation algorithm will play an important role in the estimation of parameters. In Section 5-4 is a generalisation of an estimation method based on the Expectation-Maximization algorithm developed and presented.

The second part of this thesis, Part II, will focus on applications of the Fellegi and Sunter framework on data. In Chapter 6, a simulation study will be used to explore the behaviour of the framework and the estimation methods. After this simulation study, the Dutch police road accident records will be linked with Dutch hospital data. The methods and results are compared with the current record linkage by the SWOV and discussed in Chapter 7. The possibilities of record linkage for road safety research are very broad. Chapter 8 will show how (online) news articles about road accidents can be linked to police records with the Fellegi and Sunter framework.

Part I

Theory

Chapter 2

Record Linkage

2-1 Introduction

The term *record linkage* is used to indicate the procedure of bringing together information from two or more records[1] that are believed to belong to the same entity [Newcombe et al., 1986] [Yancey, 2002] [Herzog, Scheuren and Winkler, 2007]. Record linkage is used to link data from multiple data sources or to find *duplicates* in a single data source. In computer science, record linkage is also known as *data matching*. Data matching does not restrict the data structures to records. In this document, the term record linkage is used while the term 'data matching' also satisfies.

The idea of record linkage was introduced in the mid-1900s. As far as known, Dunn [1946] was the first to use the term record linkage in his 'Book of Life' concept. A Book of Life is a personal book that starts at birth and ends with death. Records of the principal events in life, such as marriage and graduation, fill the pages. In this context, record linkage is the process of assembling the pages of the person into a volume. After this conceptual introduction of record linkage, Newcombe et al. [1959] started to extend this concept of record linkage in a mathematical way.

In record linkage, the attributes of the entity (stored in a record) are used to link two or more records. Attributes can be (unique) entity identifiers, but also attributes like (sur)name, sex, date of birth and hair colour. If the identifier of two records[2] is identical, then the records (highly likely) belong to the same entity. In general, record linkage is seen as the process of linking records for which these unique identifiers are not available. The data needs to be linked based on attributes with less distinguishing power such as (sur)name, sex, date of birth and hair colour.

[1] A record is a simple, structured computer storage object that contains information about the attributes of an entity. (See Section 2-3)

[2] Assume that both records contain the same type of identifier. For example, the Social Security Number [Puckett, 2009].

The process of generating records for entities is subject to errors. Therefore, comparing attributes of two records sometimes disagrees while the records belong to the same entity. For example, the name Marie can be a misspelling of Mary while both records belong to the same person. If all other attribute comparisons agree, these records belong highly likely to the same person. For record linkage, not all attributes have to be correctly stored in the databases. In short, the goal is to decide how many and which attribute comparisons need to agree to say that two or more records belong to the same entity.

The purpose of this chapter is to describe a workflow for performing a record linkage or deduplication operation. Section 2-2 gives a schematic overview of the workflow. Each step of the workflow plays a role in the record linkage process. Sections 2-3 till 2-8 detail each step of the workflow.

2-2 The record linkage workflow

A record linkage process consists of some steps that need to be performed in consecutive order. The linkage operation can be represented as a workflow [Christen, 2012*a*]. In Figure 2-1, the steps of the workflow are depicted schematically. The workflow presents a record linkage operation between two datasets. The same workflow can also be used for deduplication of a dataset. In that case, both datasets in the workflow represent the same dataset (the dataset deduplicate).

The workflow starts with one or two datasets. In Section 2-3 is discussed how dataset are constructed and how the data is collected. The collection is often a process on which the analyst has no influence. The first step for a record linkage operation is to prepare the data sources such that the data are in a clean, standardised and comparable format (see Data preparation, Section 2-4). The next step is to pair all records such that all record pair combinations are made. In principle, the attributes of all record pairs are compared on the available attributes. For large data sources, the number of record pairs can be gigantic. The aim of the indexing step is to exclude pairs of records that are not likely to correspond with the same entity (see Indexing, Section 2-5). For example, records need to agree on the name otherwise they are not compared on all attributes. Only pairs left after indexing are compared and used in the following steps.

The indexed record pairs are compared on a sub(set) of attributes (see Comparing record pairs, Section 2-6). Comparing information is based on the type of information. Comparing information can be done strict; the pieces of information are identical or not. However, also partial agreement is used for comparison. The compared record pairs serve as input for the classification process (see Classifying record pairs, Section 2-7). The classification step is used to decide if a pair of records belongs to the same entity or not (link or non-link). Classification classifies record pairs into one of three sets; the set of links, the set of non-links and the set of possible links. Possible links are record pairs for which it is not clear if they belong to the same entity.

The set of possible links needs to be evaluated by a reviewer in the evaluation step (see Evaluation, Section 2-8). A manual review of the comparisons can lead to a classification

into links or non-links. Besides manual reviewing the possible links, there are several post-classification and evaluation methods available.

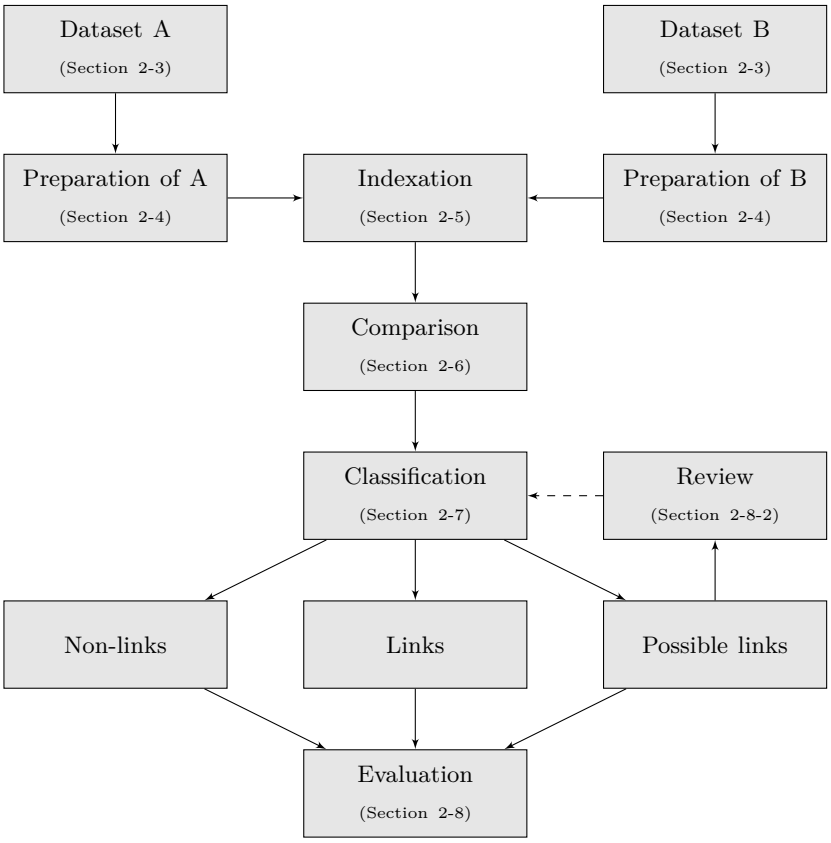


Figure 2-1: The workflow for a record linkage operation. Image based on [Christen, 2012*a*]

2-3 Datasets

The data stored by organisations, like governments and businesses, is stored in many different (digital) formats. There are many names for these structures of organised data, such as *datafiles*, *datasets*, *databases* or *record files*. In this thesis, these data structures are called datasets. In general, the organised structure of a dataset is a list of basic data structures containing information about an entity such as an individual or business. In computer science, these basic data structures are called *records*. Each attribute of the person or company fills a *field* of the record. These fields are containers of predefined characteristics of the entity. In Table 2-1, an example of a dataset with fictitious personal information is given.

In general, a dataset is an (incomplete) representation of a statistical *population*. Each record of the dataset represents a member of the population. Due to incompleteness, it may happen that not every member of the population is represented by (a record in) the dataset. In fact, the dataset represents a sample of the population. This selection of members of the

Record id	Name	Surname	Date of birth	Sex	Place of birth
rec1	Mark	Schwartz	12/19/1952	M	California
rec2	Mary	Smith	03/07/1963	F	Los Angeles
rec3	Micheal	Johnson	06/01/1982	M	NY
rec4	Marie	Smith	03/07/1963	F	LA
rec5	Richard	Levy	01/04/1991	M	New York
rec6	John	Nelson	03/11/1910	M	Texas

Table 2-1: A fictitious example of a structured dataset with personal information. Note that the records with identifier rec2 and rec4 may belong to the same person.

population can be the result of a simple random sampling[3] method or some other sampling method.

In general, each element (an entity) in a (sub)population can be identified by a selection of characteristics. For example, a population of individuals has characteristics such as name, age and sex. These characteristics can be merged into a record, the characteristics are stored in the fields of the record. This process of merging the characteristics into a record is called a *record generating process* [Fellegi and Sunter, 1969]. This process can introduce errors (e.g. typographical, mistakes) and incompleteness to the record. The characteristics are often collected and stored by humans, but also (text-)recognition software is used.

In most publications about record linkage, records are linked between two datasets [Fellegi and Sunter, 1969] [Herzog, Scheuren and Winkler, 2007] [Winkler, 1988]. In this thesis, these two datasets are dataset A and dataset B. Datasets A and B are incomplete representations of the statistical populations $\mathcal{A}$ and $\mathcal{B}$ respectively. The subpopulations, represented in the datasets, are denoted by $\mathcal{A}_s \subseteq \mathcal{A}$ and $\mathcal{B}_s \subseteq \mathcal{B}$ respectively. Mathematically, the samples $\mathcal{A}_s \subseteq \mathcal{A}$ and $\mathcal{B}_s \subseteq \mathcal{B}$ serve as input for the record generating process. The record generating functions

$$\alpha : \mathcal{A} \to A$$
$$\beta : \mathcal{B} \to B$$

maps the subpopulations $\mathcal{A}_s$ and $\mathcal{B}_s$ into the datasets A and B respectively. The elements $a \in A$ and $b \in B$ are records of the dataset. For the special case of deduplication, only one dataset and one population are used (in fact $\mathcal{A} = \mathcal{B}$).

The process of storing the characteristics of an entity into a record (the record generation) is a process that is often not visible for the analyst or *data matcher*. It means that the exact role of the functions α and β on the datasets is not always clear. The record generating process is subject to mistakes, interpretation and lack of knowledge. In the data preparation step of the record linkage workflow in Section 2-4, (typographical) variations in the record generating process are examined in more detail.

[3]The simple random sampling method selects the members of the population with equal probability.

2-4 Data preparation

In practice, many datasets contain noisy, inconsistent and missing data [Christen, 2012*a*]. This raw data needs to be converted into a format that is usable for record linkage. The conversion of data for further analysis is known as *data preparation* or *pre-processing*. Data preparation can be seen as the process of increasing the usability of the data or as enriching the data quality. Data quality is often expressed in terms of completeness, accuracy, consistency, validity and timeliness [Christen, 2012*a*]. The field of data preparation is a widely studied research field (not specific for record linkage) [Pyle, 1999]. In this section, a few of the techniques are discussed. Especially techniques that influence the process of record linkage.

2-4-1 Cleaning and standardisation

Changing and removing unwanted characters and tokens in a dataset is the first step, or one of the first steps, in the data preparation process. This step is often applied to string or numerical information such as names, business names and phone numbers. It involves the removal of special characters like dots, slashes and minus characters. The main reason to perform this step is to make the information more easily comparable. Removal of special tokens in a string reduces the number of typographical variations. Special characters, like dots, slashes and minus characters, can be removed or replaced by spaces. In case of multiple special tokens in a consecutive order, they are replaced by one space. Uppercase characters are often converted into lowercase characters and diacritical marks are replaced by their Latin alphabet equivalent. The reduction of string complexity can make the chance of agreement larger, without doing substantial concessions on the content.

For many typographical errors, abbreviations and variations, it is obvious what was meant. For example, if someone fills in 'Church Str', it was clearly meant to be 'Church Street'. Such typographical errors and variations are easy to correct. There are look-up tables available, or can be created, to correct such typographical errors and variations automatically into standardised and corrected values. In the field of data preparation, this is called *standardisation*. Most of the time, standardisation is a simple rule-based process. A part of standardisation is the labelling of data into a simple format. This is often already done in the record generating process itself. For example, the field with the sex of a person is often inserted as M and F or 1 and 0. This is a type of encoding method, i.e. a standard. It happens that multiple standards are used in a record generating process for a single dataset.

2-4-2 Inconsistent information

Sometimes, values in a dataset are very unlikely, or even impossible. For example, someone aged 180. For such inconsistent data, the data analyst needs to make a decision what to do with the value. A few of the options are; correct the value, leave untouched or change into a missing value.

Sometimes, databases contain information that is clearly incorrect. For example, the date 20/01/1970 stored in a field in format MM/DD/YYYY is clearly a mistake. In this case,

it feels quite safe to correct the date into 01/20/1970. If the number of the month and the number of the day are equal or less to 12, it is not possible to distinguish such errors.

Other inconsistencies are not so clear and trivial. The police road accident data used in Part II contains quite a large number of babies, involved in a road accident on their date of birth. This seems to be odd. It is likely that the police officer reported, by mistake, the current day as the date of birth. Although this data is quite clearly incorrect, it is not easy to correct the value. Declaring such a variable as missing seems to be a valid solution.

2-4-3 Missing data

Records can be incomplete because the information is not available. It means that not all fields of the record contain values. There are a several options to deal with missing values. One option is to remove the entire record if a field value is missing. This method may work well if the amount of missing data is small. Filling missing values by hand is another option. However, this can be labour intensive or impossible. Another option is to fill the missing values with a constant value such as the median. Such techniques, called *imputation* techniques, are widely studied in the field of statistics. There are many advanced imputation techniques available. In the context of record linkage, imputation is not common [Christen, 2012*a*]. The missing values are often left missing. Nevertheless, missing values play an important role in the classification (see The role of missing data, Section 6-5).

2-4-4 Smoothing data

In the theory on record linkage, the term *data smoothing* is sometimes used. In statistics, smoothing of data is known as a process to suppress noise. For example with a moving average. In the context of record linkage, smoothing is not only used for numerical data but also other types of data. This type of data preparation groups data together and replaces it with a value or range. Smoothing information is closely related to some (string)-comparison methods proposed in Section 2-6. Smoothing is especially of interest when the data is noisy or of low quality. Smoothing can be an important data preparation step.

A drawback of smoothing is the loss of distinguishing information. In a census database, the date of birth '01/01/1970' will agree with fewer records than the year of birth 1970. The loss of distinguishing power makes it harder to distinguish the distribution of links from the distribution of non-links. Sometimes, the reduction of distinguishing power helps the classification process. Most persons know their exact date of birth, but it is much harder for other events in the past. Examples are forms with questions about the date of first drugs use or the date of a car crash. For such questions, it is not very likely that a respondent recalls the exact date. It is more likely the respondent remembers that the date was, for instance, in the summer of 2005. Such information is often quite noisy and of poor quality, which makes it useful to smooth it before linking. A solution is to compare only the year of the car crash and not the exact date. This smoothed data makes the chance on an agreeing comparison larger at the expense of losing distinguishing power.

2-5 Indexing

After preparation of dataset A and dataset B, it is natural to compare each record in dataset A with all records in dataset B. A combination of a record of dataset A and a record of dataset B is called a *record pair*. These record pairs are pairs of records given by

$$(a, b) \in A \times B. \tag{2-1}$$

The number of records in datasets A and B are denoted as N_A and N_B respectively. The total number of record pairs depends on the number of records in dataset A and dataset B. It increases (quadratically) as function of the number of records in the datasets, i.e. the number of record pairs is $N_A \cdot N_B$.

Comparing all record pairs (on all attributes) can be computationally prohibitive. For example, two datasets with 10^6 records have 10^{12} record pairs to compare. Comparing all record pairs may lead to long computational times and substantial memory requirements. Several techniques are developed to make a smart selection of record pairs to compare on all attributes. They rely on the fact that many record pairs do not belong to the same entity [Baxter, Christen and Churches, 2003]. This way of selecting record pairs is often called *indexing* or *indexation*.

Indexing is the process of selecting a subset of $A \times B$. This record pairs in this subset are called *candidate record pairs*. In the next steps of the workflow, only the candidate record pairs are evaluated; the other pairs are left out of scope. It implies that indexing has to be done securely. If a record pair belonging to the same entity is not part of the candidate record pairs, it can not be linked anymore. However, if too many records are added to the candidate record pairs, the process becomes again computationally intensive.

The number of record pairs belonging to the same entity (the true links) scale up slower than the number of record pairs that do not belong to the same entity (the non-links). Assume that there are no duplicates in dataset A or dataset B. This means that a dataset does not contain multiple records belonging to the same entity. Assume that record $a \in A$ can be linked to at most one record $b \in B$ and vice versa. With these assumptions, the number of true links scales up linearly while the number of record pairs scales quadratically with respect to the number of records in the datasets.

Indexing is a trade-off between the reduction of record pairs and the fraction of links missed due to indexing. The goal is to reduce the number of record pairs drastically (reduction), but also to keep all linking record pairs in the set of candidate record pairs (completeness). There are several methods to measure the effect and quality of the indexing methods [Elfeky, Verykios and Elmagarmid, 2002]. First of all, the *Reduction Ratio* (RR) is a metric that quantifies the reduction of record pairs. The reduction ratio metric is

$$\mathrm{RR} = 1 - \frac{N_C}{N_A N_B}, \tag{2-2}$$

where N_C is the number of candidate record pairs. The Reduction Ratio does not say anything about the completeness of the indexing method. For the completeness is the *Pairs Completeness* (PC) ratio used. The pairs completeness metric is the ratio of linked record

pairs in the set of candidate record pairs against the total number of links without indexing[4]. The Pairs Completeness ratio is given by

$$\text{PC} = \frac{N_{M,C}}{N_M}, \tag{2-3}$$

where $N_{M,C}$ is the number of links in the set of candidate record pairs and N_M is the total number of links. There are metrics to combine the Reduction Ratio and the Pairs Completeness. The *F-score* is a popular metric for examining the quality of the indexing method. It is the harmonic mean of the Reduction Ratio and the Pairs Completeness, mathematically given by

$$\text{F-score} = \frac{2 \cdot \text{RR} \cdot \text{PC}}{\text{RR} + \text{PC}} \tag{2-4}$$

The F-score is a value between $[0, 1]$. If the F-score is 1, then the indexing method selects all, and only, the record pairs that belong to the same entity.

Over the years, several indexing techniques are developed. Some of the methods are very basic and widely usable while other are (relatively) advanced and specifically developed for the type of data. The most common indexing method is *standard indexing*, which compares only records within mutually non-overlapping blocks of $A \times B$. Standard indexing can be applied to all types of data. This method is discussed in detail in Section 2-5-1. Standard indexing turns out to be a special case of *Sorted Neighbourhood Indexing*. This method is discussed in Section 2-5-2. In Section 2-5-3, the indexing *Q-gram Indexing* method for strings is discussed. This method is useful for names and surnames.

2-5-1 Blocking

Standard indexing, also known as blocking, is a very popular and useful indexing method. With blocking, the record pairs are compared on a (single) field of the record; called a *blocking key*. All record pairs agreeing on the blocking key are assigned to the 'block' of their *blocking key value*. The blocking key value is the value of blocking key of the record pair. Every record pair can only be assigned to one block. Therefore, standard indexing divides record pairs into mutually disjoint blocks of record pairs. The choice of a suitable blocking key is important in the record linkage process. Fields with few errors and missing values are the best option [Christen, 2012*b*].

In Figure 2-2, the blocking method is displayed graphically. Consider records $a_1, \ldots, a_{16} \in A$ and records $b_1, \ldots, b_{12} \in B$ are sorted on the blocking key according to the same sorting criteria. Record pairs agreeing on the blocking key value are coloured grey. Not all records $a_1, \ldots, a_{16} \in A$ and $b_1, \ldots, b_{12} \in B$ are assigned to a block because they do not agree with any of the records on the blocking key. With blocking, the blocking key still needs to be compared for all record pairs. This might look computational intensive. However, most data processing software and computer languages have efficient algorithms for this. Most algorithms sort the data first.

[4] A reduction ratio of 0 is a *Full index*

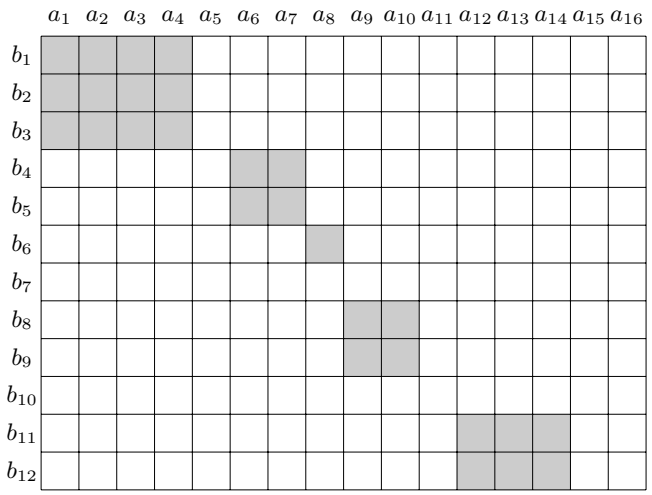


Figure 2-2: Standard indexing (or blocking) on sorted data

In general, blocking is an effective method to reduce the number of record pairs that need a full comparison. The number of candidate record pairs depends on the frequency distribution of the blocking key values and the number of records in both datasets. A quick estimate of the reduction of record pairs can be made; assuming that the frequency distribution of the blocking key values is uniform. Let k be the number of blocking key values both datasets have in common. The number of record pairs in a block is estimated as $\frac{N_A}{k}\frac{N_B}{k}$. The estimated Reduction Ratio (RR) is

$$\text{RR} = 1 - \frac{k\frac{N_A}{k}\frac{N_B}{k}}{N_A N_B} = 1 - \frac{1}{k}. \tag{2-5}$$

The number of candidate record pairs can decrease even further when using multiple blocking keys. Each 'block' contains record pairs agreeing on all blocking keys. It can increase the number of blocks and reduces the size of the blocks.

2-5-2 (Adaptive) Sorted Neighbourhood Indexing

The standard indexing method can only put identical blocking key values into a block. Especially for string values, this is a drawback. Typographical variations (called 'neighbours') in the blocking key values lead to record pairs which are not indexed that they belong to the same entity. The Sorted Neighbourhood Indexing deals with these typographical variations or 'neighbours' [Hernández and Stolfo, 1995].

In the Sorted Neighbourhood method proposed by Hernández and Stolfo [1995], the datasets A and B are concatenated. One of the attributes/fields is the *sorting key*, which is similar to the blocking key. For the sorting key are the unique values of the concatenated data extracted and sorted. Now, slide a fixed window of size $w \in \mathbb{N}^+$ over the unique sorting key values of the concatenated datasets. In contrast to standard indexing, this indexing technique ($w > 1$) creates candidate record pairs with identical or similar key values. Figure 2-3 shows how this indexing technique allows similar key values. The figure is a fictitious example with 12 sorting key values, $s_1, \ldots, s_{12}$. Adjusting the size of the window can be used to change the

allowed similarity. The standard indexing technique can be seen as a special case of the sorted neighbourhood indexing method. The blocking keys values are used as sorting key values. A window of size $w = 1$ results in the same candidate record pairs as with standard indexing or blocking.

Sorted Neighbourhood indexing is applied successfully on string and numerical data. For string data, like names and last names, it is a good method to take into account typographical variations. For numerical data, is can be applied to include some tolerance in the numerical value. For numerical data, the window is sometimes not of fixed size but dynamic. This indexing method is adaptive sorted neighbourhood indexing [Christen, 2012*a*].

Typographical variations at the beginning of a string field make the Sorted Neighbourhood indexing vulnerable to indexing errors, because the key values are often sorted with a sorting criteria that starts at the beginning of the word. A misspelling at the beginning of the string has a larger influence on the indexing than a misspelling at the end. Advanced sorting criteria are developed to overcome this problem. To overcome this issue, one can consider encoding of the string with an encoding tool that is not very sensitive to mistakes at the beginning of the string. Then, the encoded values are used as sorting key values. See Section 2-6 for information about encoding string information.

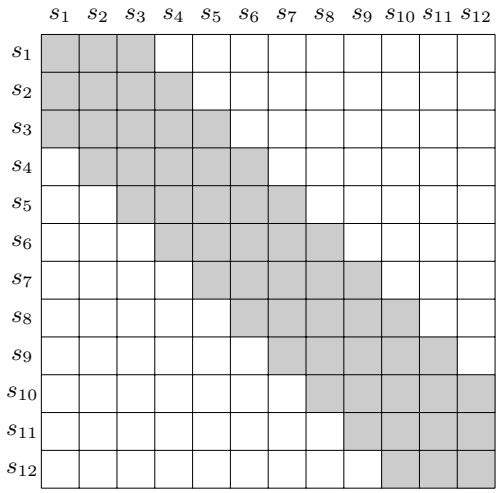


Figure 2-3: Sorted Neighbourhood Indexing with $w = 3$. The sorting key values are $s_1, \ldots, s_{12}$.

The Reduction Ratio (RR) for the Sorted Neighbourhood indexing method highly depends on the size of the window. A quick estimate of the Reduction Ratio can be performed under the assumption that the datasets are of identical length, and the method is applied to the record values instead of sorting keys. Straightforward calculations show that the reduction ratio is

$$\mathrm{RR} = 1 - \frac{w^2 + (N_A - w)(2w - 1)}{N_A \cdot N_A}. \tag{2-6}$$

2-5-3 Q-gram indexing

The Sorted Neighbourhood indexing method is a simple method to deal with typographical errors for string comparison fields. A drawback of this method is that it is vulnerable for errors at the beginning of the string. The q-gram indexing method is a method that can deal with data containing many errors and typographical abbreviations. The location of a mistake in the string has no, or less, influence on the result. In general, the q-gram indexing method is used for string information.

With q-gram indexing, a string is split into k substrings of length q [Baxter, Christen and Churches, 2003]. The substring are called q-grams. A common choice for the length of a q-gram is $q = 2$ or $q = 3$. For example, the author's name is split in to 'jo','on', 'na', 'at', 'th', 'ha', 'an' when using $q = 2$. With q-gram indexing, the q-grams of two strings are compared. If many q-grams occur in both strings, then the strings are (relatively) similar. To control the amount of q-grams needed to add the record pair to the candidate record pairs, a threshold $t \in [0, 1]$ is used. This threshold is used to set a minimum number of q-grams that need to agree. This minimum number of q-grams is given by

$$\max(1, \lfloor kt \rfloor).$$

For the name 'jonathan' and $q = 2$, the number of q-grams is 7. Set the threshold to $t = \frac{2}{3}$. This threshold implies that the number of q-grams to agree is at least $\lfloor \frac{14}{3} \rfloor = 4$. This means that the combinations are [jo, on, na, at, th, ha, an], [on, na, at, th, ha, an], [jo, na, at, th, ha, an], [jo, on, na, th, ha, an], [jo, on, na, at, ha, an], [jo, on, na, at, th, ha], [na, at, th, ha, an], ..., [at, th, ha, an]. Consider now the misspelled variant 'jonatan'. The number of q-grams is now 6. The maximum number of q-grams to compare on is $\lfloor \frac{12}{3} \rfloor = 4$. Again, the combinations of q-grams are created. There are several combinations with 4 and one with 5 q-grams found in both cases. For example, [jo, on, na, at, an] occurs in both situations. This record pair is a candidate record pair. This q-gram indexing method also works in case of letter interchanges. Consider the name variant 'joanthan', again the minimum number of q-grams needed is 4. The combination [jo, th, ha, an] is found in both cases.

The q-gram method is a good method for indexing of datasets with large number of typographical variations. In contrast to the Sorted Neighbourhood indexing method, the method is not more vulnerable to errors in the beginning of the string. A drawback of this method is the computational complexity. For each string, a q-gram is made and the lists of combinations is made. For 'jonathan' are $\sum_{i=4}^{7} \binom{7}{i} = 35+21+7+1 = 64$ possible q-gram combinations. This makes the number of comparisons much larger. Take into account that all of these values are compared with the Q-gram combinations of the other string in case of (clear) disagreement. For example, if the other string has also 64 combinations, the number of comparisons can be $64 * 64 = 4096$ comparisons. There are several computational improvements to improve this naive approach.

2-5-4 Disjunctions, conjunctions and index passes

Indexing is a crucial step in the record linkage process. A bad quality blocking or sorting key leads to record pairs that can no longer be linked, while other indexing methods or blocking and sorting keys result in (too) many candidate record pairs. There are several techniques

to deal with these problems without choosing another indexing method. These methods are based on combinations of indexing techniques.

If there are too many candidate record pairs, a conjunction can be taken of multiple comparison variables. For example, blocking can be used on the zip code and the surname in one step. Only if they both agree, the record pair is a candidate record pair. The drawback of this approach is that the Pairs Completeness can become smaller. In some practical applications, it is not useful to use a conjunction of keys. The reason is that the quality of the key is too low. To overcome this problem, repeat the entire linking process with different (combinations of) keys. Each linking operation (an indexing pass) uses another index key. The disjunction of the links is used as record linkage result. This approach is known as using *index passes* [Winkler, Yancey and Porter, 2010]. Another approach, with similarities to the indexing passes, is found by taking the disjunction of several blocking keys. A pair of records is a candidate record pairs if at least one of the indexing methods applied declares the record pair as a candidate record pair.

2-6 Comparing record pairs

In Table 2-1 was seen that *rec2* and *rec4* may belong to the same entity. Such a manual record linkage is based on (quickly) comparing the attributes of the entity found in both records. This comparing of attributes is exactly what is done in the comparison step of the record linkage workflow. A (selection) of attributes found in both records is compared. It turns out that comparing information is not always trivial. In this section is discussed how (different types of) information can be compared.

The comparison of record pairs is performed with a *comparison function* or *similarity function*. The comparison function or similarity function

$$s : A \times B \to \Gamma \tag{2-7}$$

maps the record pairs in $A \times B$ into the *comparison space* Γ. The comparison function s is a vector of functions

$$s = (s_1, \ldots, s_K)$$

where K is the number of comparisons between both records. Each element $\boldsymbol{y} \in \Gamma$ is a vector

$$\boldsymbol{y} = (y^1, \ldots, y^K)$$

of length K. This vector is called the *comparison vector* or *similarity vector*.

In the literature on record linkage, comparison vectors are often distinguished from similarity vectors. A similarity vector is a vector $\boldsymbol{y} \in [0, 1]^K$. Each element represents the 'similarity' between two attributes. If two attributes are identical, the similarity is 1. The more the attributes disagree, the closer the similarity goes to 0. Comparison vectors are vectors with information about the kind of comparison. These vectors are not necessarily values between 0 and 1 or even numerical values. The values in the comparison vector indicate what type of

comparison occurred. For example, the comparison vector can contain the cases 'identical' and 'not identical', but also 'identical sex and the sex is male'. These comparison vectors are often used in probabilistic record linkage while similarity vectors are often used in deterministic classification (see Classifying record pairs, Section 2-7).

The next step is to take a closer look at the comparison or similarity functions itself. Consider a record pair $(a, b) \in A \times B$, compared on attribute/field $i \in \{1, \ldots, K\}$. The most basic form of comparison considers only agreement and disagreement. Comparisons of this kind are mathematically denoted as

$$s_i(a^i, b^i) = \begin{cases} 1 & \text{if } a^i = b^i \\ 0 & \text{else,} \end{cases} \tag{2-8}$$

where 1 indicates agreement and 0 for disagreement. This type of comparing is sometimes called the *exact comparing*. It is a mapping $s_i : a_i \times b_i \rightarrow \{0, 1\}$.

Sometimes, exact comparing is not sufficient. For example, when comparison names and surnames with misspellings. (Phonetic) encoding can be used before comparing the names or surnames. Consider the values a^i and b^i, which are encode before being compared. The encoding function ϕ is a function that maps a value into a code or another value. These encoding functions can be used in the exact comparison function (2-8), i.e.

$$s(a^i, b^i) = \begin{cases} 1 & \text{if } \phi(a^i) = \phi(b^i) \\ 0 & \text{else.} \end{cases} \tag{2-9}$$

A well-known phonetic encoding function for names and surnames is Soundex encoding [US National archieves, 2007]. Returning to the example of Marie and Mary, both Marie and Mary have Soundex code $M600$. Formula 2-9 will return 1, i.e. agreement. There are also other encoding functions like Soundex. Most of them are language specific. Soundex works especially well for names in English-speaking countries.

2-6-1 Comparing string information

In the previous section was described how (encoded) string information can be compared with the exact comparison function. Besides encoding string information, there is another method to compare string information. The similarity between two strings can be computed. This is often a values between 0 and 1. Some of the similarity functions are specifically developed for names and surnames while others are useful for all types of string information. A few widely used string similarity metrics are shortly described below;

Truncate strings This method truncates both strings and exactly compares the truncated substrings. This method is simple and can be effective. There are several ways to truncate a string. For example, by truncating the begin or the end of the string. For names and surnames, it is useful to truncate the end of the name because fewer errors occur at the beginning of a name [Christen, 2012*a*].

Longest common substring The Longest Common Substring metric is an iterative method that searches for the longest substring both string have in common in each iteration

[Christen, 2012*a*]. This substring is removed from the both strings. For the remaining string, this step is repeated. This step is continued until there is no substring in common left or the length of the substring in common drops below a threshold; the algorithm terminates. The similarity is the length of all substrings together, divided by the length of the shortest string of both input strings.

Levenshtein distance This metric defines a distance based on the minimum number of changes needed to correct the string [Levenshtein, 1966]. The function counts the number of substitutions, insertions or deletions. There are variants of this metric were some substitutions have different importance.

***Q*-gram** The q-gram metric splits the strings into substrings of length q (Similar to q-gram indexing described in Section 2-5-3). For example, Mary is split in (Ma, ar, ry) when using $q = 2$. Both strings are split into substrings. The number of q-grams in common is divided by the number of q-grams of the shortest input string [Augsten and Bohlen, 2013]. There are several variants of this metric.

Jaro-Winkler This metric is specifically developed for names and surnames. The method combines techniques found in the q-gram metric and the Levenshtein metric [Winkler, 1999]. The metric uses that most errors in names occur at the end of the name.

Table 2-2 shows the similarity for the mentioned methods for the names 'Marie' and 'Mary'. It is clear that there are substantial differences between the metrics for the names 'Marie' and 'Mary'. The choice of a suitable metric is based on knowledge about the data. For names, the Jaro-Winkler metric can be effective while it is better to truncate the strings or use q-grams for other string information.

Metric	Similarity
Truncate string 1 : 3	1
Truncate string 1 : 4	0
Longest common substring	0.75
Levenstein	0.6
Q-gram ($q = 2$)	0.66
Q-gram ($q = 3$)	0.33
Jaro-Winkler	0.75

Table 2-2: Table with similarity values for names Marie and Mary

2-6-2 Comparing numerical information

Many datasets contain numerical information like salaries, length of persons and sports results. Exact comparison may not always satisfy for numerical comparison. For example, the registered length of a person may vary for both records as a result of rounding or increase in length. An exact comparison will fail in this example. There are (infinitely) many similarity metrics possible to compare numerical information. Often, a region of tolerance is used when comparing numerical information. In Section 2-6-4, it turns out that numerical comparison methods are also useful for time and date comparison in the context of record linkage.

A basic approach to compare numerical information is to define a maximum absolute difference $d_{\max}$. If the absolute difference between two numerical values exceeds $d_{\max}$, the comparison disagrees. Consider two pieces of numerical information a_i and b_i. The comparison is given by

$$s(a_i, b_i) = \begin{cases} 1 & \text{if } |a_i - b_i| \leq d_{\max} \\ 0 & \text{otherwise.} \end{cases} \tag{2-10}$$

It is possible to add partial agreement to the numerical comparison. If the values are exactly the same, the similarity is 1. If $|a_i - b_i| \leq d_{\max}$, then it is a value between 0 and 1. The similarity can be linear, then it is given by [Christen, 2012*a*]

$$s(a, b) = \begin{cases} 1 - \left(\frac{|a-b|}{d_{\max}}\right) & \text{if } |a - b| < d_{\max} \\ 0 & \text{otherwise.} \end{cases} \tag{2-11}$$

An example of similarity based on a quadratic principle is [Reurings and Bos, 2009]

$$s(a, b) = \begin{cases} 1 - \left(\frac{(a-b)^2}{d_{\max}^2}\right) & \text{if } |a - b| < d_{\max} \\ 0 & \text{else.} \end{cases} \tag{2-12}$$

2-6-3 Comparing categorical information

Categorical information is information for which the number of possible values is finite. Most of the time, categorical information is stored in datasets as string or numeric information, or as a combination of both. Comparing categorical information is similar with the exact string/numerical comparing. If both fields are identical, then the comparison/similarity is 1 (agreement); otherwise the comparison/similarity is 0 (disagreement). An example of such a variable is the zip code or sex.

A partial agreement can be used for categorical information. Several categories are grouped together. If the group of categories is identical for both records, then it is full or partial agreement. This approach has similarities with the method of encoding strings, seen in Formula 2-9. There is a second approach in which closely related categories get a partial agreement. This method can be applied for zip codes. In this approach, the zip code geographically located next to the zip code of the other record partially agree.

2-6-4 Comparing date and time information

Date information is commonly found in datasets. For example, the date of birth of a person is stored. The storage of date information occurs in many different formats. The date can be stored in a (string)format like MM/DD/YYYY or DD/MM/YYYY, where DD is the day number, MM the month number and YYYY the year. It is also possible to store the date, the time, or both, as a single number. Such a value is known as a timestamp. A timestamp is a number that indicates the time or days elapsed since a given timepoint in history. The most common timestamp is the Unix time, which is the number of seconds elapsed since $00:00:00$

Coordinated Universal Time (UTC), Thursday, 1 January 1970. Timestamps are commonly found in digital datasets.

There are two types of errors with date information that frequently occur [Christen, 2012*a*]. The first type of error occurs when the day and month numbers are swapped. The format (MM/DD/YYYY) was used while (DD/MM/YYYY) was requested or vice versa. This type of error was earlier discussed in the section on data preparation (Section 2-4). Another error is an incorrectness of the month number. The day and year are correct, but the number of the month is incorrect. For example, the month 9 is registered as month 10.

The two methods to store dates and times, as a string or as a number, have both advantages and disadvantages when comparing. For the dates and times stored as string, it is possible to identify a swapping of day and month number. These comparisons can get a similarity between 0 and 1 or a special type of comparison value. For the date and time stored as a number, this is much harder. The latter storage method has as an advantage that date and time can be compared with numerical comparison methods. These methods make it possible to add tolerance in the comparison (use $d_{\max}$, see Section 2-6-2).

2-6-5 Comparing geographical information

Geographical information is often stored in one of two formats; as address or as geographic coordinates. Both types are compared differently. Address information can be compared as string information. Comparing the address can be done by comparing city names, street names, zip codes and house numbers individually. Address information is, just like string information, susceptible to typographical errors. Geographic coordinates can be compared as numerical information. There are several ways to compare geographic coordinates. The geographic distance can be computed over the surface of the earth (geodesic distance) whereafter it can be compared as numerical information. Another option is to calculate the distance over the road.

2-7 Classifying record pairs

Record linkage is the process of bringing together record pairs that belong to the same entity. If a record pair belongs to the same entity, then the record pair is called a true link. If they do not belong to the same entity, then it is a true non-link. The goal of this step in the workflow is to decide which record pairs belong to the same entity and which not. The record linkage problem is, in fact, a classification problem in which record pairs are classified as links or non-links. These two sets are often called the *positive link* set and *positive non-link* set. The classification of record pairs relies on the comparison or similarity vector.

Classification can be done in a supervised way or an unsupervised way. When the classification is supervised, there is information about the true link status of the record pairs available. The information can be used to train the classification method; a process known as *supervised learning*. After supervised learning, classification can be performed on data with unknown true link status. If there is a lack of training data, the classification needs to be done unsupervised. In this thesis are several unsupervised learning algorithms discussed for

the probabilistic record linkage classification framework. Simulation of data with known true link status can help the data analyst to optimise the classification method.

Dividing the set of (candidate) record pairs into a set of links and non-links can be hard. Some comparison vectors belong highly likely to the set of links while others belong clearly to the set of non-links. There can be a group of comparison vectors that are not easy to classify into one of these groups. They can be classified as *possible links*. For these links, human review is needed to classify them. Smith and Newcombe [1979] show that manual classification is hard. It is heavily subject to the reviewer. They concluded that manual classification is not that good as probabilistic classification.

In the remaining of this section, three non-manual classification methods are discussed shortly. The classification methods are; deterministic classification, probabilistic classification and rule-based classification.

2-7-1 Deterministic classification

Record linkage based on deterministic principles is one of the simplest methods to classify record pairs into a set of positive links and positive non-links. With deterministic classification, the comparison variables are compared and the sum of the similarities is taken. This sum of similarities is called the *weight* of the record pair. A threshold divides the records, based on their weight, into the set of positive links and set of positive non-links. It is also possible to set two threshold values. If the sum of the similarities of a record pair is above the largest threshold, then the record pair is classified as a positive link. If the weight is less than the smallest threshold, then the pair of records is assigned as a positive non-link. The remaining record pairs are classified as possible links. Due to the use of thresholds, deterministic classification is sometimes called *threshold-based classification* [Christen, 2012*a*].

Mathematically, the weight of a record pair is given by

$$\sum_{i=1}^{K} y^i = \mathbf{w}^T \boldsymbol{y} \tag{2-13}$$

where $\mathbf{w} = [1 \dots 1]^T$ is a vector of ones of length K and $y^i \in [0, 1]$. To classify the record pairs into a set of links and non-links, the threshold(s) need(s) to be set. If the sum of similarity values exceeds this threshold, the record pair is classified as a positive link. Setting a threshold level can be made manually. If there is data available with knowledge about the 'true' record linkage, then this can be used to find a sufficient threshold level by minimising the number of errors on the training data to an acceptable number.

Deterministic classification works well when there is one unique identifier to link on. The method works also well when there are multiple attribute values of high quality and equal distinguishing power. The comparison values need to have equal distinguishing power because non-important comparison variables can overwhelm important variables. For example, consider the comparison variables of an individual are; a unique personal identifier, the sex, the hair colour and the eye colour. Consider the following fictional pair of records

Record id	Identifier	Sex	Hair colour	Eye colour
rec1	23795345	Female	Blond	Blue
rec2	34680346	Female	Blond	Blue

The identifying number is not identical for this record pair. The three other attributes are identical. Assume, the similarity value is 1 in case of agreement and 0 in case of disagreement. This pair of records gets weight 3. Now, consider the following example,

Record id	Identifier	Sex	Hair colour	Eye colour
rec3	90853445	Male	Blond	Blue
rec4	90853445	Male	Brown	Blue

In this case, the record identifiers agree. The hair colour is different for this pair of records. This weight of this pair of records is 3. This weight is the same as the previous pair of records. In the second example, it is very likely that the records belong to the same entity. For the first example, it is not probable that they belong to the same entity. Nevertheless, both records have the same summed similarity values. This example shows that some variables may need to be considered more important than others. The weight vector $\mathbf{w}$ in Formula 2-13 can be adjusted to solve this problem. Important attributes of the entity (attributes with high distinguishing power) get a large weight and less important fields get low weights. The choice of the weights can be set manually or by minimising the number of wrongly classified record pairs with training data.

2-7-2 Probabilistic classification

Probabilistic record linkage is nearly always seen as the record linkage method developed by Fellegi and Sunter [1969]. The framework is based on the work of Newcombe et al. [1959] and Newcombe and Kennedy [1962]. They argue that linking a pair of records without a unique identifier should be based on attributes both records have in common. Following their approach, the weight[5] of an attribute is based on the occurrence of errors, missing values and the distribution of attribute. In Chapter 3, the Fellegi and Sunter model is discussed in detail. Fellegi and Sunter prove that there is a, what they call, linkage rule that is the statistical best and optimal mapping to classify the record pairs.

The probabilistic model is well-known due to its ability to make use of the distinguishing power of identifying information [Yancey, 2002]. Two records with the same common surname should have a lower weight than the same records with an rare surname. The likelihood that two records share a rare surname is lower than the likelihood that two records share a common surname. In Chapter 5 is the distinguishing power of the Fellegi and Sunter framework discussed and several methods discussed to estimate the parameters.

[5]The weight is similar with the weight in the deterministic classification framework

2-7-3 Rule-Based classification

Rule-based classification makes use of a set of predefined rules of logical connectives to classify the record pairs into links, possible links and non-links. The logical connectives connect comparison like $y^i = 1$, $y^i = 0$ and similarities like $y^i \geq 0.7$. The rules are combinations of conditions with conjunctions and disjunctions. A rule has the form

$$(\text{condition}_1 \vee \text{condition}_2 \vee \dots) \wedge \dots \wedge (\text{condition}_{n-1} \vee \text{condition}_n \vee \dots) \tag{2-14}$$

where 'condition$_1$" is a condition of the type mentioned above. Most of the time, the disjunctions are not needed. For example, consider a vector with comparisons on the name (s_{name}), the surname (s_{surname}), the age (y_{age}), the sex (y_{sex}) and the place of birth (y_{pob}). A rule for this comparison vector is

$$(s_{\text{name}} \geq 0.7) \wedge (s_{\text{surname}} \geq 0.85) \wedge (y_{\text{age}} = 1) \wedge (y_{\text{sex}} = 1) \wedge (y_{\text{pob}} = 1). \tag{2-15}$$

The analyst has to decide what to do if this rule is true. It can be classified as a positive link, as a positive non-link or as a possible link.

2-8 Evaluation

The classification results in a set of positive links, positive non-links and possibly a set of possible links. Interpretation and evaluating the result follows the classification. This section describes several quality measures and post-classification steps.

One of these post-classification steps is the manual classification of possible links into the sets of links or non-links. This process is known as the *clerical review*. Reviewed record pairs can flow back as training data to improve the classification. In Section 2-8-2, this is discussed in detail.

The record pairs classified as links can contain multiple records of dataset A which link with a single record $b \in B$, or vice-versa. It may be desired that one record $a \in A$ links with one record in dataset B. This type of one-to-one linking can be applied as a restriction to the link result. Such restrictions to the links are discussed in detail in Section 2-8-3.

In Section 2-8-4, a method called Capture-Recapture is described to estimate the number of entities not contained in one of the datasets. Those missing records are entities of the population not represented by one of the datasets.

2-8-1 Quality measures

In the context of record linkage, quality measures are used to measure the quality of the record linkage. The measures are used to measure quality conditions like the accuracy and precision of the classification. These measures can only be used if the true link status is known. To perform quality measures, divide the set of record pairs into the following four sets [Christen and Goiser, 2007]:

True positives (TP) True links, classified as links

False positives (FP) True non-links, classified as links

True negatives (TN) True non-links, classified as non-links

False negatives (FN) True links, classified as non-links.

Denote the number of record pairs for the four sets as N_{TP}, N_{FP}, N_{TN} and N_{FN}. Altogether, N_{TP}, N_{FP}, N_{TN} and N_{FN} sum up to the number of (candidate) record pairs.

One of the widely used quality measures used in record linkage measures the accuracy of the classification. This measure is defined as

$$\text{accuracy} := \frac{N_{\text{TP}} + N_{\text{TN}}}{N_{\text{TP}} + N_{\text{FP}} + N_{\text{TN}} + N_{\text{FN}}}. \tag{2-16}$$

The closer this measure goes to 1, the better the classification method. There is a serious drawback on this measure. In general, the number of true negative N_{TN} is much larger than the number of true positives N_{TP}. Therefore, the true negatives dominate the accuracy measure. If the number of true negatives N_{TN} and the number of true positives N_{TP} are in balance, which is uncommon in record linkage. Both measures are combined in the F-score.

Another measure, the precision measure, is not dominated by the true negatives N_{TN}. This measure considers the fraction of the number of positive links against the number of true links. The precision is given by

$$\text{precision} := \frac{N_{\text{TP}}}{N_{\text{TP}} + N_{\text{FP}}}. \tag{2-17}$$

The recall, or true positives, measure is used to get the fraction of true links against the set of classified links. The recall measure is given by

$$\text{recall} := \frac{N_{\text{TP}}}{N_{\text{TP}} + N_{\text{FN}}}. \tag{2-18}$$

These fractions, the precision and recall, are used to calculate an F-score. The score is the harmonic mean of both measures and given by

$$\text{F-score} := \frac{2 \cdot \text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}} \tag{2-19}$$

It is similar to the F-score seen in the section on indexing (see Section 2-5). An F-score close to 1 indicates a good quality classification.

2-8-2 Clerical review

If the classification method can not assign a record pair to the set of links or non-links, then it can be classified as a possible link (see Section 2-7). For the class of possible links, it is not clear to which set they belong. The analyst has to decide what to do with those record pairs. Manual classification is one of the options. This type of classification is known as a *clerical review*.

There are several difficulties with clerical review as classification method. In general, clerical review is a labor- and time-intensive process. The main reason is that it has to be done by humans, and every pair of records needs to be reviewed. Manual review can result in the review of thousands of possible links when the datasets are large, the data is of poor quality or the distinguishing power is low. Besides the labor intensiveness of the process, it can be hard to make a classification into links or non-links for the reviewer. A clerical review is based on the perception of the reviewer, this implies that multiple reviewers can classify a record pair differently. Christen [2012*a*] states that the time of the day, the mood and the concentration level can influence the clerical review process.

A clerical review can be used to make the classification method better. The manually classified record pairs are used to train the classifier. The can be used for deterministic, probabilistic and rule-based classification. Using clerical reviewed data to train the classifier needs to be considered carefully, because of the thoroughness of the classification and the differences between reviewers [Christen, 2012*a*]. Reviewed data can make the classifcation process worse.

2-8-3 One-to-one linking

After classification (and clerical review), there is a set of positive links and positive non-links. A record $a \in A$ can be found in multiple linked record pairs. The same can hold for records in B. Sometimes, this agrees with the model needed, but sometimes not. In many situations, it is preferred that one record in A links at most one record in B.

There are three possible scenarios for linking the record pairs inside the linked set [Christen, 2012*a*]. The first one is *one-to-one* linking. In this scenario, one record in A is linked with at most one record in B. Another scenario is *one-to-many* linking. In this scenario, it is possible that one record of A is linked to multiple records in B, but one record out of B can be linked to at most one record in A. The last scenario is *many-to-many* linking. In that case, one record in A can be linked to multiple records in B and one record in B can be linked to multiple records in A. A many-to-many linking is the direct result of most classification methods. The *one-to-one* linking restriction is made to many record linkage applications. It is widely used in the linkage of census data. If there are multiple records for an individual and one-to-one linking seem to be the correct model, then first deduplicate the datasets.

One-to-one linking between two datasets is closely related to a problem in the mathematical field of combinatorial optimisation. The problem is one of finding a maximum weight linking in a weighted bipartite graph [Schrijver, 2003]. It is called the (linear) *assignment problem*. There are several methods to solve the assignment problem.

Jaro [1989] rewrote the record linkage problem as a linear assignment problem. It is necessary to use some additional notation to formulate the problem. Define the weight for record pair $(a, b) \in A \times B$ as w_{ab}. The link status of a record pair is given by

$$x_{ab} = \begin{cases} 1 & \text{if record } a \text{ and record } b \text{ belong to the same entity} \\ 0 & \text{else.} \end{cases}$$

The assignment problem can be written as a maximisation problem. Mathematically, this problem is formulated as

$$\begin{aligned} &\max_{x} \sum_{a \in A} \sum_{b \in B} w_{ab} x_{ab} \\ &\text{subject to} \\ &\sum_{a \in A} x_{ab} \leq 1, \text{ for } b \in B \\ &\sum_{b \in B} x_{ab} \leq 1, \text{ for } a \in A \\ &x_{ab} \in \{0, 1\} \text{ for all } a \in A, \text{ for all } b \in B. \end{aligned} \tag{2-20}$$

There are two main algorithms to solve the assignment problem, the Greedy algorithm and the Hungarian Algorithm. The Greedy algorithm optimizes the problem locally, and the Hungarian method maximizes the problem globally. The Greedy algorithm starts with the record pair with the largest weight and assigns it as a link. The record pair with the second highest weight is assigned to the linked set if none of the records links with previously linked pairs. The procedure is repeated until no record pairs are left in the linked set. Due to this top-bottom behaviour, the Greedy algorithm does not always find the maximum possible total weight. The Hungarian method is a method that finds the maximum possible weight. It evaluates the linking problem in such a way that the algorithm finds the global maximum. The computation complexity of the Hungarian algorithm is much larger than for the Greedy algorithm.

2-8-4 Capture-Recapture

Capture-Recapture methods were developed to estimate the size of a closed animal population [Herzog, Scheuren and Winkler, 2007]. An example of such a closed animal population is the number of fish in a lake. If a biologist *Captures* a large number of animals of the closed animal population, the biologist knows that the population is at least as large as the number of captures. In the Capture-Recapture method, the animals are marked and released. At a later time, the biologist captures a large amount of the animal population again. The biologist counts the number of marked animals and the number of unmarked animals. This catch is the *Recapture* step. The number of animals captured in both captures can be used to estimate the entire population.

Capture-Recapture is used in the field of record linkage to estimate the size of a population. Consider that dataset A and dataset B represent (a part of) the same closed population. Each dataset can be seen as a 'capture' of the population. The records linked between A and B can be seen as the animals marked in the first capture and recaptured in the second capture.

The are several Capture-Recapture methods to estimate the size of the population. A simple Capture-Recapture estimator is the Lincoln-Petersen estimator [Southwood and Henderson, 2009]. This estimator assumes that the datasets are independently observed from each other.

The estimate of the population size is

$$\hat{N} = \frac{N_A \cdot N_B}{N_M}, \tag{2-21}$$

where N_A and N_B are the number of records in dataset A and dataset B respectively, and N_M is the number of links.

Chapter 3

The Fellegi and Sunter framework

3-1 Introduction

In 1969, the statisticians Ivan P. Fellegi and Alan B. Sunter published "A Theory For Record Linkage". It provides a probabilistic framework for finding pairs of records in two files that represent the same entity [Fellegi and Sunter, 1969]. With their framework, Fellegi and Sunter formalise the pioneering work on probabilistic record linkage of Newcombe et al. [1959]. Howard B. Newcombe identified many of the probabilistic concepts that play a role in record linkage. These concepts are used ten years later in the framework of Fellegi and Sunter. The publication of Fellegi and Sunter [1969] is still of great importance in the field of record linkage.

Fellegi and Sunter developed a framework for the classification of a set of record pairs into links and non-links based on a randomised decision rule. In Section 3-2, a mathematically detailed and improved description of the framework is given. It turns out that the theory of Fellegi and Sunter is closely related to the Neyman-Pearson lemma. The framework of Fellegi and Sunter and the Neyman-Pearson lemma both use the likelihood ratio to define a most discriminatory statistical test [Thibaudeau, 1992]. Section 3-3 explains how the theory of Fellegi and Sunter relates to the lemma of Neyman and Pearson. In Section 3-4, several assumption and simplifications that can be made to the framework are discussed. These assumptions and simplifications can make it easier to apply the model.

3-2 The Fellegi and Sunter framework

Consider two statistical populations $\mathcal{A}$ and $\mathcal{B}$[1]. Fellegi and Sunter [1969] assume that from both populations a simple random sample is drawn. Define the random samples for both

[1] Including the special case for which $\mathcal{A} = \mathcal{B}$

populations as $\mathcal{A}_s \subseteq \mathcal{A}$ and $\mathcal{B}_s \subseteq \mathcal{B}$. Each element in a population has a number of characteristics. For example, a population of individuals has characteristics such as name, age and gender. These characteristics are merged into a *record*. The records corresponding with the random samples $\mathcal{A}_s \subseteq \mathcal{A}$ and $\mathcal{B}_s \subseteq \mathcal{B}$ are collected in file A and B respectively. The process of merging the characteristics into a record is called the *record generating process*. This process can introduce errors (e.g. typographical, mistakes) and incompleteness to the record. The record generating process is based on the record generating functions

$$\alpha : \mathcal{A} \to A$$
$$\beta : \mathcal{B} \to B.$$

These functions map the populations $\mathcal{A}$ and $\mathcal{B}$ into record files A and B respectively.

A record $a \in A$ can represent the same entity as a record $b \in B$. The idea is to pair a record $a \in A$ with a record $b \in B$ and decide if they belong to the same entity. Therefore, the set of record pairs $A \times B$ is created for which each pair contains one record out of A and one record out of B. Mathematically; a record pair is given by

$$(a, b) \in A \times B. \tag{3-1}$$

The set of record pairs $A \times B$ is divided into two distinct sets of record pairs, a subset of $A \times B$ with pairs representing the same entity and a subset of $A \times B$ for which the corresponding entity is not equal. The first subset is called the *linked set* and the second subset is called the *non-linked set*. Each pair of records has a *true link status* M. The true link status is assumed to be random variable[2]. Each pair of records in the linked set has true link status $M = 1$ and each pair of records in the non-linked set has true link status $M = 0$. As the true link status M is unknown, the goal is to find or estimate the true link status M for each pair of records.

All record pairs in $A \times B$ are compared on a selection of attribute/fields that both files have in common. See Section 2-6, for more information about comparing records. Let $K \in \mathbb{Z}^+$ be the number of fields used for comparison. A record pair is compared with a *comparison function*. The comparison function

$$s : A \times B \to \Gamma$$

maps the record pairs in $A \times B$ into the *comparison space* Γ. Each element $\boldsymbol{y} \in \Gamma$ is a K-vector

$$\boldsymbol{y} = (y^1, \ldots, y^K).$$

and is called the *comparison vector* or the *agreement pattern*. This vector is of length K and each element indicates agreement if the field comparison is identical and disagreement if not. But one can also think about more complex comparison patterns. For example, the comparison agrees and has a particular value. At this point, it is not needed to specify the exact method of comparing.

[2]The true link status of a variable is a random variable because the true link status is subject to the random process of sampling record pairs from the population(s). In some situations, the true links status is seen as a parameter of the record pair and not as a random variable. For example in hypothesis testing (See Section 3-3).

Fellegi and Sunter [1969] assume that the comparison vector $\boldsymbol{y} \in \Gamma$ is the realisation of a random variable

$$\boldsymbol{Y} = (Y^1, \ldots, Y^K) \tag{3-2}$$

for which each element is a random variable. For all $i \in \{1, \ldots, K\}$, the comparison value y^i is a realisation of Y^i. There is a vector $(\boldsymbol{Y}, M)$ of random variables for each record pair in $A \times B$. The realisation of the comparison vector $\boldsymbol{Y}$ can be observed, but true link status M is an unobserved latent variable. The true link status M is related to the realisation of $\boldsymbol{Y}$.

3-2-1 Optimal linkage rule

Fellegi and Sunter [1969] formulate their theory in terms of linkage rules. These linkage rules are used to classify record pairs into the set of links and the set of non-links. The linkage rules defined by Fellegi and Sunter are closely related to decision rules found in mathematical decision theory. A decision rule is a mapping of an observation into an action [Parmigiani and Inoue, 2009]. For the sets of links and non-links, the actions are the *positive link* action I and the *positive non-link* action III respectively. Fellegi and Sunter distinguish a third action, the *possible link* action II. This action is made if the realisation of $\boldsymbol{Y}$ is not informative enough to classify it into the set of links or the set of non-links. Together, these actions $\{\text{I}, \text{II}, \text{III}\}$ form the *action space*.

Consider a linkage function

$$d : \Gamma \to S \tag{3-3}$$

that is a mapping from the comparison space Γ into

$$S = \{\boldsymbol{p} \in [0,1]^3 : p_1 + p_2 + p_3 = 1\}.$$

Linkage function, or linkage rule, $d(\boldsymbol{y})$ yields a vector with functions $d_1(\boldsymbol{y})$, $d_2(\boldsymbol{y})$ and $d_3(\boldsymbol{y})$ corresponding with the probability of the actions I, II and III given $\boldsymbol{y} \in \Gamma$. Note that a linkage rule does not map the comparison space into an action but into a probability. This differs from a decision rule. Fellegi and Sunter point out that for some, or even all, realisations of $\boldsymbol{Y}$ the linkage function represents a degenerate random variable. For example, $d(\boldsymbol{y}) = (1, 0, 0)$ indicates that all record pairs have positive link action a_1.

A decision for an action does not necessarily correspond with the true link status of a record pair. Therefore, there are errors associated with the linkage rule defined in (3-3). A pair of records with link status $M = 1$ can have action III, i.e. a true link gets the positive non-link action. This type of error occurs with probability

$$\mathbb{E}[d_3(\boldsymbol{Y})|M = 1] = \sum_{\boldsymbol{y} \in \Gamma} P(\boldsymbol{Y} = \boldsymbol{y}|M = 1) d_3(\boldsymbol{y}). \tag{3-4}$$

The conditional probability $P(\boldsymbol{Y} = \boldsymbol{y}|M = 1)$ for $\boldsymbol{y} \in \Gamma$ plays a key role in the theory of Fellegi and Sunter. It is the probability of finding $\boldsymbol{y} \in \Gamma$ given that it is a true link. For all $\boldsymbol{y} \in \Gamma$, this probability is denoted by

$$m(\boldsymbol{y}) := P(\boldsymbol{Y} = \boldsymbol{y}|M = 1). \tag{3-5}$$

Write now Formula 3-4 as

$$\mathbb{E}[d_3(\boldsymbol{Y})|M=1] = \sum_{\boldsymbol{y}\in\Gamma} m(\boldsymbol{y}) d_3(\boldsymbol{y}). \tag{3-6}$$

In a similar way, a pair of records with link status $M = 0$ can get action I assigned to it. The probability of this type of error is

$$\mathbb{E}[d_1(\boldsymbol{Y})|M=0] = \sum_{\boldsymbol{y}\in\Gamma} P(\boldsymbol{Y}=\boldsymbol{y}|M=0) d_1(\boldsymbol{y}) \tag{3-7}$$

$$= \sum_{\boldsymbol{y}\in\Gamma} u(\boldsymbol{y}) d_1(\boldsymbol{y}) \tag{3-8}$$

where

$$u(\boldsymbol{y}) := P(\boldsymbol{Y}=\boldsymbol{y}|M=0) \tag{3-9}$$

for all $\boldsymbol{y} \in \Gamma$. Both the m-probability mass function and the u-probability mass function play an important role in the framework. The importance of these probabilities was already identified by Newcombe et al. [1959]. The linkage rule d is a linkage rule defined with the error levels

$$\mathbb{E}[d_1(\boldsymbol{Y})|M=0] = \mu \qquad \text{and} \qquad \mathbb{E}[d_3(\boldsymbol{Y})|M=1] = \lambda. \tag{3-10}$$

Fellegi and Sunter denote the linkage rule d on space Γ as $d(\mu, \lambda, \Gamma)$ for which $0 < \mu < 1$ and $0 < \lambda < 1$. The errors μ and λ are called the *error levels* of the linkage rule.

In Section 2-7, deterministic, probabilistic and rule-based classification principles are discussed. In fact, these classification techniques are mathematically equivalent to linkage rules. The class of linkage rules is infinitely large. An example of a linkage rule is $d(\boldsymbol{y}) = (0, 1, 0)$ for all $\boldsymbol{y} \in \Gamma$. This rule classifies all record pairs as possible links. Another example of a linkage rule is a rule that classifies all $\boldsymbol{y} \in \Gamma$ as positive links, positive non-links or possible links with a random process. If d is a linkage rule on Γ with the error levels μ and γ, then there are multiple linkage rules possible. Some rules are better and/or more useful than others. Fellegi and Sunter state that an optimal linkage rule is a linkage rule that minimises the probability of the possible link action for the given error levels. Definition 3-2.1 is the formal definition of an optimal linkage rule [Fellegi and Sunter, 1969].

Definition 3-2.1. *The linkage rule $d(\mu, \lambda, \Gamma)$ is said to be the optimal linkage rule if the relation*

$$\mathbb{E}[d_2(\boldsymbol{Y})] \leq \mathbb{E}[d'_2(\boldsymbol{Y})] \tag{3-11}$$

holds for every $d'(\mu, \lambda, \Gamma)$ in the class of linkage rules on Γ with error levels μ and λ.

Consider again the previously mentioned linkage rule $d(\boldsymbol{y}) = (0, 1, 0)$ for all $\boldsymbol{y} \in \Gamma$. This linkage rule classifies all record pairs as possible links. If $\mu = 0$, $\lambda = 0$ and $m(\boldsymbol{y}) > 0$, $u(\boldsymbol{y}) > 0$ for all $\boldsymbol{y} \in \Gamma$, this linkage rule is optimal. If $m(\boldsymbol{y}) = 0$ or $u(\boldsymbol{y}) = 0$ for some $\boldsymbol{y} \in \Gamma$, the linkage rule is not optimal. The rule is not optimal if the error levels are μ and λ are sufficiently larger than 0 such that it is possible to classify at least one record pair with a small probability as positive link or as positive non-link.

3-2-2 Fundamental theorem

The goal of Fellegi and Sunter was to define a linkage rule that is optimal, but also the best possible linkage rule on error levels μ and λ. The best linkage rule is the linkage rule with the most distinguishing power to distinguish the distributions of true links and true non-links. Fellegi and Sunter provide a theorem for the best, and also optimal, linkage rule. To achieve such a best linkage rule, an ordering of the comparison space Γ is necessary. The comparison vectors in Γ are stored in a sequence $(\boldsymbol{y}_i)_{i=1}^{N_\Gamma}$, where N_Γ is the number of elements in the comparison space Γ. The ordering of the comparison space Γ is based on the likelihood ratio

$$\Lambda(\boldsymbol{y}) := \frac{P(\boldsymbol{Y} = \boldsymbol{y}|M = 1)}{P(\boldsymbol{Y} = \boldsymbol{y}|M = 0)} = \frac{m(\boldsymbol{y})}{u(\boldsymbol{y})} \tag{3-12}$$

Comparison vectors in Γ are ordered such that the likelihood ratio is monotone decreasing, i.e.

$$\Lambda(\boldsymbol{y}_1) \geq \Lambda(\boldsymbol{y}_2) \geq \ldots \Lambda(\boldsymbol{y}_{N_\Gamma - 1}) \geq \Lambda(\boldsymbol{y}_{N_\Gamma}).$$

There are two special cases; the case for which multiple $\boldsymbol{y} \in \Gamma$ have the same likelihood ratio and the case for which $\boldsymbol{y} \in \Gamma$ has u-probability $u(\boldsymbol{y}) = 0$. For the first scenario, the order of the comparison vectors with identical likelihood ratios are stored in an arbitrary order. The comparison vectors $\{\boldsymbol{y} : u(\boldsymbol{y}) = 0, \boldsymbol{y} \in \Gamma\}$ are stored in an arbitrary order in beginning of the sequence. These comparison vectors are stored in the beginning of the sequence because $\boldsymbol{y}$ has a zero probability to occur in the set of non-links. They can only occur in the set of links.

The ordering of the comparison vectors is used to classify the comparison vectors into the positive link action set, positive non-link action set and possible link action set. The two error levels μ and λ are used to split the sequence of comparison vectors into the three sets. To split the sequence, choose two indicators n and n' in the ordered sequence $(\boldsymbol{y}_i)_{i=1}^{N_\Gamma}$ such that

$$\sum_{i=1}^{n-1} u(\boldsymbol{y}_i) < \mu \leq \sum_{i=1}^{n} u(\boldsymbol{y}_i) \tag{3-13}$$

$$\sum_{i=n'}^{N_\Gamma} m(\boldsymbol{y}_i) \geq \lambda > \sum_{i=n'+1}^{N_\Gamma} m(\boldsymbol{y}_i). \tag{3-14}$$

It is important that the linkage rule $d(\mu, \lambda, \Gamma)$ is admissible for the given error levels μ and λ. A decision rule is admissible if and only if there does not exist any better decision rule with identical error levels (the linkage rule is a sort of decision rule). Fellegi and Sunter claim that the linkage rule for the error levels μ and λ is always admissible if $1 < n < n' - 1 < N_\Gamma$. This restriction for n and n' prevents that a comparison vector $\boldsymbol{y} \in \Gamma$ is classified as a positive link as well as a positive non-link, because there is always at least 1 possible link separating the positive links from the positive non-links. It should be mentioned that the restriction $1 < n < n' - 1 < N_\Gamma$ can be relaxed in some situations.

The linkage rule $d(\mu, \lambda, \Gamma)$ proposed by Fellegi and Sunter is based on the ordering of the comparison space and the choice of error levels. The comparison vectors $(\boldsymbol{y}_i)_{i=1}^{n-1}$ get the

positive link action I. Because the sequence is monotone decreasing, these comparison vectors have relatively high likelihood ratios. The comparison vectors $(\boldsymbol{y}_i)_{i=n}^{n'-1}$ get the possible link action II. The subsequence $(\boldsymbol{y}_i)_{i=n'+1}^{N_\Gamma}$ has relatively small likelihood ratios and gets the positive non-link action III.

There are two comparison vectors left; the comparison vectors $\boldsymbol{y}_n$ and $\boldsymbol{y}_{n'}$. For these vectors, a random decision are made. Such a random decision is needed to achieve the exact error levels μ and λ. The random decisions make the linkage rule a variant of a random decision rule. For $\boldsymbol{y}_n$, random decisions are made between the positive link action I and the possible link action II. The random decisions are made such that error level μ is exactly met. For $i = n'$, random decisions are made between the positive non-link action III and the possible link action II such that the exact error level λ is met.

For $\boldsymbol{y}_i \in (\boldsymbol{y}_i)_{i=1}^{N_\Gamma}$, the randomised decision rule $d(\boldsymbol{y}_i)$ is summarised as

$$d(\boldsymbol{y}_i) = \begin{cases} (1,0,0) & \text{if } i \leq n-1 \\ (P_\mu, 1-P_\mu, 0) & \text{if } i = n \\ (0,1,0) & \text{if } n < i \leq n'-1 \\ (0, 1-P_\lambda, P_\lambda) & \text{if } i = n' \\ (0,0,1) & \text{if } i \geq n'+1. \end{cases} \tag{3-15}$$

This linkage rule is a mathematical formulation of what was discussed above. The random decisions are kept in the probabilities P_μ and P_λ. These probabilities need to ensure the error probabilities μ and λ are exactly met. The probabilities P_μ and P_λ are found with linear interpolation. The probability P_μ, solved from $\sum_{i=1}^{n-1} u(\boldsymbol{y}_i) + u(\boldsymbol{y}_n)P_\mu = \mu$, is

$$P_\mu = \frac{\mu - \sum_{i=1}^{n-1} u(\boldsymbol{y}_i)}{u(\boldsymbol{y}_n)}. \tag{3-16}$$

The probability

$$P_\lambda = \frac{\lambda - \sum_{i=n'+1}^{N_\Gamma} m(\boldsymbol{y}_i)}{m(\boldsymbol{y}_{n'})} \tag{3-17}$$

is derived from $\sum_{i=n'+1}^{N_\Gamma} m(\boldsymbol{y}_i) + m(\boldsymbol{y}_{n'})P_\lambda = \lambda$.

The linkage rule (3-15) is an optimal decision rule for error levels μ and λ on Γ [Fellegi and Sunter, 1969]. Fellegi and Sunter formulate the following theorem for this decision rule

Theorem 3.1. *Let $d(\mu, \lambda, \Gamma)$ be a linkage rule defined by* (3-15)*. Then $d(\mu, \lambda, \Gamma)$ is a best linkage rule on Γ at the levels (μ, λ).*

Proof. See Fellegi and Sunter [1969, p.1201-1207] □

Fellegi and Sunter prove that this linkage rule, of all randomised or non-randomised decision rules, is the best and optimal decision rule. The random decisions of this linkage rule are for some practical applications unacceptable. The corollaries in Section 3-2-3 make it easier to apply the linkage rule because the error levels are chosen in such a way that random decisions are not needed.

3-2-3 Corollaries

The random decisions of the best linkage rule are needed to achieve the error levels μ and λ. The random decisions are made for comparison vectors $\boldsymbol{y}_n$ and $\boldsymbol{y}_{n'}$. Particular choices of error levels can simplify the best (randomised) linkage rule. Suppose the error levels are chosen such that

$$(\mu, \lambda) \in \left\{ \sum_{i=1}^{n} u(\boldsymbol{y}_i), \sum_{i=n'}^{N_\Gamma} m(\boldsymbol{y}_i) : 1 \leq n < n' \leq N_\Gamma \right\} \tag{3-18}$$

Substituting these error levels into Formula 3-16 and Formula 3-17 results in $P_\mu = 1$ and $P_\lambda = 1$. For the error levels given in (3-18), the randomised best linkage rule becomes

$$d(\boldsymbol{y}_i) = \begin{cases} (1,0,0) & \text{if } 1 \leq i \leq n \\ (0,1,0) & \text{if } n < i < n' \\ (0,0,1) & \text{if } n' \leq i \leq N_\Gamma. \end{cases} \tag{3-19}$$

This linkage rule has no random decisions. Due to this choice of the error levels, there is no longer a situation in which record pair is classified as a positive link/potential link or positive non-link/possible link according to a random process. As a consequence of this choice, there is no possibility to set the error levels exactly if (μ, λ) are not part of the set in Formula 3-18.

To divide the set of comparison vectors based on the likelihood ratio, two threshold values are needed. The subsets correspond with the positive link action, the positive non-link action and the possible link action. The threshold levels

$$T_\mu := \frac{m(\boldsymbol{y}_n)}{u(\boldsymbol{y}_n)} \tag{3-20}$$

$$T_\lambda := \frac{m(\boldsymbol{y}_{n'})}{u(\boldsymbol{y}_{n'})} \tag{3-21}$$

are used divide the comparison space Γ according to the error levels given in (3-18). They divide the comparison space into

$$\Gamma_\mu = \{\boldsymbol{y} : T_\mu \leq m(\boldsymbol{y})/u(\boldsymbol{y})\} \tag{3-22}$$

$$\Gamma_\lambda = \{\boldsymbol{y} : m(\boldsymbol{y})/u(\boldsymbol{y}) \geq T_\lambda\}. \tag{3-23}$$

The choice of error levels as proposed in (3-18) leads to linkage rule

$$d(\boldsymbol{y}) = \begin{cases} (1,0,0) & \text{if } T_\mu \leq m(\boldsymbol{y})/u(\boldsymbol{y}) \\ (0,1,0) & \text{if } T_\lambda < m(\boldsymbol{y})/u(\boldsymbol{y}) < T_\mu \\ (0,0,1) & \text{if } m(\boldsymbol{y})/u(\boldsymbol{y}) \geq T_\lambda. \end{cases} \tag{3-24}$$

To avoid random decisions, the best decision rule for error levels (3-18) is used in (nearly) all record linkage applications.

3-3 The Fellegi and Sunter framework in the context of hypothesis testing

The theory of Fellegi and Sunter [1969] is closely related to the classical theory of statistical hypothesis testing. Two simple hypothesis tests divide the set of record pairs into the three sets of actions in the Fellegi and Sunter framework (the positive link action set, the positive non-link action set and the possible link action set). The Neyman-Pearson lemma plays an important role in this classification into three action sets. The lemma of Neyman and Pearson [1933] states that a hypothesis test with two simple hypotheses that rejects small values of the likelihood ratio is the most powerful test of all tests with significance level α [Rice, 2006]. In the theory of Fellegi and Sunter, the Neyman-Pearson lemma is implicitly applied to define the best linkage rule [Thibaudeau, 1992]. In fact, the best linkage rule is a combination of two most powerful hypothesis tests.

The stochastic data for a record pair in the Fellegi and Sunter framework is of the form $(\boldsymbol{Y}, M)$. The variable $\boldsymbol{Y}$ is the observed comparison vector and link status M is an unobserved latent variable. In this section, the true link status is not a random variable but as a parameter of the record pair[3]. As noted before, the Fellegi and Sunter framework is built on two simple hypothesises. Consider first the following hypothesis; a simple null hypothesis $H_0: \ M = 1$ (the record pair is a true link) against the simple alternative hypothesis $H_1: \ M = 0$ (the record pair is a true non-link). For the observed comparison vectors $\boldsymbol{y} \in \Gamma$, the hypothesis is

$$\begin{aligned} H_0: &\quad M = 1, \\ H_1: &\quad M = 0. \end{aligned}$$

The likelihood ratio for $\boldsymbol{y} \in \Gamma$ (Same as Formula 3-12)

$$\Lambda(\boldsymbol{y}) = \frac{P(\boldsymbol{Y} = \boldsymbol{y}|M = 1)}{P(\boldsymbol{Y} = \boldsymbol{y}|M = 0)} = \frac{m(\boldsymbol{y})}{u(\boldsymbol{y})}$$

is a measure for the relative plausibilities of H_0 and H_1 [Rice, 2006]. The likelihood ratio test is a hypothesis test that rejects the null hypothesis H_0 if the likelihood ratio is small. If the likelihood ratio test statistic is large, the null hypothesis is favoured. The Neyman-Pearson lemma states that this hypothesis test is the most powerful test. The function

$$\phi_1(\boldsymbol{y}) = \begin{cases} 0 & \text{if } \Lambda(\boldsymbol{y}) > k_1 \\ \gamma_1 & \text{if } \Lambda(\boldsymbol{y}) = k_1 \\ 1 & \text{if } \Lambda(\boldsymbol{y}) < k_1. \end{cases} \tag{3-25}$$

is the randomised decision rule of a likelihood ratio test for this hypothesis. The values $k_1 \in [0, \infty]$ and $\gamma_1 \in [0, 1]$ decide the outcomes of the test. The case $\Lambda(\boldsymbol{y}) = k_1$ is the randomised decision. This randomised case is used to obtain the exact significance level. It is comparable with the random decisions in the Fellegi ans Sunter model and formulated in Formula 3-17. If the likelihood ratio test statistic is small, then the decision rule ϕ_1 has

[3]There is no other notation chosen to prevent confusion, but one can think about using θ_M as parameter with values 0 and 1

outcome 1. This implies that the alternative hypothesis is better. If the likelihood ratio test statistic is large, then the decision rule ϕ_1 has outcome 0 and the null hypothesis is favoured. The significance level of this test is

$$\lambda = \mathbb{E}[\phi_1(\boldsymbol{Y})|H_0]$$

This significance level is similar to the error level in $\lambda = \mathbb{E}[d_3(\boldsymbol{Y})|M = 1]$ (Formula 3-10). The critical region for this decision rule is $\{\boldsymbol{y} : \phi_1(\boldsymbol{y}) = 1, \forall \boldsymbol{y} \in \Gamma\}$ and the acceptance region is $\{\boldsymbol{y} : \phi_1(\boldsymbol{y}) = 0, \forall \boldsymbol{y} \in \Gamma\}$. The acceptance region is in fact the set with the action I.

Fellegi and Sunter classify their model into 3 distinct subsets. Therefore, there is a similar symmetric approach needed. A second hypothesis test is needed; the simple null hypothesis $H_0 : \ M = 0$ (the record pair is a true non-link) against the simple alternative hypothesis $M = 1$ (the record pair is a true link). For $\boldsymbol{y} \in \Gamma$, the hypothesis is

$$\begin{aligned} H_0 : &\quad M = 0, \\ H_1 : &\quad M = 1. \end{aligned}$$

The likelihood ratio for $\boldsymbol{y} \in \Gamma$ is now

$$\frac{P(\boldsymbol{Y} = \boldsymbol{y}|M = 0)}{P(\boldsymbol{Y} = \boldsymbol{y}|M = 1)} = \frac{u(\boldsymbol{y})}{m(\boldsymbol{y})}. \tag{3-26}$$

It is the inverse of Formula 3-12, i.e. $1/\Lambda(\boldsymbol{y})$. Note that $1/\Lambda(\boldsymbol{y})$ is ordered monotone decreasing, while Fellegi and Sunter use a monotone increasing ordering. With this order, most non-links are in the beginning of the sequence. The randomised decision function for this hypothesis is

$$\phi_2(\boldsymbol{y}) = \begin{cases} 0 & \text{if } 1/\Lambda(\boldsymbol{y}) > k_2 \\ \gamma_2 & \text{if } 1/\Lambda(\boldsymbol{y}) = k_2 \\ 1 & \text{if } 1/\Lambda(\boldsymbol{y}) < k_2. \end{cases} \tag{3-27}$$

The values $k_2 \in [0, \infty]$ and $\gamma_2 \in [0, 1]$ are used to get the desired significance level. Also for this test holds: if the likelihood ratio test statistic is small, then the alternative hypothesis is favoured. A relative large test statistic is in favour of the null hypothesis. The significance level of this hypothesis test is

$$\mu = \mathbb{E}[\phi_2(\boldsymbol{Y})|H_0].$$

It is similar with the error level $\mu = \mathbb{E}[d_1(\boldsymbol{Y})|M = 0]$ in the Fellegi and Sunter model (Formula 3-10). The acceptance region is $\{\boldsymbol{y} : \phi_1(\boldsymbol{y}) = 0, \forall \boldsymbol{y} \in \Gamma\}$ for this hypothesis. The acceptance region is, in fact, the set with the action III.

With the two hypothesises, there are two regions of acceptance for comparison vectors $\boldsymbol{y} \in \Gamma$. One region is the set with action I (positive link action) and the other region is the set with action III (positive non-link action). The remaining $\boldsymbol{y} \in \Gamma$ are part of the set with action II (possible link set). Fellegi ans Sunter state that only for admissible error levels the linkage rule is the best possible linkage rule. In terms of hypothesis testing, the admissibility is easy to explain. It is admissible if the test fails to reject at most one of the null hypothesises. So, a comparison vector $\boldsymbol{y} \in \Gamma$ is never a positive link and a positive non-link.

3-4 Model assumptions and interpretations

Simplifications of the Fellegi and Sunter framework are used to make the record linkage model more easily applicable [Fellegi and Sunter, 1969]. A simplification that deals with the random decisions was discussed in Section 3-2-3. Fellegi and Sunter propose more simplifications to their model. An important and widely used simplification of the model is the, so called, *conditional independence assumption*. The comparison variables of the comparison vector are assumed to be mutually independent given the true link status [Herzog, Scheuren and Winkler, 2007]. This assumption is discussed in detail in Section 3-4-1. In Section 3-4-2, an assumption is discussed that assumes that the result of comparing attributes is restricted to 'agreement' and 'disagreement'.

3-4-1 Conditional independence assumption

One way to simplify the framework of Fellegi and Sunter is to assume that the components of comparison vector $\boldsymbol{Y}$ are mutually conditional independent given the true link status [Herzog, Scheuren and Winkler, 2007]. This assumption is called the *conditional independence assumption*. It can be seen as two separate assumptions. Assume that the components of the comparison vector are mutually independent if a pair is a true link. The second assumption that, if a pair is a true non-link, the components of the comparison vector are mutually independent.

The reason to make this assumption is the possibility to decompose the m-probability mass function and u-probability mass function in terms of marginal probability mass functions [Fellegi and Sunter, 1969]. Under the conditional independence assumption, the m-probability mass function in terms of marginal probabilities functions is

$$m(\boldsymbol{y}) = m_1(y^1) \cdot m_2(y^2) \cdot \ldots \cdot m_K(y^K) \tag{3-28}$$

where $m_i(y^i)$ is defined as

$$m_i(y^i) := P(Y^i = y^i | M = 1). \tag{3-29}$$

The marginal probability mass function $m_i(y^i)$ is the probability of observing $Y^i = y^i$ conditioned on $M = 1$. Also, the u-probability mass function can be written in terms of marginal probabilities with the conditional independence assumption. This results in

$$u(\boldsymbol{y}) = u_1(y^1) \cdot u_2(y^2) \cdot \ldots \cdot u_K(y^K) \tag{3-30}$$

where $u_i(y^i)$ is

$$u_i(y^i) := P(Y^i = y^i | M = 0). \tag{3-31}$$

The conditional independence assumption is often applied in record linkage [Thibaudeau, 1993]. There are several arguments to apply this assumption to the Fellegi and Sunter model. For many applications, this assumption is valid for the actual comparison data. A pair of records agreeing on the last name does not say anything, or only little, about agreement on the street name. Another reason to make this assumption is the reduction of parameters

of interest. A comparison vector $\boldsymbol{y} \in \Gamma$ of length 10 and two comparison values ('agreement' and 'disagreement') has $2^{10} = 1024$ possible comparison vectors in Γ. The m- and u-probability mass functions need to be estimated for all of these elements of Γ. The number of m and u estimates reduces to 20 when the conditional independence assumption is applied. Estimating parameters is also easier for some estimation methods when the conditional independence assumption is applied. For example, the mathematical complexity of the Expectation-Maximisation algorithm discussed in Section 4-2 is reduced under the assumption. Even if this assumption is violated, the classification is can be quite accurate for many record linkage applications [Herzog, Scheuren and Winkler, 2007]. The marginal probability functions $m(y^i)$ and $u(y^i)$ for $i \in \{1, \ldots, K\}$ are easy to understand and informative for the analyst. They tell the analyst what the probability is to find a certain comparison value, given the true link status.

Validity of the conditional independence assumption

The conditional independence assumption conditions on the unknown true link status M. Therefore, it is not possible to check directly whether the conditional independence assumption is reasonable. There are some approaches to measure conditional dependencies or conditional independence.

One of the methods is to classify the record pairs into a set of positive links and a set of positive non-links with the conditional independence assumption [Thibaudeau, 1993]. After this, a correlation matrix can be used to identify possible dependencies in the set of positive links and another correlation matrix can be used to identify dependencies in the set of positive non-links. Thibaudeau observed that among the linked pairs, little or no dependencies were found. Dependencies are especially found in the comparison vectors of non-links [Thibaudeau, 1993]. Therefore, the interesting parameters with respect to dependencies are the u-probability mass functions.

Daggy et al. [2013] describe a method to involve dependencies in the process of estimating parameters for the Fellegi and Sunter framework. Their method to measure dependencies calculates the correlation between the comparison variables for all record pairs. They include both true links and true non-links. Under this assumption, the correlation matrix is calculated for all comparisons [Daggy et al., 2013]. They implicitly assume that the number of non-links is much larger than the set of links.

In this thesis, an absolute difference is derived between the true u-probability mass function and the u-probability mass function under the conditional independence assumption. The conditional independence assumption implies that $u(\boldsymbol{y}) = \prod_{i=1}^{K} u_i(y^i)$ (Formula 3-30). If the data violates this assumption, consider the difference between $u(\boldsymbol{y})$ and $\prod_{i=1}^{K} u_i(y^i)$. The absolute difference of $u(\boldsymbol{y})$ and $\prod_{i=1}^{K} u_i(y^i)$ is

$$|u(\boldsymbol{y}) - \prod_{i=1}^{K} u_i(y^i)|. \tag{3-32}$$

This absolute difference is bounded between 0 and 1, because $0 \leq u(\boldsymbol{y}) \leq 1$ and $0 \leq \prod_{i=1}^{K} u_i(y^i) \leq 1$. A bound can be used to calculate the maximum difference between the

method under the conditional independence assumption and the true estimates. The absolute difference is given by

$$|u(\boldsymbol{y}) - \prod_{i=1}^{K} u_i(y^i)| \leq |f(\boldsymbol{y}) - \prod_{i=1}^{K} f_i(y^i)| + \frac{2\min(N_A, N_B)}{N_A \cdot N_B} \tag{3-33}$$

where $f(\boldsymbol{y}) := P(\boldsymbol{Y} = \boldsymbol{y})$ and $f_i(y^i) := P(Y^i = y^i)$. N_A and N_B are the number of records in files A and B respectively. The bound makes use of known file characteristics such as comparison values and the sizes of the files N_A and N_B. The proof is found in Appendix A. The absolute difference can be used to get an interval for the maximum difference between the estimates of the u-probability mass function in general and under the conditional independence assumption.

3-4-2 Binary assumption

Comparing the attributes of a record can result in a vector of agreement, disagreement or partial agreement (see Comparing record pairs, Section 2-6). Sometimes, record linkage techniques make use of multiple levels of agreement. They distinguish different cases of agreement such as 'agreement and the value is ...'. Fellegi and Sunter [1969] do not restrict the vector components of the comparison vector $\boldsymbol{y} \in \Gamma$ to particular types of agreement. In many applications, only 'agreement' and 'disagreement' is used, often denoted with 1 and 0 respectively[4]. This means that components of the comparison vector $\boldsymbol{y} = (y^1, \ldots, y^K)$ are restricted to $y^i \in \{0, 1\}$ for $i \in \{1, \ldots, K\}$. It results in a *binary vector* of length K. This assumption is widely used in practice and is called the *binary assumption*.

The usefulness of the binary assumption becomes clear in combination with the conditional independence assumption. Applying the two assumptions to the Fellegi and Sunter model implies that the marginal probability functions for m and u satisfy

$$\begin{aligned} m_i(0) + m_i(1) &= 1 \\ u_i(0) + u_i(1) &= 1. \end{aligned}$$

The assumption simplifies the m- and u-probability mass functions under the conditional independence assumption in Formula 3-28 and Formula 3-30. The m- and u-probability mass functions are now

$$m(\boldsymbol{y}) = \prod_{i=1}^{k} m_i(1)^{y^i} m_i(0)^{1-y^i} \tag{3-34}$$

and

$$u(\boldsymbol{y}) = \prod_{i=1}^{k} u_i(1)^{y^i} u_i(0)^{1-y^i} \tag{3-35}$$

[4]In this thesis, the label 2 is often used for agreement and the label 1 for disagreement. The label 0 is reserved for comparisons with at least one of the attributes is missing.

respectively. This notation is useful because it makes it easier to derive estimates for m- and u-probability mass functions. Another way of writing 3-34 and 3-35 is

$$m(\boldsymbol{y}) = \prod_{i=1}^{k} m_i(1)^{\mathbb{1}\{y^i=1\}} m_i(0)^{\mathbb{1}\{y^i=0\}} \tag{3-36}$$

and

$$u(\boldsymbol{y}) = \prod_{i=1}^{k} u_i(1)^{\mathbb{1}\{y^i=1\}} u_i(0)^{\mathbb{1}\{y^i=0\}}. \tag{3-37}$$

This notation can be generalised easily into a situation without the binary assumption (but with the conditional independence assumption). The notation is used in Chapter 4.

It is now possible to formulate the m-probability mass function only in terms of $m_1(1)$, $m_2(1)$, $\ldots$, $m_K(1)$. This is because $m_i(1) = 1 - m_i(0)$ for $i \in \{1, \ldots, K\}$. The same holds for the u-probability mass functions. This results in

$$m(\boldsymbol{y}) = \prod_{i=1}^{K} m_i(1)^{y^i} (1 - m_i(1))^{1-y^i} \tag{3-38}$$

$$= \prod_{i=1}^{K} m_i(1)^{\mathbb{1}\{y^i=1\}} (1 - m_i(1))^{\mathbb{1}\{y^i=0\}} \tag{3-39}$$

for the m-probability function and

$$u(\boldsymbol{y}) = \prod_{i=1}^{K} u_i(1)^{y^i} (1 - u_i(1))^{1-y^i} \tag{3-40}$$

$$= \prod_{i=1}^{K} u_i(1)^{\mathbb{1}\{y^i=1\}} (1 - u_i(1))^{\mathbb{1}\{y^i=0\}} \tag{3-41}$$

for the u-probability function. This reduces the number of estimates and the mathematical complexity for estimation of the parameters m and u. In Section 4-3, this is used for the estimation of parameters with the Expectation-Maximisation algorithm.

3-4-3 Computing weights

The monotone increasing ordering of $m(\boldsymbol{y})/u(\boldsymbol{y})$ in the Fellegi and Sunter framework is used to classify. Fellegi and Sunter [1969] state that it has some advantages to use the logarithm of $m(\boldsymbol{y})/u(\boldsymbol{y})$. It does not make any difference to the classification method. It is especially useful for interpretation. The logarithm of $m(\boldsymbol{y})/u(\boldsymbol{y})$ is called the *weight* of the comparison vector $\boldsymbol{y} \in \Gamma$ and the weight is

$$w(\boldsymbol{y}) = \log\left(\frac{m(\boldsymbol{y})}{u(\boldsymbol{y})}\right) = \log m(\boldsymbol{y}) - \log u(\boldsymbol{y}). \tag{3-42}$$

The weight is positive if $m(\boldsymbol{y}) > u(\boldsymbol{y})$ and negative if $m(\boldsymbol{y}) < u(\boldsymbol{y})$. Comparison vectors with high positive weights are likely to be links. Comparison vectors with low negative weights are

likely to be non-links. It is not necessarily true that comparison vectors with $w(\boldsymbol{y}) > 0$ are likely to be links and comparison vectors with $w(\boldsymbol{y}) < 0$ are likely to be non-links (explained later in this section). If all comparison vectors are the same or missing, the weight $w(\boldsymbol{y}) = 0$. In this situation, there is no information contained in the comparison vector.

If $u(\boldsymbol{y}) = 0$, then the weight is $w(\boldsymbol{y}) = \infty$. The case $u(\boldsymbol{y}) = 0$ occurs when the comparison vector $\boldsymbol{y} \in \Gamma$ only occurs in the set of links and not in the set of non-links. This happens when each record has an identifier such as a personal identification number. If $m(\boldsymbol{y}) = 0$, then the weight is $w(\boldsymbol{y}) = -\infty$. In this case, the comparison vector $\boldsymbol{y} \in \Gamma$ only occurs in the non-link set.

The conditional independence assumption makes it possible to calculate the weight of a single comparison variable. The weight of a comparison variable $\boldsymbol{y}_i$ for $i \in \{1, \ldots, K\}$ is given by

$$w_i(y^i) = \log\left(\frac{m_i(y^i)}{u_i(y^i)}\right) = \log m_i(y^i) - \log u_i(y^i). \tag{3-43}$$

The total weight in Formula 3-42 is the sum of the field specific weights

$$w(\boldsymbol{y}) = \log\left(\frac{m_1(y^1) \cdot m_2(y^2) \cdots m_K(y^K)}{u_1(y^1) \cdot u_2(y^2) \cdots u_K(y^K)}\right) \tag{3-44}$$

$$= w_1(y^1) + w_2(y^2) + \cdots + w_K(y^K). \tag{3-45}$$

If also the binary assumption is applied, the weight is [Herzog, Scheuren and Winkler, 2007]

$$w^i(y^i) = \begin{cases} \log\left(\frac{m_i(1)}{u_i(1)}\right) & \text{if } y^i = 1 \\ \log\left(\frac{1-m_i(1)}{1-u_i(1)}\right) & \text{if } y^i = 0. \end{cases}$$

As noted before, this way of looking to the problem is for the analyst easier to identify the important comparison variables. A large weight in the case of agreement is not directly related to a low weight for disagreement. In general, important variables have high (positive) weights in the case of agreement and low (negative) weights in the case of disagreement.

The sign of the weight $w(\boldsymbol{y})$ does not directly say something about the prevalence of being a link or a non-link. A (small) positive weight $w(\boldsymbol{y})$ does not imply that the record pair with comparison vector $\boldsymbol{y}$ is more likely to be a link than a non-link. To explain this, consider the ratio

$$\frac{P(M = 1|\boldsymbol{Y} = \boldsymbol{y})}{P(M = 0|\boldsymbol{Y} = \boldsymbol{y})} = \frac{P(\boldsymbol{Y} = \boldsymbol{y}|M = 1)P(M = 1)}{P(\boldsymbol{Y} = \boldsymbol{y}|M = 0)P(M = 0)}. \tag{3-46}$$

If this ratio is larger than 1, then the record pair is more likely to be belong to the same entity. If it is less than 1, then the record pair is more likely to be a non-link. Rewrite Formula 3-46 as

$$\frac{P(M = 1|\boldsymbol{Y} = \boldsymbol{y})}{P(M = 0|\boldsymbol{Y} = \boldsymbol{y})} = \frac{m(\boldsymbol{y})}{u(\boldsymbol{y})}\frac{\pi}{1-\pi}. \tag{3-47}$$

The term $\pi/(1-\pi)$ functions as a scaling factor between the ratio and the weight. A link prevalence $\pi = \frac{1}{2}$ means that the factor is 1. The sign of the weight $w(\boldsymbol{y})$ is then informative for the prevalence of being a link or non-link. The logarithm of Formula 3-47 is

$$\log\left(\frac{P(M=1|\boldsymbol{Y}=\boldsymbol{y})}{P(M=0|\boldsymbol{Y}=\boldsymbol{y})}\right) = w(\boldsymbol{y}) + \log\left(\frac{\pi}{1-\pi}\right). \tag{3-48}$$

In most situations, the number of true links is far less than the number of true non-links. This implies that the link prevalence is smaller than $\frac{1}{2}$. The term $\pi/(1-\pi)$ is bounded by $[0,1)$ if $\pi < \frac{1}{2}$. This means that the logarithm of the term is negative. Therefore, the weight for which $P(M=1|\boldsymbol{Y}=\boldsymbol{y}) = P(M=0|\boldsymbol{Y}=\boldsymbol{y})$ is positive.

3-4-4 Indexing and the Fellegi and Sunter framework

The main goal of indexing is to reduce the number of record pairs while keeping (most of) the links in it. This implies that indexing may change the link prevalence π. Not only the link prevalence is influenced by indexing, but the m- and u- probability mass functions can also change. The estimates are only informative for the set of comparison vectors on which the parameters are calculated. Note that this holds for all probabilistic record linkages with the Fellegi and Sunter framework.

Fellegi and Sunter [1969] describe the influence of indexing on their framework shortly. They formulate indexing as a restriction on the comparison space Γ. In other words, the candidate comparison vectors are a subspace of the comparison space $\Gamma^* \subseteq \Gamma$. They divide the comparison space into $\Gamma^* \subseteq \Gamma$ and $(\Gamma \setminus \Gamma^*) \subseteq \Gamma$ for which $\Gamma^* \cap (\Gamma \setminus \Gamma^*) = \Gamma$. They show that the error levels for μ and λ are not the same for cases with and without indexing. Consider the subspaces $\Gamma_\mu \subseteq \Gamma$ and $\Gamma_\lambda \subseteq \Gamma$ with error levels μ and λ as mentioned in Formula 3-22. The adjusted error levels are given by

$$\mu^* = \mu - \sum_{\boldsymbol{y} \in \Gamma_\mu \cap (\Gamma \setminus \Gamma^*)} m(\boldsymbol{y}) \tag{3-49}$$

and

$$\lambda^* = \lambda + \sum_{\boldsymbol{y} \in (\Gamma \setminus \Gamma_\lambda) \cap (\Gamma \setminus \Gamma^*)} u(\boldsymbol{y}). \tag{3-50}$$

Standard indexing, or blocking, is an effective tool for reduction of the number of comparison vectors [Baxter, Christen and Churches, 2003]. For blocking, the restriction on the comparison space suggested by Fellegi and Sunter is valid.

After the publication of the Fellegi and Sunter model, there are several new indexing methods developed. These indexing methods, such as Sorted Neighbourhood indexing (see Section 2-5-2) and Q-gram indexing (see Section 2-5-3), are no longer restrictions on the comparison space. These indexing methods reduce the number of record pairs, but do not necessarily reduce the number of elements in the comparison space. For example, the data is indexed with the Sorted Neighbourhood method on a string comparison field. Most of the time, some record pairs with disagreeing string comparisons are included in the candidate record pairs.

This type of indexation implies that all of the elements of the comparison space occur in candidate record pairs.

Consider a dataset with K comparison fields. The standard indexing method is applied on the first field, i.e. field $\boldsymbol{y}_1$. This implies that comparison space Γ is restricted under the assumption that $y^1 = 1$. Call this subspace for which $y^1 = 1$ the comparison space $\Gamma^* \subseteq \Gamma$. Also, assume that the data is conditional independent given the true link status and that the binary assumption holds. For $y^1 = 1$, it implies directly that

$$m_1(1) = P(Y^1 = 1|M = 1) = 1$$

and

$$u_1(1) = P(Y^1 = 1|M = 0) = 1$$

on the comparison space. Because $m_1(1)$ and $u_1(1)$ are equal to 1, the probability mass functions $m(\boldsymbol{y})$ and $u(\boldsymbol{y})$ for all $\boldsymbol{y} \in \Gamma$ are reduced to

$$m(\boldsymbol{y}) = m_2(y^2) \cdots m_K(y^K)$$

and

$$u(\boldsymbol{y}) = u_2(y^2) \cdots u_K(y^K).$$

Mention that the likelihood ratio $m_1(y^1)/u_1(y^1)$ is 1 for y^1 and the weight for this field is

$$w(y^1) = \log_2(1) = 0.$$

This implies that blocking can be applied on data without direct consequences for the estimation methods. The analyst can use the remaining fields for estimation. Even if the data is blocked before the decision maker has an influence on the data. In the Fellegi and Sunter model, this means that there is no information contained in the field.

If the data is indexed with a different indexing method on comparison variable $\boldsymbol{y}_1$, the probabilities $m_1(1)$ and $u_1(1)$ are not necessarily equal to 1. This happens with the Sorted Neighbourhood indexing method. The probabilities $m_1(1)$ and $u_1(1)$ are

$$m_1(1) = P(Y^1 = 1|M = 1) \leq 1$$

and

$$u_1(1) = P(Y^1 = 1|M = 0) \leq 1.$$

The probability functions $m(\boldsymbol{y})$ and $u(\boldsymbol{y})$ are now

$$m(\boldsymbol{y}) = m_1(y^1) \cdots m_K(y^K)$$

and

$$u(\boldsymbol{y}) = u_1(y^1) \cdots u_K(y^K)$$

for which the especially $u_1(y^1)$ is interesting. If $m_1(y^1) > u_1(y^1)$, then the weight $w(y^1)$ is positive. If $m_1(y^1) < u_1(y^1)$, then the weight $w(y^1)$ is negative. For each indexing method, the weight function $w(y^1)$ can be different because it depends on the set of comparison vectors.

3-4-5 Bayes' theorem for conditional probabilities

In some situations, the probabilities $P(M = 1|\boldsymbol{Y} = \boldsymbol{y})$ and $P(M = 0|\boldsymbol{Y} = \boldsymbol{y})$ are interesting for interpretation. It gives the probability of a true link or true non-link given the (known) comparison vector instead of conditioning on the latent variable M. In the Fellegi and Sunter model, the parameters of interest are the m, u and π probabilities. Both probabilities $P(M = 1|\boldsymbol{Y} = \boldsymbol{y})$ and $P(M = 0|\boldsymbol{Y} = \boldsymbol{y})$ are expressible in these terms by using Bayes' theorem.

Write the probability $P(M = 1|\boldsymbol{Y} = \boldsymbol{y})$ as

$$P(M = 1|\boldsymbol{Y} = \boldsymbol{y}) = \frac{P(\boldsymbol{Y} = \boldsymbol{y}|M = 1)P(M = 1)}{P(\boldsymbol{Y} = \boldsymbol{y})}$$

and substitute

$$P(\boldsymbol{Y} = \boldsymbol{y}) = P(\boldsymbol{Y} = \boldsymbol{y}|M = 1)P(M = 1) + P(\boldsymbol{Y} = \boldsymbol{y}|M = 0)P(M = 0).$$

The result is

$$P(M = 1|\boldsymbol{Y} = \boldsymbol{y}) = \frac{P(\boldsymbol{Y} = \boldsymbol{y}|M = 1)}{P(\boldsymbol{Y} = \boldsymbol{y}|M = 1)P(M = 1) + P(\boldsymbol{Y} = \boldsymbol{y}|M = 0)P(M = 0)}.$$

This formula can be written in terms of m- and u probabilities and the link prevalence $\pi := P(M = 1)$. It is given by

$$P(M = 1|\boldsymbol{Y} = \boldsymbol{y}) = \frac{m(\boldsymbol{y})\pi}{m(\boldsymbol{y})\pi + u(\boldsymbol{y})(1 - \pi)}. \tag{3-51}$$

The probability of a non-link given the comparison vector $\boldsymbol{y} \in \Gamma$ is

$$P(M = 0|\boldsymbol{Y} = \boldsymbol{y}) = \frac{u(\boldsymbol{y})(1 - \pi)}{m(\boldsymbol{y})\pi + u(\boldsymbol{y})(1 - \pi)}. \tag{3-52}$$

$P(M = 1|\boldsymbol{Y} = \boldsymbol{y})$ and $P(M = 0|\boldsymbol{Y} = \boldsymbol{y})$ add up to 1. In some situations, it is more informative for the interpretation to use $P(M = 1|\boldsymbol{Y} = \boldsymbol{y})$ and $P(M = 0|\boldsymbol{Y} = \boldsymbol{y})$ instead of m- and u-probability mass functions. Formula 3-51 and Formula 3-52 are used in Chapter 4 for the iterative estimation of parameters m, u and p.

Chapter 4

EM-Algorithm for estimation of parameters in the Fellegi and Sunter framework

4-1 Introduction

There are several methods available for unsupervised learning in the Fellegi and Sunter model [Fellegi and Sunter, 1969; Winkler, 1988, 1999]. The Expectation-Maximization (EM) algorithm applied in the context of the Fellegi and Sunter model is a popular method. The method does not need training or reviewed data to result in good estimates and, it has good convergence properties [Herzog, Scheuren and Winkler, 2007]. For this thesis, several other estimation methods are studied such as using log-linear models, Bayesian networks and the original method proposed by Fellegi and Sunter [Winkler, 1999]. Because of the good results obtained with the Expectation-Maximization algorithm in the literature, the focus of this thesis is on this estimation method. Much attention was payed to the mathematical formulation, because the EM-algorithm in the context of the Fellegi and Sunter model is very simplistic described in the available literature.

The Expectation-Maximization algorithm is an iterative algorithm used for computing maximum likelihood estimates for problems with incomplete data [McLachlan and Krishnan, 2007]. The EM-algorithm was described by Dempster, Laird and Rubin [1977]. They unified earlier work on estimating maximum likelihood estimates from incomplete data. For record linkage, the true link status is unobserved and is seen as incomplete data. Winkler [1988] describes how this iterative algorithm can be applied in the context of the Fellegi and Sunter [1969] framework. The algorithm is successfully applied to estimate the m- and u-probability mass functions for all $\gamma \in \Gamma$ and probability of randomly picking a pair of records that is a link [Winkler, 1988].

In this Chapter, the EM-algorithm in the context of the Fellegi and Sunter model is discussed

under several assumptions. In Section 4-2, the EM-algorithm is discussed in general. In Section 4-3, the EM-algorithm is discussed under the binary and conditional independence assumption. In Section 4-4, a variant of the EM-algorithm proposed by Schürle [2005] is discussed. This approach was developed to deal with conditional dependencies in the set of true links or the set of true non-links.

4-2 The Expectation-Maximization algorithm

The Expectation-Maximization problem starts with a set of observed comparison vectors $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_N$. The comparison vectors $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_N$ are realisations of random variables $\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_N$. The random variables $\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_N$ are assumed to be independently distributed. Each random variable $\boldsymbol{Y}_j$ is related to a random variable M_j for $1 \leq j \leq N$). It is the true link status of the record pair. The realisation of M_j is g_j where $g_j \in \{0, 1\}$ for $1 \leq j \leq N$). The stochastic data for each record pair in the Fellegi and Sunter model is $(\boldsymbol{Y}, M)$. Therefore, the complete data for this model is $(\boldsymbol{y}_1, g_1), \ldots, (\boldsymbol{y}_N, g_N)$. The goal is to estimate parameters depending on the complete data while only the incomplete data $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_N$ are available.

The vector of interesting parameters to estimate in the Fellegi and Sunter model is

$$\boldsymbol{\theta} = (m, u, \pi).$$

where the link prevalence $\pi := P(M = 1)$ is the probability of a randomly selected pair of records to be a true link. The m- and u-probability mass functions are parametrised according to all comparison vectors $\boldsymbol{y} \in \Gamma$. Therefore, the parameters for the m-probability mass functions span space $[0, 1]^{\Gamma}$. The parameters for the u-probability mass functions span $[0, 1]^{\Gamma}$ and the link prevalence is a parameter in $[0, 1]$. All the parameters in $\boldsymbol{\theta} \in [0, 1]^{\Gamma} \times [0, 1]^{\Gamma} \times [0, 1]$ represent probabilities.

The complete data likelihood for the Fellegi and Sunter model is given by [Herzog, Scheuren and Winkler, 2007]

$$\mathcal{L}(\boldsymbol{\theta}; g_1, \ldots, g_N, \boldsymbol{y}_1, \ldots, \boldsymbol{y}_N) = \prod_{j=1}^{N} P(\boldsymbol{Y}_j = \boldsymbol{y}_j, M_j = g_j). \tag{4-1}$$

Rewrite the complete data likelihood in Formula 4-1 as

$$\begin{aligned} &\mathcal{L}(\boldsymbol{\theta}; g_1, \ldots, g_N, \boldsymbol{y}_1, \ldots, \boldsymbol{y}_N) \\ &\quad = \prod_{j=1}^{N} P(\boldsymbol{Y}_j = \boldsymbol{y}_j, M_j = 1)^{g_j} P(\boldsymbol{Y}_j = \boldsymbol{y}_j, M_j = 0)^{1-g_j}. \end{aligned} \tag{4-2}$$

Substitute

$$\begin{aligned} P(\boldsymbol{Y} = \boldsymbol{y}, M = 1) &= P(\boldsymbol{Y} = \boldsymbol{y} | M = 1) P(M = 1) \\ &= m(\boldsymbol{y}) \cdot \pi \end{aligned} \tag{4-3}$$

and

$$\begin{aligned} P(\boldsymbol{Y} = \boldsymbol{y}, M = 0) &= P(\boldsymbol{Y} = \boldsymbol{y} | M = 0) P(M = 0) \\ &= u(\boldsymbol{y}) \cdot (1 - \pi) \end{aligned} \tag{4-4}$$

into the complete data likelihood function in Formula 4-2. Then, the complete data log-likelihood is given by

$$\begin{aligned}\log \mathcal{L}(\boldsymbol{\theta}; g_1, \ldots, g_N, \boldsymbol{y}_1, \ldots, \boldsymbol{y}_N) \\ = \sum_{j=1}^{N} g_j \log\left(\pi \cdot m(\boldsymbol{y}_j)\right) + (1 - g_j) \log\left((1-\pi) \cdot u(\boldsymbol{y}_j)\right)\end{aligned} \tag{4-5}$$

The complete data log-likelihood plays an important role in the EM-algorithm.

The EM-algorithm is an iterative algorithm for which each iteration consists of two steps; an Expectation step and a Maximization step. In the Expectation step, the expected value of the log-likelihood is calculated based on the current estimates of parameters $\boldsymbol{\theta}^{(t)}$ given the (independently) observed data $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_N$. The superscript $t \in \mathbb{N}_0$ is an integer indicating the iteration number.

For the Fellegi and Sunter model, the Expectation step is

$$Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}) = \mathbb{E}_{\boldsymbol{\theta}^{(t)}}\left[\log \mathcal{L}(\boldsymbol{\theta}; g_1, \ldots, g_N, \boldsymbol{y}_1, \ldots, \boldsymbol{y}_N) | \boldsymbol{Y}_1 = \boldsymbol{y}_1, \ldots, \boldsymbol{Y}_N = \boldsymbol{y}_N)\right]. \tag{4-6}$$

Insert the complete data log-likelihood of Formula 4-5 into the Expectation step. It results in

$$\begin{aligned}Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}) = \sum_{j=1}^{N} \mathbb{E}_{\boldsymbol{\theta}^{(t)}}\left[M_j | \boldsymbol{Y}_j = \boldsymbol{y}_j\right] \cdot \log(\pi \cdot m(\boldsymbol{y})) \\ + \sum_{j=1}^{N} \left(1 - \mathbb{E}_{\boldsymbol{\theta}^{(t)}}\left[M_j | \boldsymbol{Y}_j = \boldsymbol{y}_j\right]\right) \cdot \log\left((1-\pi) \cdot u(\boldsymbol{y})\right).\end{aligned} \tag{4-7}$$

Because M_j can only take the values 0 and 1, the expectation $\mathbb{E}_{\boldsymbol{\theta}^{(t)}}\left[M_j | \boldsymbol{Y}_j = \boldsymbol{y}_j\right]$ is

$$\mathbb{E}_{\boldsymbol{\theta}^{(t)}}\left[M_j | \boldsymbol{Y}_j = \boldsymbol{y}_j\right] = P_{\boldsymbol{\theta}^{(t)}}(M_j = 1 | \boldsymbol{Y}_j = \boldsymbol{y}_j).$$

This can be substituted into Formula 4-7, it results in

$$\begin{aligned}Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}) = \sum_{j=1}^{N} P_{\boldsymbol{\theta}^{(t)}}(M_j = 1 | \boldsymbol{Y}_j = \boldsymbol{y}_j) \cdot \log(\pi \cdot m(\boldsymbol{y}_j)) \\ + \sum_{j=1}^{N} P_{\boldsymbol{\theta}^{(t)}}(M_j = 0 | \boldsymbol{Y}_j = \boldsymbol{y}_j) \cdot \log\left((1-\pi) \cdot u(\boldsymbol{y}_j)\right).\end{aligned} \tag{4-8}$$

Formulate the probabilities $P_{\boldsymbol{\theta}^{(t)}}(M_j = 1 | \boldsymbol{Y}_j = \boldsymbol{y}_j)$ and $P_{\boldsymbol{\theta}^{(t)}}(M_j = 0 | \boldsymbol{Y}_j = \boldsymbol{y}_j)$ in terms of $\boldsymbol{\theta}^{(t)}$. The probabilities are closely related to Formula 3-51 and Formula 3-52. They are given by

$$P_{\boldsymbol{\theta}^{(t)}}(M = 1 | \boldsymbol{Y} = \boldsymbol{y}) = \frac{\pi^{(t)} \cdot m^{(t)}(\boldsymbol{y})}{\pi^{(t)} \cdot m^{(t)}(\boldsymbol{y}) + (1 - \pi^{(t)}) \cdot u^{(t)}(\boldsymbol{y})} \tag{4-9}$$

and

$$P_{\boldsymbol{\theta}^{(t)}}(M = 0 | \boldsymbol{Y} = \boldsymbol{y}) = \frac{(1 - \pi^{(t)}) \cdot u^{(t)}(\boldsymbol{y})}{\pi^{(t)} \cdot m^{(t)}(\boldsymbol{y}) + (1 - \pi^{(t)}) \cdot u^{(t)}(\boldsymbol{y})}. \tag{4-10}$$

In the Maximization step, the conditional expectation of the complete data log likelihood, $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)})$, is maximized with respect to parameters $\boldsymbol{\theta}$. The step is given by

$$\boldsymbol{\theta}^{(t+1)} = \arg\max_{\boldsymbol{\theta}} Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}). \tag{4-11}$$

For each component θ_n in $\boldsymbol{\theta}$, the maximized parameter is given by

$$\theta_n^{(t+1)} = \left(\arg\max_{\boldsymbol{\theta}} Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)})\right)_n. \tag{4-12}$$

All the parameters, the m-probabilities, u-probabilities and π, in the Fellegi and Sunter model can be estimated with the EM-algorithm. This type of general EM-algorithm is a solution to estimate the parameters of interest. The advantage of the general EM-algorithm discussed in this section is that there are no additional assumptions made, which are potentially violated. In the Section 4-2-1, the convergence properties are discussed and complications associated with this method.

4-2-1 Convergence properties and starting values

Wu [1983] wrote an article in which the convergence properties of the EM-algorithm are discussed. Wu considers a sequence of parameters $(\boldsymbol{\theta}^{(t)})_{t\geq 0}$ derived with the EM-algorithm. If $\boldsymbol{\theta}^{(t)} = \boldsymbol{\theta}^{(t+1)}$, then the set of parameters $\boldsymbol{\theta}^{(t)}$ is a stationary point of the incomplete data log-likelihood $\log \mathcal{L}(\boldsymbol{\theta}; \boldsymbol{y}_1, \ldots, \boldsymbol{y}_N)$. Wu also shows that a sequence of the incomplete data log-likelihood converges monotonically to a stationary point. It is important to note that the EM-algorithm converges to a stationary point, but not necessarily to a global maximum or not even to a local maximum.

In the context of record linkage in the Fellegi and Sunter model, the EM-algorithm can converge to incorrect local maximum values [Herzog, Scheuren and Winkler, 2007]. This convergence problem happens because there can be multiple stationary solutions. Convergence to incorrect stationary points also depends strongly on the choice of starting values. Different starting values may lead to different stationary points. An option for the choice of starting values is to use the ECM-algorithm described in Section 4-3 first to get reasonable starting values. This algorithm assumes conditional independence given the true link status. These estimates can be used as starting values for this algorithm[Herzog, Scheuren and Winkler, 2007].

The EM-algorithm as described in this section is rarely used in probabilistic record linkage applications, mainly because the Maximization step has no (known) closed form [Herzog, Scheuren and Winkler, 2007]. It makes the Maximization step computationally intensive, because iterative fitting methods, such as the Newton-Raphson method, are needed to solve it [Winkler, of the Census et al., 1993]. It requires many parameters to estimate. Consider N_Γ is the number of comparison vectors in Γ, then there are 2^{N_Γ} parameters for the m-probabilities, 2^{N_Γ} parameters for the u-probabilities and the link prevalence π.

4-3 The Expectation/ Conditional Maximization algorithm

A variant of the EM-algorithm proposed by Dempster, Laird and Rubin [1977] is the Expectation/Conditional Maximization (ECM) algorithm developed by Meng and Rubin [1993]. This algorithm is the normal EM-algorithm with an additional constraint to the likelihood. The ECM-algorithm makes use of the conditional independence assumption (see Section 3-4-1) in the log-likelihood. This assumption makes it possible to simplify the likelihood and formulate the Maximization step in a closed form. This makes the ECM-algorithm easier to apply. The ECM-algorithm has good convergence properties in contrast with the convergence properties of the EM-algorithm discussed in Section 4-2-1.

In Section 3-4-1 was seen that the m- and u-probability mass functions can be written in terms of marginal probability functions. For clarity, the m- and u-probability mass functions are given again;

$$m(\boldsymbol{y}) = m_1(y^1) \cdot m_2(y^2) \cdots m_K(y^K)$$
$$u(\boldsymbol{y}) = u_1(y^1) \cdot u_2(y^2) \cdots u_K(y^K).$$

Under the binary assumption, Formula 3-38 and Formula 3-40 explained that the number of parameters can be reduced even further. Both formulae are given again;

$$m(\boldsymbol{y}) = \prod_{i=1}^{K} m_i(1)^{y^i}[1 - m_i(1)]^{1-y^i}$$
$$u(\boldsymbol{y}) = \prod_{i=1}^{K} u_i(1)^{y^i}[1 - u_i(1)]^{1-y^i}.$$

Due to the binary assumption, it is sufficient to take only the comparison of agreement $y^i = 1$ into account (See Section 3-4-2). Straightforward calculations lead to the disagreement estimates ($m_i(0) = 1 - m_i(1)$). Therefore, the desired parameters for the Fellegi and Sunter model under this assumption is

$$\boldsymbol{\theta} = (m_1(1), \ldots, m_K(1), u_1(1), \ldots, u_K(1), \pi).$$

The number of parameters to estimate is $2K + 1$. The m- and u-probability mass functions under the conditional dependence and binary assumption are inserted in the Expectation step (Formula 4-8). The probabilities $P_{\boldsymbol{\theta}^{(t)}}(M = 1|\boldsymbol{Y} = \boldsymbol{y})$ and $P_{\boldsymbol{\theta}^{(t)}}(M = 0|\boldsymbol{Y} = \boldsymbol{y})$ can be written in terms of parameters of $\boldsymbol{\theta}^{(t)}$ and the observed comparison vectors. The probabilities $P_{\boldsymbol{\theta}^{(t)}}(M = 1|\boldsymbol{Y} = \boldsymbol{y})$ and $P_{\boldsymbol{\theta}^{(t)}}(M = 0|\boldsymbol{Y} = \boldsymbol{y})$ are then

$$P_{\boldsymbol{\theta}^{(t)}}(M = 1|\boldsymbol{Y} = \boldsymbol{y}) = \frac{\pi^{(t)} \prod_{i=1}^{K} (m_i^{(t)}(1))^{y^i} (1 - m_i^{(t)}(1))^{1-y^i}}{\pi^{(t)} \prod_{i=1}^{K} (m_i^{(t)}(1))^{y^i} (1 - m_i^{(t)}(1))^{1-y^i} + (1 - \pi^{(t)}) \prod_{i=1}^{K} (u_i^{(t)}(1))^{y^i} (1 - u_i^{(t)}(1))^{1-y^i}} \tag{4-13}$$

and

$$P_{\boldsymbol{\theta}^{(t)}}(M=0|\boldsymbol{Y}=\boldsymbol{y}) = \frac{(1-\pi^{(t)})\prod_{i=1}^{K}(u_i^{(t)}(1))^{y^i}(1-u_i^{(t)}(1))^{1-y^i}}{\pi^{(t)}\prod_{i=1}^{K}(m_i^{(t)}(1))^{y^i}(1-m_i^{(t)}(1))^{1-y^i}+(1-\pi^{(t)})\prod_{i=1}^{K}(u_i^{(t)}(1))^{y^i}(1-u_i^{(t)}(1))^{1-y^i}} \tag{4-14}$$

respectively. These probabilities are used to compute $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)})$ (Formula 4-8, the Expectation step).

The Maximization step maximizes the function $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)})$ with respect to the parameters of interest $\boldsymbol{\theta}$ (Formula 4-12). To maximize $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)})$, take the partial derivative of each parameter in $\boldsymbol{\theta}$ of $\boldsymbol{\theta} \mapsto Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)})$ and set it equal to zero. The partial derivative for parameter $m_i(1) \in \boldsymbol{\theta}$ to zero is

$$\frac{\partial}{\partial m_i(1)}\sum_{j=1}^{N} P_{\boldsymbol{\theta}^{(t)}}(M_j=1|\boldsymbol{Y}_j=\boldsymbol{y}_j)\log(\pi\prod_{i=1}^{n} m_i(1)^{y_j^i}[1-m_i(1)]^{1-y_j^i})=0.$$

The link prevalence π does not depend on the m-marginal probability mass functions and can be removed. The following simplification can be made,

$$\frac{\partial}{\partial m_i(1)}\sum_{j=1}^{N} P_{\boldsymbol{\theta}^{(t)}}(M_j=1|\boldsymbol{Y}_j=\boldsymbol{y}_j)\sum_{i=1}^{n}[y_j^i\log m_i(1)+(1-y_j^i)\log(1-m_i(1))=0.$$

The partial derivative of this expression is

$$\frac{\sum_{j=1}^{N} P_{\boldsymbol{\theta}^{(t)}}(M_j=1|\boldsymbol{Y}_j=\boldsymbol{y}_j)y^i}{m_i(1)} - \frac{\sum_{j=1}^{N} P_{\boldsymbol{\theta}^{(t)}}(M_j=1|\boldsymbol{Y}_j=\boldsymbol{y}_j)(1-y^i)}{1-m_i(1)}=0$$

Solving this equation for $m_i(1)$ results in

$$m_i^{(t+1)}(1)=\frac{\sum_{j=1}^{N} P_{\boldsymbol{\theta}^{(t)}}(M_j=1|\boldsymbol{Y}_j=\boldsymbol{y}_j)\ y_j^i}{\sum_{j=1}^{N} P_{\boldsymbol{\theta}^{(t)}}(M_j=1|\boldsymbol{Y}_j=\boldsymbol{y}_j)} \tag{4-15}$$

where $P_{\boldsymbol{\theta}^{(t)}}(M_j=1|\boldsymbol{Y}_j=\boldsymbol{y}_j)$ is given by Formula 4-13 and $P_{\boldsymbol{\theta}^{(t)}}(M_j=1|\boldsymbol{Y}_j=\boldsymbol{y}_j)>0$. The estimate of the marginal m-probability mass function is now in a closed form [Herzog, Scheuren and Winkler, 2007]. Note that $\sum_{j=1}^{N} P_{\boldsymbol{\theta}^{(t)}}(M_j=1|\boldsymbol{Y}_j=\boldsymbol{y}_j)>0$, which is the case if $P_{\boldsymbol{\theta}^{(t)}}(M=1|\boldsymbol{Y}=\boldsymbol{y})$ is not equal to zero for all $\boldsymbol{y}\in\Gamma$. The second partial derivative of $\boldsymbol{\theta} \mapsto Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)})$ with respect to $m_i(1)$ to ensure that this is indeed a maximum. The second derivative is

$$-\frac{\sum_{j=1}^{N} P_{\boldsymbol{\theta}^{(t)}}(M_j=1|\boldsymbol{Y}_j=\boldsymbol{y}_j)y^i}{(m_i(1))^2} - \frac{\sum_{j=1}^{N} P_{\boldsymbol{\theta}^{(t)}}(M_j=1|\boldsymbol{Y}_j=\boldsymbol{y}_j)(1-y^i)}{(1-m_i(1))^2}.$$

The denominators are both positive and at least one of the probabilities in the numerators is positive. Therefore,

$$\frac{\partial^2 Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)})}{\partial(m_i(1))^2}<0$$

and the estimate $m_i^{(t+1)}(1)$ is a local maximum.

In a similar way are the marginal u-probability mass functions derived. For field $i \in \{1, \ldots, K\}$, the marginal u-probability mass function is

$$u_i^{(t+1)}(1) = \frac{\sum_{j=1}^{N} P_{\boldsymbol{\theta}^{(t)}}(M = 0 | \boldsymbol{Y}_j = \boldsymbol{y}_j)\ y_j^i}{\sum_{j=1}^{N} P_{\boldsymbol{\theta}^{(t)}}(M = 0 | \boldsymbol{Y}_j = \boldsymbol{y}_j)} \tag{4-16}$$

for which $P_{\boldsymbol{\theta}^{(t)}}(M_j = 0 | \boldsymbol{Y}_j = \boldsymbol{y}_j)$ is given by Formula 4-14. The maximisation with respect to the link prevalence results π in

$$\pi^{(t+1)} = \frac{\sum_{j=1}^{N} P_{\boldsymbol{\theta}^{(t)}}(M = 1 | \boldsymbol{Y}_j = \boldsymbol{y}_j)}{N}. \tag{4-17}$$

This is the result of solving the partial derivative of $\boldsymbol{\theta} \mapsto Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)})$ with respect to π,

$$\frac{\sum_{j=1}^{N} P_{\boldsymbol{\theta}^{(t)}}(M_j = 1 | \boldsymbol{Y}_j = \boldsymbol{y}_j)}{\pi} - \frac{\sum_{j=1}^{N} P_{\boldsymbol{\theta}^{(t)}}(M_j = 0 | \boldsymbol{Y}_j = \boldsymbol{y}_j)}{(1 - \pi)} = 0, \tag{4-18}$$

to zero. The link prevalence is the average of the $P_{\boldsymbol{\theta}^{(t)}}(M = 1 | \boldsymbol{Y}_j = \boldsymbol{y}_j)$ probability calculations for all comparison vectors $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_N$.

There is a slightly different way of writing formula's 4-15, 4-16 and 4-17. This different way of writing uses the frequency of occurrence of a realisation $\boldsymbol{y} \in \Gamma$. Define the function $f : \Gamma \to \mathbb{N}_0$. This function counts the number of realisations of $\boldsymbol{y} \in \Gamma$ in the observed comparison vectors $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_N$. Formula 4-15 can be written as [Herzog, Scheuren and Winkler, 2007]

$$m_i^{(t+1)}(1) = \frac{\sum_{d=1}^{N_\Gamma} P_{\boldsymbol{\theta}^{(t)}}(M_d = 1 | \boldsymbol{Y}_d = \boldsymbol{y}_d)\ f(\boldsymbol{y}_d)\ y_d^i}{\sum_{d=1}^{N_\Gamma} P_{\boldsymbol{\theta}^{(t)}}(M_d = 1 | \boldsymbol{Y}_d = \boldsymbol{y}_d)\ f(\boldsymbol{y}_d)} \tag{4-19}$$

where $1, \ldots, N_\Gamma$ are indices for the elements in the comparison space Γ. The u-probability mass function in terms of frequencies is

$$u_i^{(t+1)}(1) = \frac{\sum_{d=1}^{N_\Gamma} P_{\boldsymbol{\theta}^{(t)}}(M_d = 0 | \boldsymbol{Y}_d = \boldsymbol{y}_d)\ f(\boldsymbol{y}_d)\ y_d^i}{\sum_{d=1}^{N_\Gamma} P_{\boldsymbol{\theta}^{(t)}}(M_d = 0 | \boldsymbol{Y}_d = \boldsymbol{y}_d)\ f(\boldsymbol{y}_d)} \tag{4-20}$$

In these terms, the link prevalence is

$$\pi^{(t+1)} = \frac{\sum_{d=1}^{N_\Gamma} P_{\boldsymbol{\theta}^{(t)}}(M = 1 | \boldsymbol{Y}_d = \boldsymbol{y}_d) f(\boldsymbol{y}_d)}{\sum_{d=1}^{N_\Gamma} f(\boldsymbol{y}_d)}. \tag{4-21}$$

There are two main reasons to write the parameters estimates in this way. Firstly, it can be easier to analyse the problem in the context of the comparison space Γ instead of N record pair comparisons. This is because a pair of records needs to be classified only based on the configuration of the comparison vector $\boldsymbol{y} \in \Gamma$. Two pairs of records can contain completely different information while having identical comparison vectors. Only for the comparison vector is a classification needed and not for both record pairs independently.

Regarding the computational implementation of the algorithm, this formulation is useful because fewer computations and memory are necessary to execute the algorithm itself. The

number N is often large, but the number of comparison vectors in Γ (N_Γ) is far less. The big $\mathcal{O}$ for each iteration is $\mathcal{O}(K \cdot N)$ operations against $O(K \cdot N_\Gamma)$ operations. It should be mentioned that the grouping of $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_N$ is of cost $\mathcal{O}(N)$, but this has to be done once instead of for each iteration. A second advantage is the reduction of computer memory usage. This is for large data sources of great advantage.

4-3-1 Convergence properties and starting values

The ECM-algorithm is known for its good convergence properties in the context of the Fellegi and Sunter model. Usually, the ECM-algorithm converges to stationary solutions that are unique for the parameter set $\boldsymbol{\theta}$ [Herzog, Scheuren and Winkler, 2007]. The ECM-algorithm has relatively fewer parameters to optimize. Therefore, the number of starting values is low.

An advantage of applying the ECM-algorithm is the good understandability and estimability of the (starting) m- and u probabilities. For field $i \in \{1, \ldots, K\}$, the probability $m_i(0)$ is the probability of an error in comparison field i given that the entities belong to each other. It is often easier to estimate this value from knowledge about the dataset, instead of estimating the probability of occurrence of comparison vector $\boldsymbol{y} \in \Gamma$ in the true links or true non-links. The starting values depend on knowledge about the data quality. In literature, the $m_i(1)$-marginal probabilities are chosen close to 1 and $u_i(1)$-marginal probabilities are chosen close to 0. Common values are 0.9 for the marginal $m_i(1)$ probabilities and values between 0.1 and 0.5 for the marginal $u_i(1)$ probabilities.

4-4 Conditional dependent parameter estimation with the EM-algorithm

Schürle [2005] applied another constraint to the likelihood in the Expectation-Maximization algorithm for the Fellegi and Sunter framework. Schürle developed the method to record linkage problems for which the conditional independence assumption is not reasonable for the available data. The method deals with dependencies between comparison variables given the true link status. To involve dependencies, Schürle uses the chain rule to write the m- and u-probability mass functions into products of conditional probabilities. This section describes the method proposed by Schürle. The binary assumption is applied to this assumption. In this thesis, a notation different to Schürle is used to make it possible to drop the binary assumption in the future.

Schürle [2005] uses the probability chain rule (or general product rule) to write the m- and u-probability mass functions in factors of conditional probabilities. The m- and u-probability mass functions are

$$m(\boldsymbol{y}) = m_1(y^1) m_2(y^2|y^1) \cdots m_K(y^K|y^1, \ldots, y^{K-1}) \tag{4-22}$$

and

$$u(\boldsymbol{y}) = u_1(y^1) u_2(y^2|y^1) \cdots u_K(y^K|y^1, \ldots, y^{K-1}) \tag{4-23}$$

respectively. For $i \in \{2, \ldots, K\}$, the conditional probability in (4-22) and (4-23) is defined as

$$m_i(y^i|y^1, \ldots, y^{i-1}) := P(Y^i = y^i|Y^1 = y^1, Y^2 = y^2, \ldots, Y^{i-1} = y^{i-1}, M = 1) \tag{4-24}$$

and

$$u_i(y^i|y^1, \ldots, y^{i-1}) := P(Y^i = y^i|Y^1 = y^1, Y^2 = y^2, \ldots, Y^{i-1} = y^{i-1}, M = 0) \tag{4-25}$$

respectively. For each m_i- and u_i-probability function, there are many configurations of the conditionalized variables $Y^1, Y^2, \ldots, Y^{i-1}$. This is where Schürle [2005] applied the binary assumption is his paper. The binary assumption implies that there are 2^{i-1} configurations of field $i \in \{2, \ldots, K\}$ for Formula 4-24 and 2^{i-1} configurations of field $i \in \{2, \ldots, K\}$ for Formula 4-25.

Each realisation $Y^1 = y^1, Y^2 = y^2, \ldots, Y^{i-1} = y^{i-1}$ needs a unique indicator that indicates the configuration of the realisation. Consider a function

$$z_i : \Gamma \to \{\mathbf{e}_1, \ldots, \mathbf{e}_\nu\} \tag{4-26}$$

where $\mathbf{e}_1, \ldots, \mathbf{e}_\nu$ are vectors in the standard basis of $\mathbb{R}^\nu$ and $1 \leq \nu \leq 2^{i-1}$. ν is equal to 2^{i-1} if all possible comparison vectors are found in Γ. The function z_i is a mapping of the comparison vector $\boldsymbol{y}$ on a unique vector. The mapping is not further specified.

Now redefine Formula 4-24 and Formula 4-25 by using those unique indicating vectors. The m- and u-probability mass functions are now

$$m_i(y^i|z_i(\boldsymbol{y})) := m_i(y^i|y^1, \ldots, y^{i-1}) \tag{4-27}$$

$$u_i(y^i|z_i(\boldsymbol{y})) := u_i(y^i|y^1, \ldots, y^{i-1}). \tag{4-28}$$

Due to the binary assumption, the probability $m_i(1|z_i(\boldsymbol{y}))$ is equal to $1 - m_i(0|z_i(\boldsymbol{y}))$ (see Section 3-4-2). Therefore, only one of these terms need to be included in the set of parameters to estimate.

Assign all possible configurations of the marginal probabilities to the following m- and u-probability vectors

$$\mathbf{m}_i(1) := \big(m_i(1|\mathbf{e}_1), \ldots, m_i(1|\mathbf{e}_\nu)\big)^T \qquad i \in \{2, \ldots, K\} \tag{4-29}$$

$$\mathbf{u}_i(1) := \big(u_i(1|\mathbf{e}_1), \ldots, u_i(1|\mathbf{e}_\nu)\big)^T \qquad i \in \{2, \ldots, K\} \tag{4-30}$$

The case $i = 1$ is a special case because it has no conditioning. Define the m_i- and u_i-probability vectors for $i = 1$ as

$$\mathbf{m}_1(1) := m_1(1) \qquad \text{and} \qquad \mathbf{u}_1(1) := u_1(1)$$

respectively.

Using this notation, it is possible to express the probabilities $m(\boldsymbol{y})$ and $u(\boldsymbol{y})$ in terms of the just mentioned notation for all $\boldsymbol{y} \in \Gamma$. The m- and u-probability functions (4-22) and (4-23) are

$$m(\boldsymbol{y}) = \prod_{i=1}^{K} [(z_i(\boldsymbol{y}))^T \cdot \mathbf{m}_i(1)]^{y^i} [1 - (z_i(\boldsymbol{y}))^T \cdot \mathbf{m}_i(1)]^{1-y^i} \tag{4-31}$$

and

$$u(\boldsymbol{y}) = \prod_{i=1}^{K} [(z_i(\boldsymbol{y}))^T \cdot \mathbf{u}_i(1)]^{y^i} [1 - (z_i(\boldsymbol{y}))^T \cdot \mathbf{u}_i(1)]^{1-y^i} \tag{4-32}$$

respectively. This expression with the chain rule plays an important role in the Expectation-Maximization algorithm of Schürle [2005].

The parameters of interest for this version of the EM-algorithm are

$$\boldsymbol{\theta} = (\mathbf{m}_1(1), \ldots, \mathbf{m}_K(1), \mathbf{u}_1(1), \ldots, \mathbf{u}_K(1), \pi).$$

When the comparison values are restricted to 'agreement' and 'disagreement', then the vectors $\mathbf{m}_1(1), \ldots, \mathbf{m}_K(1)$ contain in total $2^K - 1$ parameters. Also, the u-probabilities require $2^K - 1$ parameters. With the link prevalence, there are $2^{K+1} - 1$ parameters $\boldsymbol{\theta} \in [0,1]^{2^{K+1}-1}$.

The conditional expectation of the log-likelihood $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)})$ is identical to the Expectation step in the general EM-algorithm. For clarity, the $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)})$ (Formula 4-8) is given by

$$\begin{aligned} Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}) = &\sum_{j=1}^{N} P_{\boldsymbol{\theta}^{(t)}}(M_j = 1|\boldsymbol{Y}_j = \boldsymbol{y}_j) \log(\pi\, m(\boldsymbol{y})) \\ &+ \sum_{j=1}^{N} P_{\boldsymbol{\theta}^{(t)}}(M_j = 0|\boldsymbol{Y}_j = \boldsymbol{y}_j) \log\big((1-\pi) u(\boldsymbol{y})\big). \end{aligned}$$

The m- and u-probability mass functions in $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)})$ are given by (4-31) and (4-32) respectively. In the Maximization step, the argument of the maximum of $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)})$ is computed. The partial derivative with respect to parameter $m_i(1|\mathbf{e})$ for $\mathbf{e} \in \{\mathbf{e}_1, \ldots, \mathbf{e}_\nu\}$ is given by

$$\begin{aligned} \frac{\partial Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)})}{\partial m_i(1|\mathbf{e})} &= \sum_{j=1}^{N} P_{\boldsymbol{\theta}^{(t)}}(M_j = 0|\boldsymbol{Y}_j = \boldsymbol{y}_j) \frac{\partial}{\partial m_i(1|\mathbf{e})} \ln m(\boldsymbol{y}_j) \\ &= \sum_{j=1}^{N} P_{\boldsymbol{\theta}^{(t)}}(M_j = 0|\boldsymbol{Y}_j = \boldsymbol{y}_j) \left(\frac{y_j^i}{(z_i(\boldsymbol{y}_j))^T \cdot \mathbf{m}_i(1)} - \frac{1 - y_j^i}{1 - (z_i(\boldsymbol{y}_j))^T \cdot \mathbf{m}_i(1)} \right) \\ &= \sum_{\{j=1|z_i(\boldsymbol{y}_j)=\mathbf{e}\}}^{N} P_{\boldsymbol{\theta}^{(t)}}(M_j = 0|\boldsymbol{Y}_j = \boldsymbol{y}_j) \frac{y_j^i - m_i(1|\mathbf{e})}{m_i(1|\mathbf{e})(1 - m_i(1|\mathbf{e}))} \end{aligned}$$

Set this derivative to zero and solve for $m_i(1|\mathbf{e})$. It results in

$$m_i^{(t+1)}(1|\mathbf{e}) = \frac{\sum_{\{j=1|z_i(\boldsymbol{y}_j)=\mathbf{e}\}}^{N} P_{\boldsymbol{\theta}^{(t)}}(M_j = 1|\boldsymbol{Y}_j = \boldsymbol{y}_j) y_j^i}{\sum_{\{j=1|z_i(\boldsymbol{y}_j)=\mathbf{e}\}}^{N} P_{\boldsymbol{\theta}^{(t)}}(M_j = 1|\boldsymbol{Y}_j = \boldsymbol{y}_j)} \tag{4-33}$$

For the (marginal) u-probabilities, the same approach leads to

$$m_i^{(t+1)}(1|\mathbf{e}) = \frac{\sum_{\{j=1|z_i(\boldsymbol{y}_j)=\mathbf{e}\}}^{N} P_{\boldsymbol{\theta}^{(t)}}(M_j = 0|\boldsymbol{Y}_j = \boldsymbol{y}_j) y_j^i}{\sum_{\{j=1|z_i(\boldsymbol{y}_j)=\mathbf{e}\}}^{N} P_{\boldsymbol{\theta}^{(t)}}(M_j = 0|\boldsymbol{Y}_j = \boldsymbol{y}_j)} \tag{4-34}$$

The maximisation of parameter $\pi = P(M = 1)$ gives the same result as for the ECM-algorithm (Formula 4-17). Therefore, the estimate for $\pi^{(t+1)}$ is

$$\pi^{(t+1)} = \frac{\sum_{j=1}^{N} P_{\boldsymbol{\theta}^{(t)}}(M_j = 0|\boldsymbol{Y}_j = \boldsymbol{y}_j)}{N}. \tag{4-35}$$

Note that $P_{\boldsymbol{\theta}^{(t)}}(M_j = 0|\boldsymbol{Y}_j = \boldsymbol{y}_j)$ differs from the method described in the ECM-algorithm.

4-4-1 Convergence properties and starting values

The application of the EM-algorithm proposed by Schürle [2005] has a remarkable convergence property. This version of the iterative EM-algorithm converges to a stationary point in one iteration [Schürle, 2005]. A proper choice of the starting values is important for the conditional dependent EM-algorithm. The algorithm has multiple stationary points.

To prove that the conditional dependent EM-algorithm converges in one iteration, Schürle [2005] showed that the probabilities $P_{\boldsymbol{\theta}^{(t)}}(M = 1|\boldsymbol{Y} = \boldsymbol{y})$ and $P_{\boldsymbol{\theta}^{(t)}}(M = 0|\boldsymbol{Y} = \boldsymbol{y})$ do not change after the first iteration. This is reported in the following theorem

Theorem 4.1 ([Schürle, 2005])**.** *If the conditonal dependence method is applied, then*

$$P_{\boldsymbol{\theta}^{(t+1)}}(M = 1|\boldsymbol{Y} = \boldsymbol{y}) = P_{\boldsymbol{\theta}^{(t)}}(M = 1|\boldsymbol{Y} = \boldsymbol{y})$$

and

$$P_{\boldsymbol{\theta}^{(t+1)}}(M = 0|\boldsymbol{Y} = \boldsymbol{y}) = P_{\boldsymbol{\theta}^{(t)}}(M = 0|\boldsymbol{Y} = \boldsymbol{y})$$

for $t = 0, 1, 2, \ldots$.

Proof. See Schürle [2005, p.442] □

This theorem results in the following corollary.

Corollary 4.1.1 ([Schürle, 2005])**.** *If the conditional independence assumption is applied, then Theorem 4.1 implies in conjunction with Formula 4-33, Formula 4-34 and Formula 4-35 that*

$$\boldsymbol{\theta}^{(t+1)} = \boldsymbol{\theta}^{(t)}$$

for $t = 0, 1, 2, \ldots$. *In conjunction with the theorem of Wu [1983, p.98], described in Section 4-2-1, it follows that the conditional dependence method leads to a stationary point of* $\mathcal{L}(\boldsymbol{\theta})$ *after the first iteration.*

Proof. See Schürle [2005, p.442] □

This corollary shows that the conditional dependent EM-algorithm converges in a single Expectation-Maximization step. The initial parameters restrict the solution to a small subspace of solutions for $\boldsymbol{\theta}$. Therefore, the starting values of this conditional dependent EM-algorithm play an important role [Schürle, 2005].

The choice of good starting values is a process for which there is, so far, no best solution. Schürle proposes a method in which conditional independent estimates of the probability mass functions m and u are used as starting values for the conditional dependent EM-algorithm. For all fields $i \in \{1, \ldots, K\}$ and for all $z_i(\boldsymbol{y})$, the starting values for $m_i(1|z_i(\boldsymbol{y}))$ and $u_i(1|z_i(\boldsymbol{y}))$ are given by

$$m_i^{(0)}(1|z_i(\boldsymbol{y})) = m_i(1)$$

and

$$u_i^{(0)}(1|z_i(\boldsymbol{y})) = u_i(1)$$

respectively. The initial value for parameter π is

$$\pi^{(0)} = \pi.$$

A reason to define the starting values in this way is because it involves far less initial estimates. Instead of $2(2^K - 1) + 1$ initial estimates, there are $2K + 1$ estimates needed. Schürle uses knowledge about the data and file characteristics for initial, conditional independent, estimates for $\boldsymbol{\theta}$. Although not mentioned by Schürle, also the ECM-estimates can be used as starting values.

Chapter 5

Estimation of parameters in the Fellegi and Sunter framework based on the distribution of characteristics

5-1 Introduction

One of the popular aspects of the Fellegi and Sunter [1969] framework is the possibility to use the distribution of characteristics in the populations for classification. Records with rare entity characteristics, such as persons with rare names, are more likely to be links than entities with common attribute values. A returning example in the theory of record linkage are the names Zabrinksy and Smith [Winkler, 1988]. The name Smith is quite common in the USA while Zabrinksy is not. There are not many Zabrinksy's, so two randomly picked records with the name Zabrinksy are more likely to belong to the same person than two records with the name Smith. Not only names can be used to add additional information to the classification, but also characteristics like the zip code, hair colour and sex are useful. Fellegi and Sunter use the distribution of characteristics for estimation of the m- and u-marginal probability mass functions. Sometimes, using the distribution of characteristics for estimation is called *frequency based estimation* [Winkler, 1999].

Frequency based estimation is performed on one characteristic of the entity. All the possible (unique) characteristics of the entity found in $\mathcal{A} \cup \mathcal{B}$ are collected and given by

$$v_1, v_2, \ldots, v_Q$$

where Q is the number of unique attribute values found in $\mathcal{A} \cup \mathcal{B}$. Each value occurs

$$f_{\mathcal{A},1}, f_{\mathcal{A},2}, \ldots, f_{\mathcal{A},Q}$$

times in population $\mathcal{A}$. The sum of the frequency of occurrence

$$\sum_{i=1}^{Q} f_{\mathcal{A},i} = N_{\mathcal{A}}$$

is equal to the number of entities $N_{\mathcal{A}}$ in population $\mathcal{A}$. The frequency of occurrence in population $\mathcal{B}$ is

$$f_{\mathcal{B},1}, f_{\mathcal{B},2}, \ldots, f_{\mathcal{B},Q}$$

for which the sum of the frequencies

$$\sum_{i=1}^{Q} f_{\mathcal{B},i} = N_{\mathcal{B}}$$

is the size of the population $\mathcal{B}$.

The set of true links and true non-links also have certain distribution for this characteristic. It is sufficient to know one of them. The frequency that an attribute value is observed in the population of links $\mathcal{M}$ is given by

$$f_{\mathcal{M},1}, f_{\mathcal{M},2}, \ldots, f_{\mathcal{M},Q}$$

where the sum of the frequencies

$$\sum_{i=1}^{Q} f_{\mathcal{M},i} = N_{\mathcal{M}}$$

is the number of true links $N_{\mathcal{M}}$ between both populations.

The datasets A and B are incomplete representations of the populations $\mathcal{A}$ and $\mathcal{B}$. Besides incompleteness, also errors may occur in the data. There are three types of errors and incompleteness was identified by Fellegi and Sunter [1969].

e_A **and** e_B The probability that an attribute is misreported in dataset A or dataset B respectively. Assume that a misreport is independent of the particular value. For names this assumption is easily violated, complicated, or uncommon, names are more often misspelled.

e_{A_0} **and** e_{B_0} The probability that an attribute is not reported in dataset A or dataset B respectively. A not-reported value is independent of the particular value.

e_T The probability that an attribute is different in dataset A and dataset B, but it is not a mistake. This could happen when the value of the field changes between the generation of the record in dataset A and in dataset B. Think about a change of name, marital status or profession.

In the next two sections, two methods are discussed for the estimation of parameters that make use of the distribution of the characteristic in the populations. The first estimation method, described in Section 5-2, is a method proposed by Fellegi and Sunter. The second is a related method that tries to deal with the lack of knowledge about $f_{\mathcal{M},1}, f_{\mathcal{M},2}, \ldots, f_{\mathcal{M},Q}$.

Both methods estimate the m- and u-marginal probability mass functions for three cases of comparison outcomes. The three cases are; the comparison agrees and the attribute in both records is one of the attributes $v_1, v_2, \ldots, v_Q$, the comparison disagrees and in one, or both, of the records is the attribute missing. For simplicity, the indicator/label 0 is used for disagreement, -1 for a pair of records with missing attribute values and $1, \ldots, Q$ for records pairs that agree on $v_1, v_2, \ldots, v_Q$ respectively.

5-2 Frequency based estimation of parameters (Fellegi and Sunter)

Fellegi and Sunter [1969] distinguish the three cases described above for "frequency based" estimation. Therefore, the labels $\{-1, 0, 1, 2, \ldots, Q\}$ are of interest for estimation of parameters. For simplicity, define the m- and u-marginal probability mass functions for these labels as

$$m_i(q) := P(Y^i = q | M = 1) \tag{5-1}$$

and

$$u_i(q) := P(Y^i = q | M = 0) \tag{5-2}$$

for which q is the label of the comparison and $i \in \{1, \ldots, K\}$. The labels correspond to the earlier mentioned types of comparison.

Situation: The attribute in both records is v_q for which $q \in \{1, \ldots, Q\}$
The m-marginal probability mass function is given by

$$m(q) = \frac{f_q}{N_{\mathcal{M}}}(1 - e_A)(1 - e_B)(1 - e_T)(1 - e_{A_0})(1 - e_{B_0}). \tag{5-3}$$

The u-marginal probability mass function is given by

$$u(q) = \frac{f_{\mathcal{A},q} f_{\mathcal{B},q}}{N_{\mathcal{A}} N_{\mathcal{B}}}(1 - e_A)(1 - e_B)(1 - e_T)(1 - e_{A_0})(1 - e_{B_0}). \tag{5-4}$$

Both Formula 5-3 and Formula 5-4 can be used to calculate the weight of this variable and attribute v_c (Section 3-4-3). The weight is given by

$$w(q) = \log\left(\frac{f_{M,q} N_{\mathcal{A}} N_{\mathcal{B}}}{f_{\mathcal{A},q} f_{\mathcal{B},q} N_{\mathcal{M}}}\right). \tag{5-5}$$

The weight does not depend on any of the errors defined above. In Section 2-5 was mentioned that the number of links scales linearly and the number of non-links scales quadratically. With Formula 5-3 and Formula 5-4 can be seen that this property is exploited to get frequency based weights.

Fellegi and Sunter [1969] state that the proportions $f_{\mathcal{A},q}/N_{\mathcal{A}}$, $f_{\mathcal{B},q}/N_{\mathcal{B}}$ and $f_{\mathcal{M},q}/N_{\mathcal{M}}$ for $q \in \{1, \ldots, Q\}$ can be assumed identical in most of the cases[1]. If the proportions are assumed

[1] This is were Fellegi and Sunter make use of the simple random sampling assumption made at the beginning of Chapter 3. This assumption makes it possible to estimate the proportions given that both datasets represent the same population. The assumption is quickly violated in practice.

to be identical in the population of links $\mathcal{M}$ and populations $\mathcal{A}$ and $\mathcal{B}$, and then define

$$p_q := \frac{f_{\mathcal{A},q}}{N_{\mathcal{A}}} = \frac{f_{\mathcal{B},q}}{N_{\mathcal{B}}} = \frac{f_{\mathcal{M},q}}{N_{\mathcal{M}}}. \tag{5-6}$$

The weight of agreement on the value with index c is

$$w(q) = \log\left(\frac{1}{p_q}\right). \tag{5-7}$$

This method is sometimes assigned to Gill [2001] who formalised it. Formula 5-7 shows that an uncommon attribute value (value with a low p_q) has a larger weight than a common value. For example, the name Smith occurs has proportion 1/10 in all populations. Then the weight is log(10). The proportion of people with the name Zabrinksy is 1/100, then the weight is log(100). So, Zabrinksy gets a much larger weight.

Situation: The attribute in both records disagree
The m-marginal probability mass function for a disagreeing comparison is given by

$$m(0) = [1 - (1 - e_A)(1 - e_B)(1 - e_T)](1 - e_{A_0})(1 - e_{B_0}) \tag{5-8}$$

and the u-marginal probability mass function by

$$u(0) = [1 - (1 - e_A)(1 - e_B)(1 - e_T)\sum_{j=1}^{Q} \frac{f_{\mathcal{A},j} f_{\mathcal{B},j}}{N_{\mathcal{A}} N_{\mathcal{B}}}](1 - e_{A_0})(1 - e_{B_0}). \tag{5-9}$$

The weight for a disagreeing comparison is given by

$$w(0) = \log\left(\frac{e_A + e_B + e_T}{1 - (1 - e_A - e_B - e_T)\sum_{j=1}^{Q} \frac{f_{\mathcal{A},j} f_{\mathcal{B},j}}{N_{\mathcal{A}} N_{\mathcal{B}}}}\right) \tag{5-10}$$

The weight is identical for all types of disagreement. There is no difference in different kinds of disagreement between values.

Situation: At least one attribute in the pair of records is missing
The m-marginal probability mass function is given by

$$m(-1) = 1 - (1 - e_{A_0})(1 - e_{B_0}) \tag{5-11}$$

and the u-marginal probability mass function by

$$u(-1) = 1 - (1 - e_{A_0})(1 - e_{B_0}). \tag{5-12}$$

Both probabilities are identical. Therefore, the weight

$$w(-1) = \log(1) = 0 \tag{5-13}$$

is zero.

5-3 Frequency based estimation of parameters (Winkler)

Winkler [2000] proposes a slightly different way of estimating the m- and u-marginal probability mass functions based on information on the distribution of attributes in the populations. His method proposes an approach to deal with the lack of knowledge about the distribution of attributes in the population of true links. Winkler's approach uses

$$f_{\mathcal{M},q} = \begin{cases} \min(f_{\mathcal{A},q}, f_{\mathcal{B},q}) & \text{if } f_{\mathcal{A},q} > 1 \text{ or } f_{\mathcal{B},q} > 1 \\ \frac{2}{3} & \text{if } f_{\mathcal{A},q} = 1 \text{ and } f_{\mathcal{B},q} = 1. \end{cases} \tag{5-14}$$

for the number of times a comparison with label $c \in \{1, \ldots, Q\}$ is found in the population of true links. If a comparison attribute value occurs in both populations and at least more than once in one of both populations, then $f_{\mathcal{M},q}$ is the minimum frequency of this value for both populations. This implies that the number of links for attribute c is never more than $\min(f_{\mathcal{A},q}, f_{\mathcal{B},q})$. The idea is that one record with attribute c in a population (or dataset) can only link with one record with attribute c in another population (or dataset). So, there are never be more links with attribute c than the minimum number of occurrences in one of the datasets. If both datasets contain exactly one record with value c, then Winkler sets $f_{\mathcal{M},q}$ to $\frac{2}{3}$. He explains this as a $\frac{2}{3}$ chance of being a link for this record pair. This choice is based on experience and is not justified in Winkler [2000].

Consider now the same comparison cases as proposed by Fellegi and Sunter [1969].

Situation: The attribute in both records is v_c for which $q \in \{1, \ldots, Q\}$
The m-marginal probability mass function in Winkler's method is given by

$$m(q) = \frac{f_{\mathcal{M},q}}{N_{\mathcal{M}}}(1 - e_A)(1 - e_B)(1 - e_T)(1 - e_{A_0})(1 - e_{B_0}). \tag{5-15}$$

It does not differ from the method of Fellegi and Sunter [1969] in Section 5-2, only note that the function $f_{\mathcal{M},q}$ is defined by Formula 5-14. The u-marginal probability mass function differs from the method discussed in Section 5-2. The u-marginal probability mass function is

$$u(q) = \frac{f_{\mathcal{A},q} f_{\mathcal{B},q} - f_{\mathcal{M},q}}{N_{\mathcal{A}} N_{\mathcal{B}} - N_{\mathcal{M}}}(1 - e_A)(1 - e_B)(1 - e_T)(1 - e_{A_0})(1 - e_{B_0}). \tag{5-16}$$

The weight for this comparison value is given by

$$w(q) = \log\left(\frac{f_{\mathcal{M},q}}{f_{\mathcal{A},q} f_{\mathcal{B},q} - f_{\mathcal{M},q}} \, \frac{N_{\mathcal{A}} N_{\mathcal{B}} - N_{\mathcal{M}}}{N_{\mathcal{M}}}\right). \tag{5-17}$$

The quadratic scaling of the number of non-links is corrected with the linear scaling of the links (i.e. $f_{\mathcal{A},q} f_{\mathcal{B},q} - f_{\mathcal{M},q}$). Note that here is again implicitly assumed that the data is one-to-one linked. If the value c occurs once in both datasets, then $f_{\mathcal{A},q} f_{\mathcal{B},q} - f_{\mathcal{M},q} = \frac{1}{3}$. This explains why Winkler does not choose $f_{\mathcal{M},q} = 1$ in Formula 5-14. Due to the choice $f_{\mathcal{M},q} < 1$, the probability $u(q)$ never becomes zero, and there is still a chance on a non-link.

Situation: Situation: The attribute in both records disagree
The m-marginal probability mass function for disagreement is given by

$$m(0) = [1 - (1 - e_A)(1 - e_B)(1 - e_T)](1 - e_{A_0})(1 - e_{B_0}) \tag{5-18}$$

This formula is identical with the Fellegi and Sunter method. Again, the u-marginal probability mass function differs. The u-probability for disagreement is

$$u(0) = [1 - (1 - e_A)(1 - e_B)(1 - e_T) \sum_j \frac{f_{\mathcal{A},j} f_{\mathcal{B},j} - f_{\mathcal{M},j}}{N_{\mathcal{A}} N_{\mathcal{B}} - N_{\mathcal{M}}}](1 - e_{A_0})(1 - e_{B_0}) \tag{5-19}$$

and the weight of disagreement is given by

$$w(0) = \log \left(\frac{e_A + e_B + e_T}{1 - (1 - e_A - e_B - e_T) \sum_j \frac{f_{\mathcal{A}_j} f_{\mathcal{B}_j} - f_j}{N_{\mathcal{A}} N_{\mathcal{B}} - N_{\mathcal{M}}}} \right). \tag{5-20}$$

The correction for the number of non-links is used again in this weight.

Situation: At least one attribute in the pair of records is missing
The weight of a pair of records with at least one missing attribute (on the relevant characteristic) is identical with the weight in the method of Fellegi and Sunter, i.e. $w(-1) = 0$.

The proportions $f_{\mathcal{A},q}/N_{\mathcal{A}}$ and $f_{\mathcal{B},q}/N_{\mathcal{B}}$ for $q \in \{1, \ldots, Q\}$ can be estimated from the dataset. The number of links and the variables e_A, e_B and e_T stay unknown. Winkler [2000] proposes to use the ECM-algorithm to estimate $m(0)$ and π. Use this in combination with Formula 5-18 and set $m(0)$ equal to $e_A + e_B + e_T$. The link prevalence π can be used to estimate the number of links.

5-4 Frequency based estimation of parameters with the EM-algorithm

The methods in Section 5-2 and Section 5-3 have drawbacks. There is not a direct method to estimate the m- and u-marginal probability mass function because the parameters N_M and $e_A + e_B + e_T$ cannot be estimated directly from file characteristics. For the estimation method proposed by Fellegi and Sunter, the distribution of characteristics in the population of true links is also unknown. For this thesis, the EM-algorithm is used to estimate the parameters directly by adjusting the likelihood function. The method also assumes that the m- and u-marginal probability mass functions are conditional independent given the true link status (see Section 3-4-1).

The developed method has the possibility to distinguish different types of agreement but also disagreement. For example, comparing the date of birth $05 - 06 - 2000$ with $06 - 05 - 2000$ disagrees. However, the month and the day can be swapped, one of the goals is to assign this disagreement a different m- and u-marginal probability mass function than for comparing $05 - 06 - 2000$ with $01 - 10 - 1980$. This last comparison is obviously not identical.

Another (theoretical) advantage of this method is the possibility to estimate the m- and u-marginal probability mass function if one value is missing. Fellegi and Sunter and Winkler

argue that this weight is zero. The method in this thesis does not fix this to zero but estimates the m-and u-probabilities for this case. This is useful if the proportion of missing values is not equal for the dataset of true links and true non-links. For example, when there are relatively more missing values among the non-links, then it is more likely that a randomly picked pair of records with a missing value is a non-link. In fact, the method can distinguish different types of missing value comparisons.

Assume that attribute $i \in \{1, \ldots, K\}$ has a set of unique and distinct comparisons types V_i. Each comparison type for field i has an indicator $v_i \in V_i$. For each comparison is a m_i- and u_i-marginal probability mass function defined, where the m_i- and u_i-marginal probability mass functions are given by Formula 5-1 and Formula 5-2 respectively. The sum of all m_i-marginal probability mass functions

$$\sum_{v_i \in V_i} m_i(v_i) = 1 \tag{5-21}$$

and the sum of all u_i-marginal probability mass functions is

$$\sum_{v_i \in V_i} u_i(v_i) = 1. \tag{5-22}$$

For each comparison type $v \in V_i$, the probabilities $m_i(v)$ and $u_i(t)$ can be written as

$$m_i(y^i) = \prod_{v_i \in V_i} m_i(v_i)^{\mathbb{1}\{y^i = v_i\}} \tag{5-23}$$

and

$$u_i(y^i) = \prod_{v_i \in V_i} u_i(v_i)^{\mathbb{1}\{y^i = v_i\}} \tag{5-24}$$

respectively. Applied the conditional independence assumption (see Section 3-4-1) and to express the m- and u-probability mass functions as

$$m(\boldsymbol{y}) = \prod_{i=1}^{K} \prod_{v_i \in V_i} m_i(v_i)^{\mathbb{1}\{y^i = v_i\}} \tag{5-25}$$

and

$$u(\boldsymbol{y}) = \prod_{i=1}^{K} \prod_{v_i \in V_i} u_i(v_i)^{\mathbb{1}\{y^i = v_i\}} \tag{5-26}$$

respectively. If $V_i = \{0, 1\}$ for all fields, then the m- and u-probability mass functions are the same as used for ECM-algorithm (see Section 4-3). The ECM-algorithm is a special case of the more general EM-algorithm derived in this section.

The vector of parameters $\boldsymbol{\theta}_m$ contains all parameters for the problem. The parameters are the m-probabilities $m_i(v)$ for all fields $i \in \{1, \ldots, K\}$ and for all possible comparison types for that field $v \in V_i$. The same is done for the u-probabilities. They are found in the vector $\boldsymbol{\theta}_u$. The parameters for this problem are given by

$$\boldsymbol{\theta} = (\boldsymbol{\theta}_m, \boldsymbol{\theta}_u, \pi).$$

For clarity, the Expectation step for the EM-algorithm in the context of record linkage, Formula 4-8, is given again;

$$Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}) = \sum_{j=1}^{N} \mathbb{E}_{\boldsymbol{\theta}^{(t)}}\left[M_j|\boldsymbol{Y}_j = \boldsymbol{y}_j\right] \cdot \log(\pi \cdot m(\boldsymbol{y})) + \sum_{j=1}^{N} \left(1 - \mathbb{E}_{\boldsymbol{\theta}^{(t)}}\left[M_j|\boldsymbol{Y}_j = \boldsymbol{y}_j\right]\right) \cdot \log\left((1-\pi) \cdot u(\boldsymbol{y})\right).$$

for which the probabilities $P_{\boldsymbol{\theta}^{(t)}}(M = 1|\boldsymbol{Y} = \boldsymbol{y})$ and $P_{\boldsymbol{\theta}^{(t)}}(M = 0|\boldsymbol{Y} = \boldsymbol{y})$ are given by

$$P_{\boldsymbol{\theta}^{(t)}}(M = 1|\boldsymbol{Y} = \boldsymbol{y}) = \frac{\pi^{(t)} \cdot m^{(t)}(\boldsymbol{y})}{\pi^{(t)} \cdot m^{(t)}(\boldsymbol{y}) + (1 - \pi^{(t)}) \cdot u^{(t)}(\boldsymbol{y})} \tag{5-27}$$

and

$$P_{\boldsymbol{\theta}^{(t)}}(M = 0|\boldsymbol{Y} = \boldsymbol{y}) = \frac{(1 - \pi^{(t)}) \cdot u^{(t)}(\boldsymbol{y})}{\pi^{(t)} \cdot m^{(t)}(\boldsymbol{y}) + (1 - \pi^{(t)}) \cdot u^{(t)}(\boldsymbol{y})}. \tag{5-28}$$

The goal is now to maximise the conditional expectation of the complete data log-likelihood, $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)})$, with respect to parameters $\boldsymbol{\theta}$. To maximize this problem, there is a set of constraints that must be taken into account during the maximization. Consider a function $\boldsymbol{g} : \boldsymbol{\Theta} \to \mathbb{R}^{2K}$ for which $g_i(\boldsymbol{\theta})$ is given by

$$g_i(\boldsymbol{\theta}) = \sum_{v \in V_i} m_i(v) \tag{5-29}$$

and

$$g_{K+i}(\boldsymbol{\theta}) = \sum_{v \in V_i} u_i(t) \tag{5-30}$$

for comparison of attribute $i \in \{1, \ldots, K\}$.

The Maximization step can be formulated as

$$\begin{aligned} &\max\ Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}) \\ &\text{subject to } \boldsymbol{g}(\boldsymbol{\theta}) = 1. \end{aligned} \tag{5-31}$$

This maximization problem can be solved with the Lagrange multiplier method. The problem is one of solving

$$\nabla Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}) - \sum_{j=1}^{2K} \lambda_j \nabla \boldsymbol{g}_j(\boldsymbol{\theta}) = 0. \tag{5-32}$$

Consider comparison variable Y^i and comparison indicator $v \in V_i$, the partial derivation of Formula 5-32 with respect to $m_i(v)$ is

$$\frac{\partial Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)})}{\partial m_i(v)} - \sum_{j=1}^{2K} \lambda_j \frac{\partial \boldsymbol{g}_j(\boldsymbol{m}, \boldsymbol{u})}{\partial m_i(v)}. \tag{5-33}$$

This partial derivative is taken and set to 0 to maximise the problem,

$$\frac{\sum_{j=1}^{N} P_{\boldsymbol{\theta}^{(t)}}(M_j = 1|\boldsymbol{Y}_j = \boldsymbol{y}_j)\mathbb{1}\{y_j^i = t\}}{m_i(v)} - \lambda_j = 0. \tag{5-34}$$

All derivatives for parameters in $\boldsymbol{\theta}_m$ are of this form. For the parameters in $\boldsymbol{\theta}_u$, use λ_{K+i} instead of λ_i. The $m_i(v)$ probability is given by

$$m_i(v) = \frac{\sum_{j=1}^{N} P_{\boldsymbol{\theta}^{(t)}}(M_j = 1|\boldsymbol{Y}_j = \boldsymbol{y}_j)\mathbb{1}\{y_j^i = v\}}{\lambda_i}. \tag{5-35}$$

For field i, all the m-probabilities sum up to 1. It makes it possible to solve λ_i. The term λ_i is

$$\lambda_i = \sum_{v_i \in V_i} \sum_{j=1}^{N} P_{\boldsymbol{\theta}^{(t)}}(M_j = 1|\boldsymbol{Y}_j = \boldsymbol{y}_j)\mathbb{1}\{y_j^i = v_i\} \tag{5-36}$$

Because all indicator functions are non-overlapping and represent the entire set V_i, they sum up to 1. Therefore,

$$\lambda_i = \sum_{j=1}^{N} P_{\boldsymbol{\theta}^{(t)}}(M_j = 1|\boldsymbol{Y}_j = \boldsymbol{y}_j). \tag{5-37}$$

Substitute this formula into Formula 5-35, then the $m_i(v)$-probability function is

$$m_i^{(t)}(v) = \frac{\sum_{j=1}^{N} P_{\boldsymbol{\theta}^{(t)}}(M_j = 1|\boldsymbol{Y}_j = \boldsymbol{y}_j)\mathbb{1}\{y_j^i = v\}}{\sum_{j=1}^{N} P_{\boldsymbol{\theta}^{(t)}}(M_j = 1|\boldsymbol{Y}_j = \boldsymbol{y}_j)}. \tag{5-38}$$

Derive the u-marginal probability mass function for this estimation method in a similar way. It is given by

$$u_i^{(t)}(v) = \frac{\sum_{j=1}^{N} P_{\boldsymbol{\theta}^{(t)}}(M_j = 0|\boldsymbol{Y}_j = \boldsymbol{y}_j)\mathbb{1}\{y_j^i = v\}}{\sum_{j=1}^{N} P_{\boldsymbol{\theta}^{(t)}}(M_j = 0|\boldsymbol{Y}_j = \boldsymbol{y}_j)}. \tag{5-39}$$

The link prevalence π is similar with the link prevalence of the estimation methods in Chapter 4; i.e.

$$\pi^{(t+1)} = \frac{\sum_{j=1}^{N} P_{\boldsymbol{\theta}^{(t)}}(M_j = 0|\boldsymbol{Y}_j = \boldsymbol{y}_j)}{N}. \tag{5-40}$$

The only difference is that the m- and u-probability mass functions in $P_{\boldsymbol{\theta}^{(t)}}(M_j = 0|\boldsymbol{Y}_j = \boldsymbol{y}_j)$ are now given by Formula 5-25 and Formula 5-26. Note that the result of this parameter estimation method is closely related with the ECM-algorithm. In fact, it is a generalisation. This section does not contain a proof that the result is indeed a maximization of the problem (see Conclusion and Discussion, Chapter 9).

Part II

Simulations and applications

Chapter 6

Simulations

6-1 Introduction

In this chapter, a simulation study about probabilistic record linkage is presented. The goal of this simulation study is to explore the behaviour of the classification framework by Fellegi and Sunter and the estimation methods discussed in Part I. A simulation study is of interest, because the 'true' record linkage is not known for most record linkage applications. This means that there is no complete data available to train the classifier and to evaluate the results. The complete data in this simulation study is not used to train the classifier, but for analysis and evaluation of the classification results.

This simulation study focuses on the data quality, the classification method and the estimation methods. The following points are highlighted in this study:

- The Fellegi and Sunter classification framework (see Section 6-2).
- The number of errors made and the error levels μ and λ (see Section 6-2).
- The number of comparison variables versus the quality of the variables (see Section 6-3).
- The accuracy of the discussed Expectation-Maximisation algorithms (see Section 6-4).
- The convergence properties and starting parameters of the Expectation-Maximisation algorithms (see Section 6-4).
- The additional distinguishing power of estimates based on the distribution of characteristics in the population (see Section 6-4-2).
- The influence of missing data (see Section 6-5).

To study and evaluate the points above, data is needed for which the 'true' record linkage is known. A common simulation method in the literature is to sample data from existing datasets [Christen, 2005]. The sampled data is distributed over two datasets and error can be added to the datasets. These datasets are compared and indexation methods can be applied.

In this simulation study, a different simulation approach is chosen. In this approach, the

comparison vectors are simulated directly. Due to this choice, all steps of the record linkage workflow, described in Chapter 2, before the classification step are left out of scope. For each simulated set of data, N comparison vectors are simulated. Each comparison vector represents the comparison of a record pair. The number of comparison variables is K. A predefined distribution of comparison vectors for the true links is used to simulate N_M ($N \leq N$) comparison vectors. The remaining $N - N_M$ comparison vectors are simulated from a predefined distribution of comparison vectors for the true non-links. If not stated, the number of true links in a dataset is $N_M = 500$.

In this simulation study, the structure of some sets of simulated comparison vectors is used multiple times. For example, datasets with good or bad quality. For convenience, each type of datasets used in this simulation study is described and named below. The different sets of simulated comparison vectors are:

Good: Good data quality, binary and conditional independence assumption
For this set of comparison vectors, the comparison vectors represent the comparison of two good quality datasets of records. Good quality implies that the number of errors in the records is relatively small in both datasets. For this dataset is the set of parameters simulated under the conditional independence and binary assumption. Each parameter $m_1(1), \ldots, m_K(1)$ is a realisation of the uniform distribution $U(0.85, 0.99)$. The comparison vectors for the true links are simulated with these parameters. Each element $i \in \{1, \ldots, K\}$ in a comparison vector is a realisation of the Bernoulli distribution $Ber(m_i(1))$. The parameters $u_1(1), \ldots, u_K(1)$ are realisations of the distribution $U(0.02, 0.5)$. The $u_1(0), \ldots, u_K(0)$ are calculated from these parameters with Formula 3-30. The comparison vectors for the true non-links are simulated with these parameters. Each element $i \in \{1, \ldots, K\}$ in a comparison vector is a realisation of the Bernoulli distribution $Ber(u_i(1))$. The links prevalence is not part of a random process.

Low: Low data quality, binary and conditional independence assumption
This set of comparison vectors is simulated in the same way as the simulation of the 'good' dataset but now for data of low quality. This set of comparison vectors represents two datasets of records with plenty of errors. For the simulation of true link comparison vectors, the $m_i(1)$-probabilities are realisations of $U(0.7, 0.85)$. For the true non-links, the $u_i(1)$-probabilities are realisations of $U(0.02, 0.5)$.

Poor: Poor data quality, binary and conditional independence assumption
These comparison vectors represent the comparison of data of poor quality. Poor quality implies that there are many of errors in both datasets. The $m_i(1)$-probabilities are realisations of $U(0.6, 0.7)$ and the $u_i(1)$-probabilities are realisations of $U(0.02, 0.5)$.

SWOV: Skewed data quality, binary and conditional independence assumption
This set of comparison vectors is simulated in the same way as the previously mentioned dataset. The data has now two comparison variables with high quality, while the other variables are of poor quality. Comparison variables Y^1 and Y^2 for a true link comparison vector are realisations of $Ber(m_i(1))$ for which $m_i(1)$ was the realisation of $U(0.9, 0.99)$. The other comparison variables are simulated with m-probabilities drawn from $U(0.6, 0.8)$. For the true non-links, the $u_i(1)$-probabilities are realisations of $U(0.02, 0.5)$.

Frequency : Good data quality, conditional independence assumption

For this set of comparison vectors, not all the comparison variables are binary variables. The set is a set of good quality, for with the variable Y^1 can take values 20 different comparison types. For comparison variable Y^1 belonging to the linked set, the probability of a comparison with label $1, \ldots, 20$ is given by $\mathbf{p} = [p_1, \ldots, p_{20}]^T$ for which $\sum_{p \in \mathbf{p}} p = 1$. For the set of non-links, the probability on a comparison with label $1, \ldots, 20$ is given the probability vector $\mathbf{p} = [p_1^2/(\mathbf{p}^T\mathbf{p}), \ldots, p_{20}^2/(\mathbf{p}^T\mathbf{p})]^T$.

Dependent: Good data quality, dependencies, binary assumption
This set of comparison vectors is a set of good quality data. The difference for this set of comparison variables is that two variables are correlated with each other. There are several types of dependencies, like 2-way dependencies and 3-way dependencies. For this dataset, the comparison variables $\boldsymbol{Y}_1$ and $\boldsymbol{Y}_2$ for the set of comparison vectors belonging to the true non-links are correlated. The simulated data has a Pearson's correlation coefficient of $\rho(\boldsymbol{Y}_1, \boldsymbol{Y}_2) \approx 0.3$

6-2 The Fellegi and Sunter framework

In this section, the basic aspects of the record linkage framework by Fellegi and Sunter are discussed using simulations. The simulation study focuses on the performance of the classification framework and the associated error levels μ and λ. For the analysis in this section, a set of comparison vectors of 'good' quality is simulated. The set contains $N = 10^6$ comparison vectors, representing two files of 1000 records (and $N = 10^6$ record pairs). Each comparison vector contains $K = 8$ comparisons. Between both files, there are 500 record pairs representing the same entity, i.e. there are 500 true links and 999500 true non-links. For each comparison vector, the m- and u-probability mass functions are derived from the parameters used for simulation.

For each comparison vector, the weight (see Formula 3-42) is calculated with the parameters used for simulation of the dataset. Figure 6-1 shows the weights of all $N = 10^6$ record pairs. Note that it is a mixture of two distributions; the distribution of weights of the true links and the distribution of weights of the true non-links. In practice, the true link status is not known and both distributions are not distinguished. A large number of record pairs has a negative weight. Most of these comparison vectors represent true non-links. For most of the true non-links (999500 comparison vectors), it is likely that the u-probability is larger than the m-probability and, therefore, the weight is negative. The peak at the left in the histogram are the record pairs for which each comparison of the attribute disagrees. The 500 record pairs belonging to the same entity are almost not visible in this histogram, because they are overwhelmed by the 999500 non-links.

Most of the true links have positive weights. This is because it is likely that the m-probability of a comparison vector for a true link is larger than the u-probability of the vector. Figure 6-2 gives the weights for the same mixture of distributions as in Figure 6-1, but now for record pairs with weights equal or greater than 0. Observe that the 500 true links are overrepresented at the high(er) positive weights. Most of the record pairs do not contain errors and have the highest weights in the histogram. Making one error in the record generating process decreases the weight. Making more mistakes in the record generation decreases the weight even further. Making multiple mistakes in the record generation process is also less likely than making

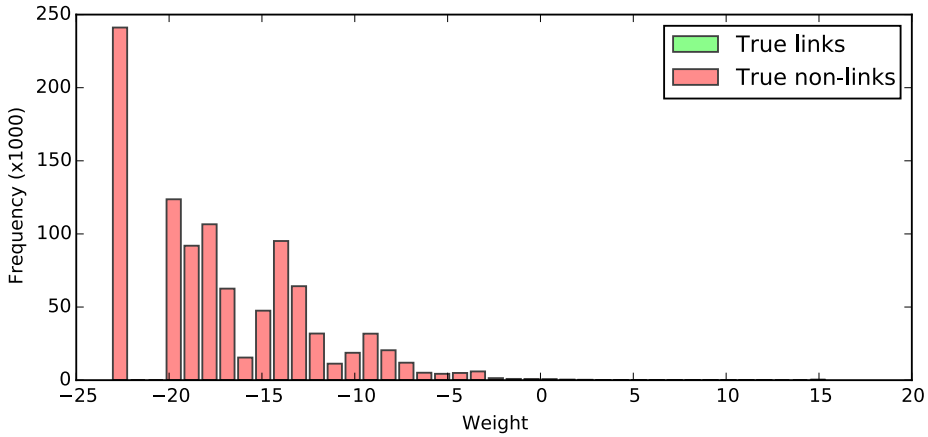


Figure 6-1: A histogram with the computed weights of $N = 10^6$ simulated comparison vectors. The 999500 true non-links overwhelm the 500 true links.

one mistake or no mistakes, therefore the distribution of weights for the true links has this shape.

For this simulated set of data, Table 6-1 shows information about the m- and u-probability mass functions for a part of the comparison space. Also, the weight w of the comparison vectors is given. The comparison vectors are sorted in a decreasing order of weight. Agreement/disagreement is labeled with 1/0. If all 8 comparisons agree, then this comparison vector has the highest weight. The comparison vector with the second highest weights, $\boldsymbol{y} = (1, 1, 1, 1, 1, 0, 1, 1)$, has one disagreeing comparison of the 6th field. This mistake on comparison variable Y^6 causes the least decrease in weight. If the comparison vector has more, or other, disagreements, the weight becomes less.

In Figure 6-2, it is clearly visible that the distributions of weights for the true links and true non-links overlap. The part of the histogram with weights between 4 and 9 shows

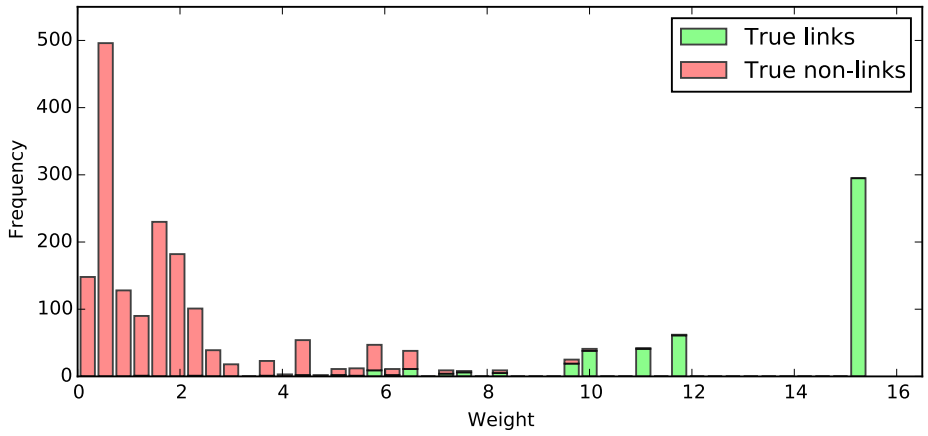


Figure 6-2: A subset of Figure 6-1 with comparison vectors with weight $w \geq 0$. Most of the true links have (relatively) large positive weights.

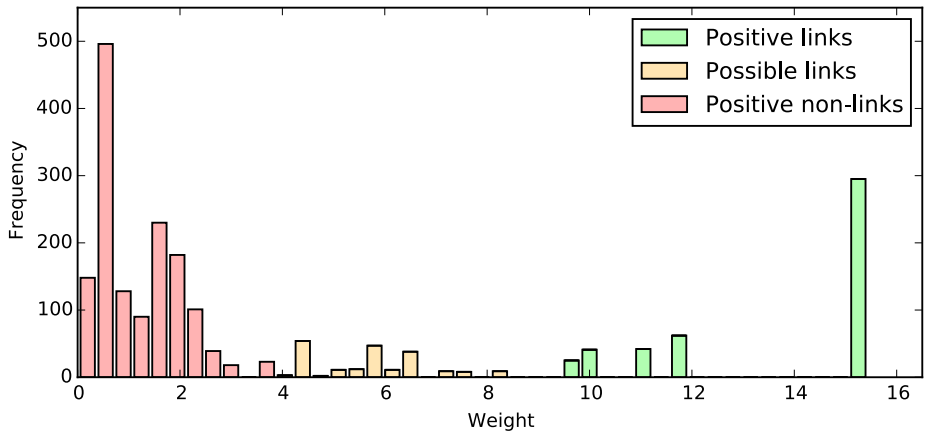


Figure 6-3: A histogram with the weight of comparison vectors with a positive weight. The set is divided into 3 actions based on manual classification; the positive links action, the positive non-link action and the possible link action.

clearly overlap between the distributions. In practice, it is not possible to distinguish both distributions, because the true link status is not known. Fellegi and Sunter divide the two distributions based on the weight into three (action-)sets; the positive link actionset, the positive non-link actionset and the possible link actionset. The framework uses two threshold weights to divide the set of comparison vectors into the three sets. The choice of the two threshold values is related to the error levels μ and λ.

For this simulated dataset, the thresholds are set based on the histogram with complete data. The comparison vectors with weights between 4 and 9 are assigned to the possible link actionset (actionset II). Most of the comparison vectors with weights higher than 9 are true links, this set of comparison vectors gets the positive link action (actionset I). Most comparison vectors with weights lower than 4 are non-links, they get the positive non-link action (actionset III). This example of a classification into three actionsets is displayed in Figure 6-3.

In Chapter 3, it was described how the error levels μ and λ relate to the classification into three actionsets. The two error levels in the Fellegi and Sunter framework are $\mu = \mathbb{E}[d_1(\boldsymbol{Y})|M = 0]$ and $\lambda = \mathbb{E}[d_3(\boldsymbol{Y})|M = 1]$. Error level μ is the probability of classifying a comparison vector with the positive link action while it is true non-link. This error level cuts the comparison space between the positive link action and the possible link action. Note that this may look counterintuitive. Error level λ is the probability on classifying a comparison vector as a positive non-link while it is true link. This error level cuts the comparison space between the positive non-link action and the possible link action. It is clear that with the classification described above, there are errors associated with the classification (see Figure 6-2).

In Table 6-1, the error level μ and λ are calculated based on the known m- and u-probability functions for each comparison vector $\boldsymbol{y} \in \Gamma$. The error levels are the errors for a non-random classification for which the comparison vector is included in actionset I and actionset III respectively. The error levels are calculated with the formula for non-random error levels (Formula 3-10). Note that $\boldsymbol{y} = (1, 1, 1, 1, 1, 1, 1, 1)$ has $\lambda = 1$ in Table 6-1. This violates with the non-randomised decision rule in which at least one vector is classified with action I. For

y^1	y^2	y^3	y^4	y^5	y^6	y^7	y^8	$w(\boldsymbol{y})$	$m(\boldsymbol{y})$	$u(\boldsymbol{y})$	λ	μ
1	1	1	1	1	1	1	1	15.43	6.08e-01	1.21e-07	1.00e+00	1.21e-07
1	1	1	1	1	0	1	1	11.83	5.42e-02	3.96e-07	3.92e-01	5.18e-07
1	1	1	0	1	1	1	1	11.68	6.91e-02	5.83e-07	3.37e-01	1.10e-06
1	0	1	1	1	1	1	1	11.09	5.16e-02	7.86e-07	2.68e-01	1.89e-06
0	1	1	1	1	1	1	1	10.97	3.14e-02	5.39e-07	2.17e-01	2.43e-06
1	1	1	1	1	1	1	0	10.11	6.32e-02	2.57e-06	1.85e-01	5.00e-06
1	1	1	1	1	1	0	1	9.98	6.71e-03	3.10e-07	1.22e-01	5.31e-06
1	1	0	1	1	1	1	1	9.63	3.00e-02	1.97e-06	1.15e-01	7.27e-06
1	1	1	1	0	1	1	1	9.56	8.71e-03	6.14e-07	8.53e-02	7.89e-06
1	1	1	0	1	0	1	1	8.08	6.16e-03	1.90e-06	7.66e-02	9.79e-06
1	0	1	1	1	0	1	1	7.49	4.60e-03	2.56e-06	7.05e-02	1.24e-05
0	1	1	1	1	0	1	1	7.38	2.80e-03	1.76e-06	6.59e-02	1.41e-05
1	0	1	0	1	1	1	1	7.35	5.86e-03	3.77e-06	6.31e-02	1.79e-05
...	...	...	...	...	...	...	...	...	...	...	...	...
1	0	1	1	1	1	0	1	5.65	5.68e-04	2.01e-06	2.03e-02	9.80e-05
0	1	1	1	1	1	0	1	5.53	3.47e-04	1.38e-06	1.97e-02	9.94e-05
1	0	0	1	1	1	1	1	5.30	2.55e-03	1.27e-05	1.94e-02	1.12e-04
1	0	1	1	0	1	1	1	5.22	7.38e-04	3.97e-06	1.68e-02	1.16e-04
0	1	0	1	1	1	1	1	5.18	1.55e-03	8.73e-06	1.61e-02	1.25e-04
0	1	1	1	0	1	1	1	5.11	4.50e-04	2.72e-06	1.45e-02	1.28e-04
1	1	1	1	1	1	0	0	4.67	6.97e-04	6.56e-06	1.41e-02	1.34e-04
1	1	0	1	1	1	1	0	4.32	3.12e-03	4.16e-05	1.34e-02	1.76e-04
1	1	1	1	0	1	1	0	4.24	9.05e-04	1.30e-05	1.03e-02	1.89e-04
1	1	0	1	1	1	0	1	4.19	3.31e-04	5.02e-06	9.37e-03	1.94e-04
1	1	1	1	0	1	0	1	4.11	9.60e-05	1.57e-06	9.04e-03	1.95e-04
1	1	0	1	0	1	1	1	3.77	4.30e-04	9.95e-06	8.94e-03	2.05e-04
1	0	1	0	1	0	1	1	3.75	5.22e-04	1.23e-05	8.51e-03	2.18e-04
0	1	1	0	1	0	1	1	3.63	3.18e-04	8.43e-06	7.99e-03	2.26e-04
0	0	1	1	1	0	1	1	3.04	2.38e-04	1.14e-05	7.67e-03	2.37e-04
...	...	...	...	...	...	...	...	...	...	...	...	...
0	1	0	1	0	0	0	0	-15.05	2.27e-09	7.77e-03	1.74e-08	3.30e-01
1	0	0	0	0	1	0	0	-15.07	4.74e-09	1.67e-02	1.51e-08	3.47e-01
0	1	0	0	0	1	0	0	-15.19	2.89e-09	1.14e-02	1.04e-08	3.58e-01
0	0	0	1	0	1	0	0	-15.78	2.16e-09	1.54e-02	7.46e-09	3.74e-01
0	0	0	0	1	0	0	0	-17.26	1.53e-09	4.78e-02	5.30e-09	4.21e-01
0	0	1	0	0	0	0	0	-17.33	4.43e-10	1.49e-02	3.78e-09	4.36e-01
0	0	0	0	0	0	1	0	-17.68	1.98e-09	9.46e-02	3.33e-09	5.31e-01
0	0	0	0	0	0	0	1	-17.81	2.10e-10	1.14e-02	1.35e-09	5.42e-01
1	0	0	0	0	0	0	0	-18.67	4.23e-10	5.45e-02	1.14e-09	5.97e-01
0	1	0	0	0	0	0	0	-18.79	2.58e-10	3.73e-02	7.18e-10	6.34e-01
0	0	0	1	0	0	0	0	-19.38	1.93e-10	5.03e-02	4.60e-10	6.84e-01
0	0	0	0	0	1	0	0	-19.53	2.45e-10	7.41e-02	2.67e-10	7.58e-01
0	0	0	0	0	0	0	0	-23.13	2.19e-11	2.42e-01	2.19e-11	1.00e+00

Table 6-1: A table with a part of the comparison space Γ for the simulated set of comparison vectors. The m- and u-probabilities are included, as well as the error levels. The comparison vectors in the comparison space are sorted in a decreasing order of weight.

clarity, in the table is the value displayed to show that they sum op to 1 (Assigning all $\boldsymbol{y} \in \Gamma$ to actionset III results in a probability of 1 for misclassification). The same arguments holds for the μ error level in combination with comparison vector $\boldsymbol{y} = (0, 0, 0, 0, 0, 0, 0, 0)$.

Return to the manual classification based on Figure 6-2. If the weights between $w(\boldsymbol{y}) = 4$ and $w(\boldsymbol{y}) = 9$ are classified as possible links, then the associated error levels can be found in Table 6-1. For this classification, $\lambda = 8.94 \cdot 10^{-3}$ and $\mu = 7.89 \cdot 10^{-6}$. Note that the error level μ is far less than the error level λ. This is not surprising, the number of true non-links (in this simulation) is much larger than the number of true links. Therefore, a few (in absolute sense) misclassified true non-links has a large influence on the error level on the positive link actionset I, while a few true links classified as positive non-links are overwhelmed by the large number of record pairs classified with the positive non-link status. There are 669 records classified as possible links. For these record pairs, a clerical review is an option for classification (See Section 2-8-2). Reducing the number of comparison vectors with the possible link action implies that there are fewer record pairs left for clerical review, but the error levels increase. Increasing the number of comparison vectors with the possible link action may imply fewer errors, but it results in more possible links.

So far, only an analysis of the classification framework was given with one simulated dataset. To get a comprehensive analysis, more datasets are simulated and classified. For each of the sets of comparison vectors 'good', 'low', and 'poor', described in Section 6-1, are 1000 datasets simulated. Each dataset contains $N = 10^6$ record pairs and 500 links. For each dataset, the comparison vectors are classified by assuming that the number of links in the dataset is known (500 links). It rarely occurs that comparison space can be split exactly such that there are 500 positive links. If this is not possible, there is a need to add random record pairs to the set of positive links to gain exactly 500 positive links. In this simulation, these random decisions are left out of scope. If it is not possible to get exactly 500 possible links without random decisions, the comparison vector for which this is needed is classified as possible link. The comparison vectors with higher weights are classified as positive links, while the others are classified as positive non-links.

For each dataset, the error levels μ and λ are calculated with Formula 3-10. These error levels are defined as μ_{sim} and λ_{sim} ('sim' from simulation). The error level are also calculated based on the complete data of the simulated dataset. For this complete data, the error levels are given by μ_{exact} and λ_{exact}. Also, the F_{score} is calculated based on the complete data (See Section 2-8-2). This score is a measure of the quality of the classification and therefore also about the quality of the dataset. In Table 6-2 are the results given for the mentioned variables. The mean and standard deviation are given, as well as the minimum and maximum value for each of the described variables.

Observe in Table 6-2 that, on average, 470 comparison vectors are classified as positive links. The number of positive links does not change a lot for the different datasets. The average number of possible links varies between 55 and 61. The standard deviation is of the same order as the mean. This means that the number of possible links fluctuates heavily. The data quality does not play a large role in the number of possible links. More about this in Section 6-3. Observe that the differences between μ_{sim} and μ_{exact} are minimal. Also, the differences between λ_{sim} and λ_{exact} are minimal. This indicates the error levels formulated by Fellegi and Sunter are valid representations of the errors found with simulations. Note that the data quality influences the number of misclassifications. Better data quality results in

Dataset		I	II	III	μ_{exact}	μ_{sim}	λ_{exact}	λ_{sim}	F_{score}
Good	mean	471.75	60.22	999468.03	7.81e-05	7.87e-05	1.76e-01	1.76e-01	0.824
	std	30.07	57.67	43.86	4.24e-05	4.29e-05	9.78e-02	9.76e-02	0.099
	min	327.00	0.00	999051.00	8.00e-06	1.01e-05	1.80e-02	2.01e-02	0.441
	max	500.00	471.00	999500.00	2.54e-04	2.68e-04	5.56e-01	5.34e-01	0.983
Low	mean	469.69	60.87	999469.43	2.22e-04	2.23e-04	4.85e-01	4.85e-01	0.512
	std	37.70	64.43	42.81	5.84e-05	5.99e-05	1.44e-01	1.43e-01	0.148
	min	229.00	0.00	999038.00	7.50e-05	8.51e-05	1.50e-01	1.62e-01	0.124
	max	500.00	557.00	999500.00	4.07e-04	4.17e-04	8.58e-01	8.63e-01	0.850
Poor	mean	473.06	55.24	999471.70	3.32e-04	3.33e-04	7.07e-01	7.07e-01	0.289
	std	33.92	52.50	35.41	5.32e-05	5.59e-05	1.25e-01	1.24e-01	0.128
	min	242.00	0.00	999142.00	1.67e-04	1.73e-04	3.44e-01	3.62e-01	0.020
	max	500.00	442.00	999500.00	4.70e-04	4.87e-04	9.62e-01	9.56e-01	0.658

Table 6-2: This table contains the results 1000 classifications and error levels for the datasets 'good', 'low' and 'poor'.

lower error probabilities (see Section 6-3). The number of true non-links classified as positive links is estimated on $0.176 * 500 = 88$ for the dataset of good quality. For the dataset of poor quality, the number of non-links classified as positive links is estimated on $0.707 * 500 = 353.5$ which is quite bad in contrast to the good dataset.

6-3 Comparison variables and data quality

In the previous section, a closer look was given to the classification framework. In this section, the focus is on the data to classify. In particular, the number of variables available for comparison and the quality of the data is examined. The Fellegi and Sunter framework can be used with incomplete or incorrect data. The quality of the data depends on the number of missing or incorrect values. Poor data quality obviously has an influence on the classification process.

Each comparison variable is of (different) importance in the classification. This is because of the distinguishing power of the variable. Some variables are very useful to identify the entity while others are not. For example, the hair colour of a person does not identify someone while a personal identifier does. An aggregation of *quasi-identifiers* (variables such as hair colour) can be used to identify a person. Think about variables such as hair colour, (sur)names, sex, place of birth and date of birth. In the framework of Fellegi and Sunter, the distinguishing power of a variable is kept in the u-marginal probability mass function. When the u-marginal probability mass function is zero for a comparison variable (assume the conditional independence assumption), then there is no agreeing comparison in the set of non-links. This case of u equal to zero is found for (unique) personal identifiers in deduplicated datasets. For quasi-identifiers, the u-probability is not close to 0. For example, the comparison variable sex often has a u-marginal probability mass function of $\frac{1}{2}$ when linking (census) records. The reason for this: there is a 1/4 probability of linking when the attribute is 'male' and 1/4 when the attribute 'female'. This means that about half of the comparisons on the sex in the set of non-links agrees on this value.

The data quality plays an important role in the classification process. This can be explained with a small example. Consider $K = 8$ comparison variables for which the conditional independence assumption holds. Both simulated datasets do not contain errors. This lack or errors means that a record pair representing a true link always has the comparison vector $\boldsymbol{y} = (1, 1, 1, 1, 1, 1, 1, 1)$. This implies that the distribution of weights for the true links is a degenerate distribution. The following table with m- and u-marginal probability mass functions describes this problem;

	y^1	y^2	y^3	y^4	y^5	y^6	y^7	y^8
$m_i(1)$	1	1	1	1	1	1	1	1
$u_i(1)$	0.1	0.3	0.2	0.15	0.5	0.05	0.1	0.1
$w_i(1)$	2.30	1.20	1.61	1.89	0.69	3.00	2.30	2.30
$w_i(0)$	$-\infty$	$-\infty$	$-\infty$	$-\infty$	$-\infty$	$-\infty$	$-\infty$	$-\infty$

A mistake always results in a $-\infty$ weight. So a pair of records with a mistake is never assigned as a positive link. In general, there are errors in the data whereby the weight for disagreement does not become $-\infty$. The fraction of errors in the data influences the weight of disagreement $w(0)$. Small amounts of errors in comparison field i results in relatively low disagreement weights $w_i(0)$. Therefore, better quality data forces the distribution of weights of the true non-links to the negative axis. Therefore, the distributions can be better distinguished in the Fellegi and Sunter model.

The number of variables plays a role to distinguish the two distributions besides the distinguishing power of the variables and the data quality. Using more quasi-identifiers (names, sex, address information) can result in a good classification. The general idea is that using more comparison variables is always better for the classification. To indicate the role of the number of comparison variables, three batches of 1000 datasets of type 'low' are simulated. Each time, a different number of comparison variables K is used. The number of comparison variables are 6, 8 or 10. In Table 6-3 are the results given. Observe that the error levels μ_{sim} and μ_{exact} are nearly identical for each of the cases, as well as λ_{sim} and λ_{exact}. The number of comparison variables is inversely related to the error probabilities. Using more comparison variables results in fewer errors and vise verse. Also, note that the F_{score} indicates a better classification when more variables are used. Observe in Table 6-3 that the number comparison variables is strongly related to the number of possible links. This relation is because the number of possible comparison vectors in Γ depends on the number of variables. The variables are spread over the comparison space and therefore there are fewer comparison vectors (of the 10^6) with the possible link action, action II, when more variables are used.

In Appendix B, a similar table is given for more types of datasets mentioned in Section 6-1. Observe that the data quality plays an important role in the classification. Datasets of type 'low' with $K = 10$ comparison variables have on average more misclassifications than datasets with $K = 6$ comparison variables of type 'good'. Although it is not possible to give a relation between the data quality and the number of variables, it is clear that the data quality plays an important role in the classification. The 'SWOV' datasets, datasets with two comparison variables of good quality while the others are of poor quality, are comparable with the 'low' quality datasets.

K		I	II	III	μ_{exact}	μ_{sim}	λ_{exact}	λ_{sim}	F_{score}
6	mean	440.17	167.41	999392.42	9.96e-05	9.98e-05	2.51e-01	2.50e-01	0.751
	std	42.60	106.60	109.98	4.64e-05	4.72e-05	9.20e-02	9.03e-02	0.092
	min	318.00	0.00	998469.00	1.10e-05	1.29e-05	3.40e-02	3.39e-02	0.481
	max	500.00	1035.00	999500.00	2.52e-04	2.73e-04	5.20e-01	4.76e-01	0.965
8	mean	471.75	60.22	999468.03	7.81e-05	7.87e-05	1.76e-01	1.76e-01	0.824
	std	30.07	57.67	43.86	4.24e-05	4.29e-05	9.78e-02	9.76e-02	0.099
	min	327.00	0.00	999051.00	8.00e-06	1.01e-05	1.80e-02	2.01e-02	0.441
	max	500.00	471.00	999500.00	2.54e-04	2.68e-04	5.56e-01	5.34e-01	0.983
10	mean	494.72	10.50	999494.79	4.65e-05	4.54e-05	9.64e-02	9.38e-02	0.904
	std	9.87	16.57	10.24	3.05e-05	3.19e-05	6.45e-02	6.59e-02	0.065
	min	411.00	0.00	999391.00	3.00e-06	9.93e-07	6.00e-03	2.48e-03	0.453
	max	500.00	129.00	999500.00	2.48e-04	2.62e-04	5.34e-01	5.37e-01	0.994

Table 6-3: This table gives the result of 1000 classifications with 'low' quality comparison variables. The column K is the number of comparison variables used for classification. Observe that the quality of the classification is better when there are more variables used (see the F-score).

6-4 Estimation methods

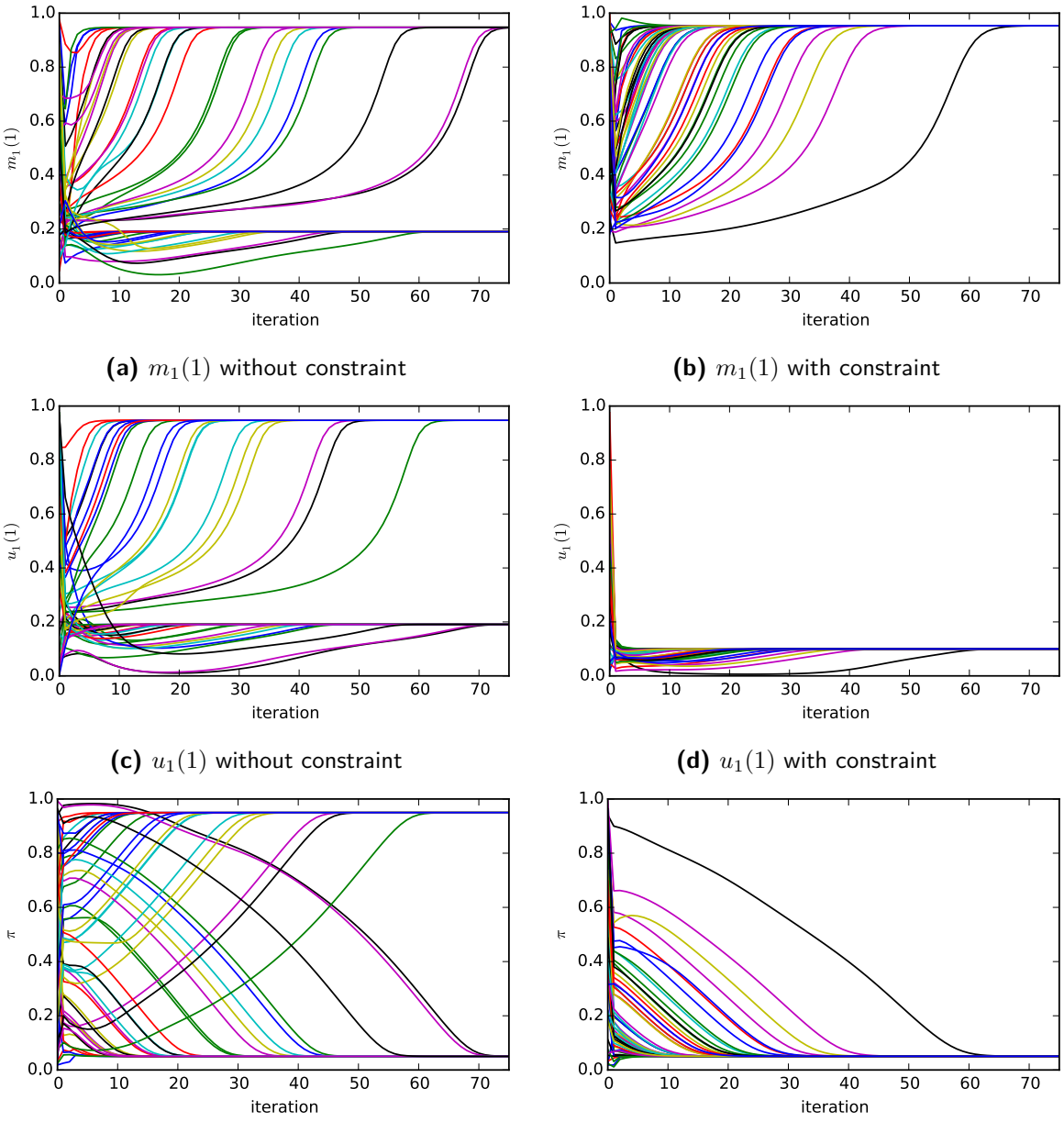
The m- and u-probability mass functions and the link prevalence π need to be estimated if there is no complete data available. Several estimation methods are discussed in Chapter 4 and Chapter 5. All estimation methods are tested and implemented. For this simulation study, the three most promising estimation methods are used to evaluate with simulations. The used methods are the ECM-algorithm described in Section 4-3, the algorithm proposed by Schürle described in Section 4-4 and the EM-algorithm used for frequency based estimates developed for this thesis. The focus is on the accurateness of the classification, the convergence properties and the starting parameters.

6-4-1 Estimation of parameters with the ECM-algorithm

The ECM-algorithm is used to estimate parameters of interest in the Fellegi and Sunter model when the data is assumed to be conditional independent given the true link status (See Section 4-3). In Section 4-3 was mentioned that the accuracy of the estimates and the converging properties of the ECM-algorithm are known to be good. The algorithm converges to a stationary point. In this section, the choice of starting parameters is discussed first.

Consider a simulated dataset of $N = 10^6$ comparison vectors for which 500 comparison vectors represent the same entity. Each comparison vector is of length 8 and the conditional independence assumption and binary assumption holds. This implies that there are 17 parameters interesting in this model; $m_1(1), \ldots, m_8(1)$,$u_1(1), \ldots, u_8(1)$ and the link prevalence π. Each of these parameters needs a starting value to start the ECM-algorithm. For this simulation, the ECM-algorithm is applied 100 times to the same dataset with different starting values. All the 17 starting values are chosen randomly between 0 and 1. In Figure 6-4 are the parameters $m_1(1)$, $u_1(1)$ and π given for each iteration and for each of the 100 random sets of starting points.

From the Figures, it is clear that ECM-algorithm tends to converge. The ECM-algorithm



(a) $m_1(1)$ without constraint

(b) $m_1(1)$ with constraint

(c) $u_1(1)$ without constraint

(d) $u_1(1)$ with constraint

(e) π without constraint

(f) π with constraint

Figure 6-4: The convergence behaviour of the ECM-algorithm for different starting points. The figures on the left show the convergence behaviour of $m_1(1)$, $u_1(1)$ and π for random starting values between 0 and 1. The figures on the right show the convergence behaviour of $m_1(1)$, $u_1(1)$ and π for random starting values between 0 and 1 with the restriction that $m_i(1) > u_i(1)$ for all $i \in \{1, \ldots, K\}$.

converges in a few iterations, up to 75 iterations. There are two stationary points to which each of the estimates of $m_1(1), u_1(1), \pi$ converges. The same convergence behaviour is noticed for $m_2(1)$, ..., $m_8(1)$ and $u_2(1)$, ..., $u_8(1)$. For some of the starting values, the parameters estimates converge to the likely values of $m_1(1), \ldots, m_8(1), u_1(1), \ldots, u_8(1), \pi$. Other parameter estimates converge to clearly incorrect values. For the parameters that do not converge to the correct value, the algorithm converges to the incorrect true link status M. The $m_i(1)$-probability estimates converge to the $u_i(1)$-probability estimates and the $u_i(1)$-probability estimates converge to the $m_i(1)$-probabilities estimates. The link prevalence π converges to $(1-\pi)$, i.e. it converges to $P(M=0)$ instead of $P(M=1)$. This behaviour can be explained with Formula 4-8 and Formula's 4-9 and 4-10. If the m and u probability mass functions are swapped and the link prevalence π is replaced by $1-\pi$, then Formula 4-8 remains the same. This implies that the formula for the Expectation step of the EM-algorithm, Formula 4-8, is maximised for both situations.

The goal is to choose the starting values such that the algorithm converges to the desired parameter estimates. Overall, identical comparisons occur often in the true link set. This means that the m-marginal probability mass functions for agreement should be large probabilities. In the true non-link set, there is relatively much less disagreement in the comparison vectors. The u-probabilities mass functions for agreement should be low probabilities. For this reason, an arbitrary choice of the starting values is not recommended. Restricting the starting m- and u-marginal probabilities to $m > u$ prevents that estimates converge to the wrong convergence point. See Figure 6-4 for the same simulated data, but now with this restriction to the starting points. For all used starting values, the parameter estimates converge to the same stationary point. This stationary point is the 'correct' stationary point. As discussed in Section 4-3-1, a good starting value for the $m_i(1)$-probabilities is 0.9 and for the $m_i(0)$-probabilities 0.1. This satisfies the mentioned condition.

So far, it is observed that the ECM-algorithm convergences to a stationary point. Now, the accuracy of the classification with the algorithm is studied. Consider 1000 datasets of type 'good' with $N = 10^6$ comparison vectors and $K = 8$ variables. Each set contains 500 true links. For each of the vectors is ECM-algorithm applied with starting values 0.9 for agreeing m-marginal probability mass functions, 0.1 for agreeing u-marginal probability mass functions and 0.01 for the link prevalence. The iterative ECM-algorithm is applied until the algorithm converged. In Table 6-4 are the results of the classification and estimation presented. The table has the same structure as the table used in Section 6-2 and Section 6-3. The only difference is now that the number of estimated links N_M is given, i.e. the link prevalence π multiplied by the number of comparison vectors. The average number of estimated links N_M is 13728.70. This estimation is by far not close to the 500 true links in the set of comparison vectors. Note that the 25% percentile gives reasonable results, but there are also a lot of simulated datasets for which the algorithm did not work well. The same process is redone with sets of $N = 10^4$ and $N = 10^5$ comparison vectors with 500 true links in it. For $N = 10^5$ comparison vectors, the number of times the algorithm worked well is much larger. The 95% percentile is still a reasonable estimate. For $N = 10^4$ comparison vectors, all estimates seem to be reasonable. The mean of N_M for all datasets is 500.90, which is very accurate. The standard deviation is 6.97, which is relatively low.

Observe that the estimation of the average error level λ is not very accurate for the datasets with $N = 10^5$ and $N = 10^6$ comparison vectors. The error level λ is conditioned on the true

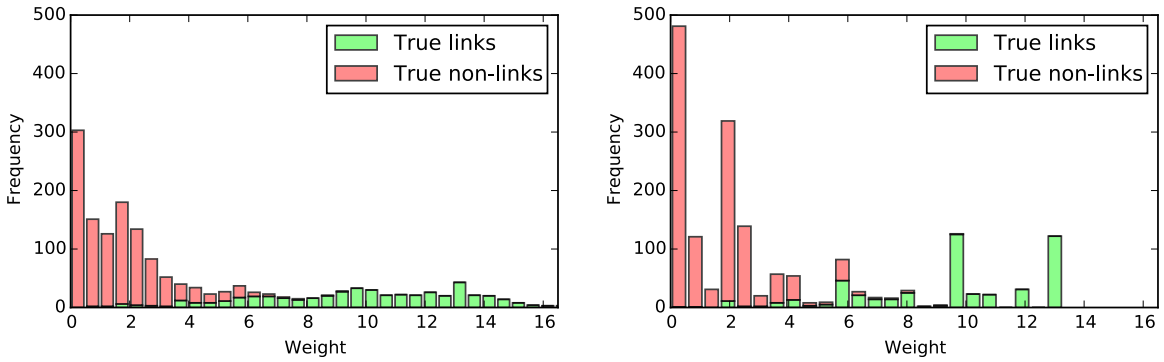
N		I	II	III	N_M	μ_{exact}	μ_{sim}	λ_{exact}	λ_{sim}
10^6	mean	13439.95	566.52	985993.53	13728.70	1.30e-02	1.22e-02	4.10e-02	5.08e-01
	std	19479.48	915.35	20204.19	19855.86	1.95e-02	1.83e-02	4.50e-02	3.40e-01
	min	422.00	0.00	862162.00	460.00	5.00e-06	9.85e-06	0.00e+00	1.71e-02
	5%	477.00	2.00	944379.75	488.00	2.50e-05	2.51e-05	0.00e+00	5.46e-02
	25%	503.00	17.00	979553.25	510.00	5.40e-05	5.13e-05	6.00e-03	1.07e-01
	50%	4322.50	187.00	995433.50	4384.00	3.83e-03	3.66e-03	1.60e-02	6.78e-01
	75%	19715.50	767.00	999482.00	20311.25	1.92e-02	1.83e-02	7.40e-02	8.30e-01
	95%	52780.20	2293.00	999508.05	54003.75	5.23e-02	4.93e-02	1.34e-01	8.67e-01
	max	136577.00	8438.00	999532.00	136718.00	1.36e-01	1.26e-01	1.74e-01	8.87e-01
10^5	mean	616.89	18.18	99364.94	626.31	1.60e-03	1.46e-03	7.50e-02	9.33e-02
	std	853.17	32.20	874.85	862.30	8.53e-03	7.43e-03	4.31e-02	1.02e-01
	min	392.00	0.00	88016.00	446.00	3.02e-05	2.36e-05	0.00e+00	7.13e-03
	5%	464.90	0.00	99397.95	481.00	1.01e-04	1.16e-04	1.60e-02	2.15e-02
	25%	488.00	3.00	99484.00	494.00	2.11e-04	2.19e-04	4.20e-02	4.42e-02
	50%	497.00	9.00	99495.00	501.00	3.42e-04	3.47e-04	6.80e-02	6.95e-02
	75%	505.00	17.00	99502.00	509.00	5.33e-04	5.15e-04	1.00e-01	1.03e-01
	95%	535.05	80.00	99514.00	560.20	1.20e-03	1.14e-03	1.62e-01	2.13e-01
	max	11696.00	288.00	99554.00	11885.00	1.13e-01	9.65e-02	2.40e-01	7.01e-01
10^4	mean	499.21	3.53	9497.26	500.90	1.73e-03	1.70e-03	3.14e-02	3.10e-02
	std	7.18	4.46	7.89	6.97	1.29e-03	1.18e-03	2.32e-02	2.24e-02
	min	463.00	0.00	9437.00	473.00	0.00e+00	1.60e-04	0.00e+00	1.70e-03
	5%	488.00	0.00	9484.00	491.00	3.16e-04	4.47e-04	6.00e-03	7.13e-03
	25%	496.00	0.00	9494.00	497.00	8.42e-04	8.82e-04	1.60e-02	1.57e-02
	50%	499.00	2.00	9498.00	501.00	1.37e-03	1.42e-03	2.60e-02	2.57e-02
	75%	503.00	6.00	9501.00	504.00	2.21e-03	2.14e-03	4.00e-02	3.97e-02
	95%	510.00	13.00	9508.00	512.00	4.22e-03	4.03e-03	7.61e-02	7.56e-02
	max	537.00	28.00	9523.00	538.00	9.79e-03	9.72e-03	1.66e-01	1.92e-01

Table 6-4: A table containing information about classifications and error levels. For $N = 10^4$, $N = 10^5$ and $N = 10^6$ are 1000 datasets simulated. For each set is the ECM-algorithm applied to estimate parameters and classify the record pairs. N_M is the estimated number of links ($N \times \pi$).

link status. The number of links is totally incorrect, so there are many true non-links found in this set. The m-probability mass functions are not correctly estimated and therefore, the error level estimates of λ are also incorrect. The effect of incorrect parameter estimates for a large number of record pairs was described by Yancey [2002]. If the proportion of true links (of all record pairs) drops below 0.05, then the parameter estimates may be not relevant for the record linkage problem. This proportion is observed in the simulations in this thesis. The simulation study in this thesis shows that the proportion of 0.05 is chosen quite strict. Even for a proportion of 0.005 (the set with $N = 10^5$ comparison vectors), most of the time are the parameter estimates relevant for the record linkage problem. In the paper of Yancey [2002], a method is described that selects a subset of record pairs to enrich the dataset and to adjust the proportion. This reduction of record pairs can be done with indexing.

6-4-2 Estimation of frequency based parameters with the EM-algorithm

The EM-algorithm proposed in Section 5-4 is a generalisation of the ECM-algorithm. This algorithm was developed to extend the ECM-algorithm with additional distinguishing power. Adding additional power can be done by distinguishing different types of agreement, such as



(a) Histogram of weights based on frequency based parameters.

(b) Histogram of weights under the binary assumption.

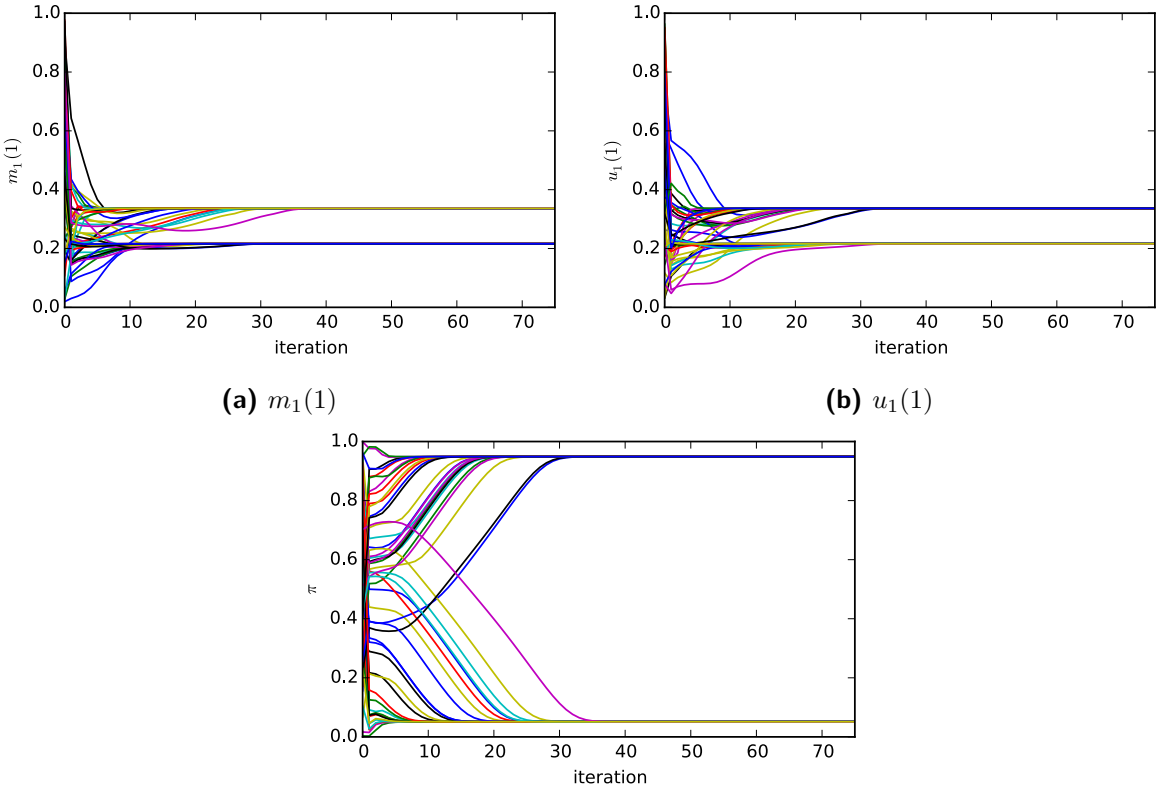
Figure 6-5: Two histograms with the weights of the comparison vectors. The right histogram shows the weights when the classification is performed under the binary assumption. The left histogram uses additional distinguishing power.

agreement and the attribute in both records is '...'. In Section 5-4 was described that also other types of comparisons can be used for distinguishing power. In this section, this is left out of scope to avoid ambiguity. In Section 6-5 is such a special type of comparison discussed with simulations.

The first step is to analyse the role of the binary assumption. Does this assumption influence the distinguishing power and if it does: how strong is the influence? To analyse this; consider a simulated dataset of $N = 10^6$ comparison vectors of type 'freq' with 8 comparison variables. The set contains 500 true links. For this simulated dataset, the m- and u-marginal probability mass functions used for simulation of the dataset are known. In Figure 6-5a, the weights of the comparison vectors are displayed in a histogram. Only weights larger than 3 are displayed because the non-links overwhelm the histogram. Note that the distribution for the true links is smoother than seen before under the binary assumption in Section 6-2.

To compare the behaviour with the binary assumption, the same dataset is converted into a dataset under the binary assumption. This implies that all types of agreement are converted into agreement. The remaining comparisons are disagreements. In Figure 6-5b is for the same dataset the binary assumption applied. Note the differences between Figure 6-5b and Figure 6-5a. The histogram is much smoother without the binary assumption. This is because there are much more comparison vectors found in the comparison space without the binary assumption. On first sight, the distribution of weights for the true links and true non-links in Figure 6-5a are not easier to distinguish than for Figure 6-5b.

The choice of starting parameters for the EM-algorithm is one of interest. In the ECM-algorithm and the algorithm by Schürle, the starting values are chosen under assumption of binary comparison vectors and conditional independence. For this application of the EM-algorithm, this can not be done, because there are multiple types of comparisons distinguished. There are several options to choose starting values. In Figure 6-6 are the starting values randomly chosen between 0 and 1. Notice the same convergence behaviour is observed as in Section 6-4-1. Some starting values converge to the incorrect true link status and, therefore, the m-marginal probability mass functions and u-marginal probabilities are interchanged.



(a) $m_1(1)$ **(b)** $u_1(1)$

(c) π

Figure 6-6: The convergence behaviour of the EM-algorithm for frequency based estimates for different starting points. The figures show the convergence behaviour of $m_1(1)$, $u_1(1)$ and π for random starting values between 0 and 1.

Also, the true link status π becomes $1 - \pi$ for some starting values.

A simple statement like $m_i(1) > u_i(1)$ for all $i \in \{1, \ldots, K\}$ as seen in Section 6-4-1 does not work. The starting values can be chosen based on file characteristics such a the distributions of comparison types. The easiest way is to choose starting values based on file characteristics and compute the frequency based application of the EM-algorithm. If the link prevalence is unlikely high, then the starting values are very likely to be incorrect.

In Figure 6-5a was seen that the two distributions are not directly better distinguishable than with the binary assumption. Table 6-5 represents the classification of 1000 simulated datasets of type 'freq'. Each set contains $N = 10^4$ comparison vectors and $K = 8$ comparison variables. For each dataset, a classification is performed with the parameters used to simulate the dataset. Also, the frequency based application of the EM-algorithm is applied for the classification. The results show that the EM-algorithm estimates the parameters well. This shows that the algorithm can work for the estimation of parameters.

The ECM-algorithm is applied to the same data with the binary assumption. The data is also classified with the parameters used for simulation. The algorithms show similar results for the number of links N_M and the error levels μ and λ. For the frequency based EM-algorithm,

Algorithm		I	II	III	N_M	μ_{exact}	μ_{sim}	λ_{exact}	λ_{sim}	F_{score}
Binary	mean	493.70	13.04	9493.25	500.00	3.67e-03	3.76e-03	7.22e-02	7.29e-02	0.928
	std	5.79	8.87	6.46	0.00	1.44e-03	1.46e-03	2.83e-02	2.79e-02	0.028
	min	459.00	0.00	9463.00	500.00	3.16e-04	4.49e-04	8.00e-03	1.22e-02	0.828
	max	500.00	57.00	9500.00	500.00	9.05e-03	9.18e-03	1.72e-01	1.62e-01	0.993
ECM	mean	493.77	12.81	9493.42	499.94	3.71e-03	3.71e-03	7.33e-02	7.17e-02	0.927
	std	13.60	8.90	14.42	12.48	1.62e-03	1.39e-03	3.06e-02	2.77e-02	0.028
	min	422.00	0.00	9415.00	448.00	3.16e-04	9.51e-04	8.00e-03	7.94e-03	0.838
	max	563.00	57.00	9541.00	572.00	1.09e-02	9.31e-03	1.88e-01	1.73e-01	0.992
Frequency	mean	498.90	2.40	9498.70	500.00	3.70e-03	3.79e-03	7.06e-02	7.09e-02	0.929
	std	2.76	5.34	3.40	0.00	1.42e-03	1.45e-03	2.69e-02	2.65e-02	0.027
	min	474.00	0.00	9467.00	500.00	4.21e-04	5.12e-04	8.00e-03	1.21e-02	0.829
	max	500.00	34.00	9500.00	500.00	8.95e-03	8.89e-03	1.72e-01	1.49e-01	0.992
Frequency based EM	mean	498.01	2.40	9499.59	499.08	3.76e-03	3.64e-03	7.38e-02	6.79e-02	0.927
	std	11.93	5.30	12.50	11.81	1.57e-03	1.35e-03	3.01e-02	2.52e-02	0.028
	min	446.00	0.00	9444.00	446.00	3.16e-04	7.73e-04	8.00e-03	8.04e-03	0.841
	max	549.00	36.00	9554.00	554.00	9.79e-03	8.39e-03	1.96e-01	1.45e-01	0.992

Table 6-5: A table containing information about classifications and error levels. For each type of algorithm are 1000 datasets simulated and classified. The situation 'binary' and 'frequency' are classifications based on the parameters used for simulation of the comparison vectors.

the number of comparison vectors with action II is less than for the ECM-algorithm. This is because of the number of elements (the comparison vectors) in the comparison space is much larger. Most of the comparison vectors do not occur many times in the set of all comparison vectors. This is the reason that the actionset II is smaller for the frequency based EM-algorithm. It should be mentioned that error levels may be related to the size of the actionset II. If the size of actionset II is small, more comparison vectors are classified into actionset I or actionset III and the risk on misclassifications becomes larger. Because the error levels are nearly identical for both algorithms, it may indicate that the frequency based EM-algorithm makes fewer misclassifications than the ECM-algorithm.

It is important to realise that the ECM-algorithm and the EM-algorithm for frequency based estimates estimate the comparison vectors for more vectors than may occur in the comparison space. This can be illustrated with an example, consider $\{(1,1),(1,0),(0,0)\} \in \Gamma$. The ECM-algorithm estimates the parameters $m_1(1)$, $m_2(1)$, $u_1(1)$, $u_2(1)$ which describe the set of comparison vectors $\{(1,1),(0,1),(1,0),(0,0)\}$. This is important for the estimation of error levels, because the missing comparison vector $(0,1)$ needs to be included in the calculation of error levels. For the ECM-algorithm, this does not often occur because the number of elements in the comparison space is not large. Therefore, most of the vectors are found in the set of comparison vectors. For the EM-algorithm for frequency based estimates, the number of elements in the comparison space is large and not all comparison vectors need to be observed. Therefore, this should be taken into account in the computation of the error levels.

6-4-3 Estimation of parameters with the algorithm of Schürle

In Section 4-4 was the estimation method proposed by Schürle described. The algorithm is developed to improve the classification of records pairs for which conditional dependencies

are found between the comparison variables. This algorithm is an EM-algorithm for which the likelihood can take into account dependencies between the comparison variables given the true link status. In Herzog, Scheuren and Winkler [2007] was described that the ECM-algorithm may result in good estimates when using it on a dataset with comparison vectors with dependencies. The algorithm has the unique property that it converges in one step.

In this section, 1000 'dependent' datasets are simulated. Each set contains $N = 10^4$ comparison vectors with $K = 8$ comparison variables. 500 comparison vectors are true links. First, the ECM-algorithm is applied to see how the classification performs if the conditional independence assumption is applied while that is not representative for the data. In Table 6-6, the results of the classification with the ECM-algorithm given. The average Pearson correlation between the two correlated comparison variables for 1000 datasets is 0.192. The estimated number of links N_M is 519.46. This estimated number of links is higher than the the number of true links in the data. Observe that the 95% quantile is still a reasonable estimate. The maximum number of estimated links for one of the datasets was 1493. This seems to indicate that the ECM-algorithm does not always converge to the correct value when there are dependencies in the data. The estimated error levels μ_{sim} and λ_{sim} differ from the simulated error levels μ_{exact} and λ_{exact} respectively. Nevertheless, the ECM-algorithm performed relatively well on the data.

Table 6-6 gives also the results with the algorithm by Schürle. The starting values of the algorithm are the parameters estimated with the ECM-algorithm. The estimated number of links N_M is the same as with the ECM algorithm. This is what was expected, because the Expectation step and Maximization step are the same for both algorithms (assuming that the starting parameters in the algorithm by Schürle are chosen conditional independent).

Algorithm		I	II	III	N_M	μ_{exact}	μ_{sim}	λ_{exact}	λ_{sim}
ECM	mean	517.00	4.94	9478.06	519.46	3.67e-03	2.27e-03	3.24e-02	3.74e-02
	std	78.10	5.50	78.22	78.08	8.26e-03	1.67e-03	2.30e-02	2.58e-02
	min	471.00	0.00	8505.00	480.00	1.05e-04	1.56e-04	0.00e+00	2.57e-03
	5%	492.00	0.00	9449.00	495.00	6.32e-04	6.06e-04	6.00e-03	8.96e-03
	25%	501.00	0.00	9480.00	503.00	1.37e-03	1.16e-03	1.60e-02	1.91e-02
	50%	506.00	4.00	9490.00	508.00	2.42e-03	1.87e-03	2.60e-02	3.12e-02
	75%	514.00	8.00	9496.00	517.00	3.79e-03	2.89e-03	4.20e-02	4.77e-02
	95%	542.05	15.05	9503.00	546.05	7.48e-03	5.21e-03	7.80e-02	8.93e-02
	max	1492.00	37.00	9520.00	1493.00	1.06e-01	1.45e-02	1.58e-01	2.01e-01
Schürle	mean	516.61	5.70	9477.69	519.46	4.32e-03	6.27e-03	4.47e-02	1.13e-01
	std	77.87	6.87	78.50	78.08	8.14e-03	4.89e-03	2.32e-02	8.63e-02
	min	471.00	0.00	8505.00	480.00	4.21e-04	9.63e-04	0.00e+00	1.55e-02
	5%	491.95	0.00	9447.00	495.00	1.26e-03	1.87e-03	1.40e-02	3.42e-02
	25%	500.00	0.00	9479.00	503.00	2.21e-03	3.22e-03	2.80e-02	5.94e-02
	50%	506.00	4.00	9489.00	508.00	3.16e-03	4.87e-03	4.20e-02	8.98e-02
	75%	514.25	8.00	9496.00	517.00	4.53e-03	7.64e-03	5.65e-02	1.36e-01
	95%	542.05	20.00	9503.00	546.05	7.79e-03	1.57e-02	8.80e-02	2.68e-01
	max	1492.00	67.00	9520.00	1493.00	1.06e-01	5.60e-02	1.58e-01	9.90e-01

Table 6-6: A table with the classifications results and estimated error levels. 1000 datasets are simulated and classified with both estimation methods, i.e. the ECM-algorithm and the algorithm by Schürle. The average Pearson correlation between the two correlated comparison variables for 1000 datasets is 0.192.

The results in the table show that this algorithm could not prevent/correct the complete misclassification discussed before. The estimated error levels μ_{sim} and λ_{sim} differ again from the simulated error levels μ_{exact} and λ_{exact} respectively. The error level estimates are not better with this method and sometimes even worse. Maybe, the creation of starting parameters with the ECM-algorithm causes the problems. See the discussion in Chapter 9 for more about the starting parameters for this algorithm.

6-5 The role of missing data

Many real world datasets contain records with missing data. In Section 2-4-3 were some pre-processing methods discussed to handle missing data such as imputation and removing the record. Another option was to leave the attribute missing. In Section 5-2 was described that Fellegi and Sunter [1969] set the weight for the comparison to zero if either of the attributes is missing. This was under the conditional independence assumption and the assumption that the missing values are equally distributed over the true links and true non-links. Several estimation methods for record linkage can only use binary comparison vectors and can not deal with missing values. For example, the ECM-algorithm applied to binary comparison vectors. Under the binary assumption, a comparison with a missing value is not distinguished from agreement/disagreement. If the comparison contains a missing value, the comparison is often seen as disagreement. In this section, the influence of this choice on the classification is studied.

Consider a 'good' dataset with 6 attributes of the entity. The first attribute contains missing values. The missing values are distributed over the datasets according a Bernoulli distributed random process. The Bernoulli random variable $X \sim Ber(p)$ is independent of the value of the comparison and the true link status of $A \times B$. The comparisons with missing values are labelled as disagreement. The ECM-algorithm is applied to the data. Each time, a different percentage of missing values is used (0%, 20%, 40%, 60%, 80%, 100%). In Table 6-7, the weight of each comparison field is given for agreement and disagreement (label 2 and 1). For $w_2(y^i)$, $w_3(y^i)$, $w_4(y^i)$, $w_5(y^i)$ and $w_6(y^i)$, the weights do not differ much in relation to the number of missing values. For the first comparison variable, variable 1, the weight for disagreement $w_1(y^1)$ increases when the number of missing values increases. For agreement, the weight remains almost constant.

In the set of comparison vectors belonging to the true links, the new $m_1(2)$-probability with missing values becomes $m_1(2)P(X = 0)$. For the set of comparison vectors belong to the true non-links, the $u_1(2)$ probability with missing values is $u_1(2)P(X = 1)$. The weight of agreement is

$$\frac{m_1(2)P(X=0)}{u_1(2)P(X=0)} = \frac{m_1(2)}{u_1(2)}.$$

The implies that the ratio and, therefore, the weight are not affected by the missing values. For the disagreeing comparisons, this does not hold. The probability on disagreement with missing values is $m_1(1) + m_1(2)P(X = 1)$. For the probability $u_1(0)$ holds $u_1(1) + u_1(2)P(X = 1)$.

% missing	y^i	$w_1(y^i)$	$w_2(y^i)$	$w_3(y^i)$	$w_4(y^i)$	$w_5(y^i)$	$w_6(y^i)$	N_M
0%	2	0.67	1.74	2.42	1.42	1.66	1.57	504
	1	-2.74	-3.28	-3.67	-1.78	-2.13	-2.71	
20%	2	0.66	1.74	2.42	1.42	1.65	1.57	502
	1	-0.98	-3.29	-3.81	-1.79	-2.13	-2.74	
40%	2	0.66	1.75	2.42	1.42	1.65	1.57	506
	1	-0.52	-3.27	-3.63	-1.78	-2.10	-2.72	
60%	2	0.69	1.75	2.41	1.43	1.65	1.57	504
	1	-0.27	-3.46	-3.41	-1.81	-2.12	-2.76	
80%	2	0.73	1.75	2.41	1.43	1.65	1.57	507
	1	-0.12	-3.41	-3.37	-1.81	-2.10	-2.74	
100%	2		1.75	2.41	1.43	1.65	1.57	505
	1	0.00	-3.40	-3.44	-1.81	-2.11	-2.75	

Table 6-7: The *field based* weights for a 'good' dataset of $N = 10^4$ comparison fields and $K = 6$ comparison variables. Each time, there are missing values added to the first comparison variable that gets the disagreement status.

The weight for disagreement under missing values is

$$\frac{m_1(1) + m_1(2)P(X = 1)}{u_1(1) + u_1(2)P(X = 1)}$$

which is not proportional with

$$\frac{m_1(2)}{u_1(2)}.$$

This behaviour is observed in Table 6-7. The agreement values can be estimated well while the disagreement values are influenced by the missing values.

One can also declare all comparisons with missing values as agreeing comparisons. In this case, the same behaviour is observed, but then the weights $w_1(2)$ remains constant and the weights $w_1(1)$ decreases. The weights for y^i for $2 \leq i \leq 6$ are nearly identical with the weights in Table 6-7. This is because the weights for agreement are

$$\frac{m_1(0) + m_1(1)P(X = 1)}{u_1(0) + u_1(1)P(X = 1)}.$$

The ratio is influenced by the missing values. For disagreement, the ratio

$$\frac{m_1(1)P(X = 0)}{u_1(1)P(X = 0)} = \frac{m_1(1)}{u_1(1)}$$

is not influenced by the missing values.

The estimates of both approaches can be used to combine the weights for agreement and disagreement and then set the weights for missing values to zero. This is a valid choice for

the weight if the missing values are independently distributed over the data. The ratio of a comparison with missing values is

$$\frac{P(X=1)}{P(X=1)} = 1.$$

Therefore, the weight, the log of 1, is zero.

With frequency based EM-algorithm described in Section 5-4, it is also possible to estimate the parameters in one classification. In Table 6-8, the results of a classification with the comparison types agreement/disagreement/either missing are given (labelled 2/1/0). The weight $w_1(0)$ is nearly zero for each amount of missing values. The weights $w_1(1)$ and $w_1(2)$ are nearly constant and identical with the weights found above. The frequency based EM-algorithm can be used effectively for this case with many missing values.

% missing	y^i	$w_1(y^i)$	$w_2(y^i)$	$w_3(y^i)$	$w_4(y^i)$	$w_5(y^i)$	$w_6(y^i)$	N_M
	2	0.67	1.74	2.42	1.42	1.66	1.57	
0%	1	-2.74	-3.28	-3.67	-1.78	-2.13	-2.71	504
	0	0.00						
	2	0.66	1.74	2.42	1.42	1.66	1.57	
20%	1	-2.65	-3.31	-3.68	-1.77	-2.15	-2.72	502
	0	0.02						
	2	0.66	1.75	2.42	1.42	1.66	1.57	
40%	1	-2.59	-3.33	-3.57	-1.78	-2.15	-2.73	506
	0	0.00						
	2	0.69	1.75	2.42	1.43	1.65	1.57	
60%	1	-3.01	-3.34	-3.56	-1.81	-2.09	-2.72	504
	0	0.00						
	2	0.73	1.75	2.41	1.43	1.65	1.57	
80%	1	-3.38	-3.36	-3.45	-1.82	-2.09	-2.73	507
	0	0.01						
	2		1.75	2.41	1.43	1.65	1.57	
100%	1		-3.40	-3.44	-1.81	-2.11	-2.75	505
	0	0.00						

Table 6-8: The *field-based* weights for a 'good' dataset of $N = 10^4$ comparison fields and $K = 6$ comparison variables. Each time, there are missing values added to the first comparison variable which has the disagreement status.

Chapter 7

Linking police and hospital road accident records

7-1 Introduction

In most countries, the police records road accidents [Amoros et al., 2011]. This data is of great importance in understanding road safety and the development of countermeasures [Rosman, 2001]. More and more countries start to use additional sources of information related to road accidents to enrich the police accident reports [IRTAD, 2011]. The most popular additional source is hospital data. The hospital data can give information about the medical consequences of road accidents. Other additional sources with road accident information are; Fire services, Insurance claims, Ambulance services and Mortality registers IRTAD [2011]. Analysis of additional data sources can lead to more knowledge about road safety problems.

There are two main reasons to use hospital data in road safety statistics and analysis [Amoros et al., 2011]. The first reason is to gain more information about the injuries of the road casualty. A police officer at the scene can not know the exact injuries of the road casualty. Therefore, the police data represent only one part of the road accident. The hospital data tells another part. The second reason to use hospital data is to deal with the incompleteness of the police road accident datasets. In most countries, it is known that there is an under-registration of road accidents (with serious injuries) [IRTAD, 2011]. The hospital data can be used to estimate the under-registration by the police (see Capture-Recapture, 2-8-4).

For clarity, the definition of a road accident (in the Netherlands) is: "An occurrence on a public road, related to traffic and causing damage to objects or injury to persons and in which at least one moving vehicle is involved" [SWOV, 2015*b*]. The official definition in Dutch [Ministerie van Infrastructuur en Milieu, 2015]: "*Een gebeurtenis op een voor het rij- en ander verkeer openstaande weg, die verband houdt met het verkeer ten gevolge waarvan schade is ontstaan en/of ten gevolge waarvan één of meerdere weggebruikers zijn overleden*

en/of gewond geraakt met uitzondering van een gebeurtenis waarbij uitsluitend voetgangers zijn betrokken. Onder voetgangers worden niet verstaan personen die zich voortbewegen met een hulpmiddel zoals rollerskates en skateboards." The definition of a severe road casualty in the Netherlands is [SWOV, 2015*b*]: "A casualty who has sustained a severe injury in a road crash. The injury must have a minimum severity of 2 on the Abbreviated Injury Scale (AIS). The score of an injury on this scale represents the severity of that injury. The Maximum AIS (MAIS) value represents the most severe injury a casualty has sustained." In this thesis, the injury scale AIS is not discussed in detail. For the reader, a good idea of a serious road accident is a road casualty which is hospitalised for at least one day excluding those who are hospitalised only for observation.

The Dutch road safety institute *Stichting Wetenschappelijk Onderzoek Verkeersveiligheid* (SWOV) performs each year a record linkage between the Dutch police road accident data and the Dutch hospital road accident data [Reurings and Bos, 2009]. The SWOV only analyses road accidents in the Netherlands with severe road casualties. The record linkage is based on deterministic classification principles. Both datasources do not contain personal identifiers, only quasi-identifiers. The reason for this is because there are too many small accidents and the registration of these accidents is (very) poor. In this thesis, the two sources of information are linked with the probabilistic record linkage framework by Fellegi and Sunter. In Section 7-2 and Section 7-3 are the police dataset and the hospital dataset discussed. In Section 7-5 are the two sources linked.

7-2 Police road accident data (BRON)

If the Dutch police is notified after a road accident in the Netherlands, they come to the scene if there are persons injured, there is major damage or drank abuse [Politie, 2015]. The police officer reports information like the vehicles and persons involved and the cause of the accident. Back at the police station, these information is inserted in a registration system by the police officer. Over the years, the police in The Netherlands used multiple systems for this registration [SWOV, 2015*a*]. The information collected by the police is send to the Dienst Verkeer en Scheepvaart (DVS), part of the Ministry of Infrastructure and Environment (I&M). Since 2004, the information is stored in the database "Bestand geRegistreerde Ongevallen in Nederland" (in English: File Registered road accidents in The Netherlands), shortly BRON. The database BRON is part of the open data portal of the Dutch governance [Ministerie van Infrastructuur en Milieu, 2015]. It is freely available but does not contain personal identifying information such as names, identifying numbers and licence plate numbers.

The BRON dataset is divided into seven microdatasets [Ministerie van Infrastructuur en Milieu, 2015]. Three microdatasets contain information about road accidents, the other datasets are reference files. The microdatasets with road accidents are; a mircodataset with road accident data, a file with information about the road network at the location of the accident and a file with detailed information about the vehicles involved in the accident. An overview of characteristics of road accidents in BRON of interest for this study are

- Date and time of the road accident
- Date of birth and sex of all the persons involved in the road accident
- Severity of the injury of the persons involved

- Hospital to which a road casualty is transported
- Vehicles involved
- Location of road accident
- Causes of road accident
- Weather and road conditions

If the registration by the police is incomplete or clearly incorrect, the Dienst Verkeer en Scheepvaart (DVS) does not look for additional information about the road accident. Only in case of fatalities or severe injuries, the DVS asks the police to correct or add additional information about the accident to the record. The DVS claims that they deduplicate the police data to prevent duplicate records in the data [Reurings and Bos, 2009].

7-2-1 The quality of BRON

For fatal road accidents, the data collected by the Dutch police is believed to be of good quality [SafetyNet, 2008]. For fatal accidents in the Netherlands, 90% of these accidents is registered in BRON [Reurings and Bos, 2009]. Overall, the road accident data in the Netherlands is believed to be of medium or low quality. The data is suffering a large under-registration, i.e. many road accidents are not found in BRON. For non-fatal incidents, the registration rate is worse. The registration rate also depends on the type of accident. For example, the registration rate of bicycle road accidents is much worse than for car accidents. For motorised vehicles, Reurings and Bos [2009] estimate the registration rate for seriously injured road casualties on 0.59 in 2008. For non-motorised vehicles, the registration rate was only 0.04 in 2008.

Road safety researchers at the SWOV have ideas about the cause of this under-registration of road accidents in BRON. It is known that not all victims of a road accident, especially non-motorised road accident victims, inform the police. They go to the hospital themselves. For road accidents where the police is called, understaffing and complicated/incomplete registration systems play a role in the under-registration. The SWOV thinks that the under-registration is not caused due to negligence of police officers at the scene.

The quality of the data itself depends on the type of characteristic Reurings and Bos [2009]. The information about the vehicles and the persons involved is believed to be of medium or good quality. The police registration of the severity of the injuries of the road causality is of poor quality. Another type of information of interest is the hospital to which the road casualty is brought; this information is often missing.

7-3 Hospital data (LMR)

Privacy plays an important role in the field of medical care. In the Netherlands, patient information such as health records are confidential. Patient (health) records are a huge source of information which are useful for hospitals and (medical) research [IRTAD, 2011]. For these purposes, Dutch Hospital Data started with the Landelijke Medische Registratie[1] (LMR) (National Registration Hospital Data) [Dutch Medical Data, 2015]. Under preservation of

[1] In 2013, LMR is phased out and replaced by Landelijke Basisregistratie Ziekenhuiszorg (LBZ).

privacy, Dutch Hospital Data provides anonymized data for research proposes. This database contains patient information of many Dutch hospitals. There is data available since 1964. The Stichting Wetenschappelijk Onderzoek Verkeersveiligheid (SWOV) orders each year a selection of the anonymized data for road safety research.

The LMR dataset is divided in three microdatasets [Centraal Bureau voor de Statistiek, 2015]. One microdataset contains basic information about the patient, the hospitalisation and the hospital. This dataset contains information like an identification number, the date of birth, the sex and zip code, but also information about the time of hospitalisation. This microdataset contains 48 variables, which are not all available for research proposes. The available variables of this microdataset contain vulnerable information for record linkage. The other microdatasets contain information about the diagnosis and treatment. This information is especially useful for analysis, but not for the record linkage process.

The following variables are registered in LMR and are useful for record linkage:

- Registration year
- Hospital number
- Date of birth and sex of the patient
- Date and time of hospitalisation
- Reason for hospitalisation
- Date and time of release (dead or alive)
- E-code (encoding of injury with external cause)

The E-code is used to label the external cause of injury and poisoning. This E-code has the form $Exxx.x$, where x is a number between 0 and 9. The first 3 numbers indicate the type of external cause of injury and poisoning. The numbers are also used to classify several types of accidents, for example road accidents. If it was a road accident, the meant of transport is denoted behind the E-code with a single digit equal or between 0 and 9. This encoding is based on the International Classification of Diseases [World Health Organization, 2004]. The means of transport in LMR are:

0 Pedestrian
1 Cyclist
2 Moped
3 Motorcyclist
4 Driver in passenger car
5 Passenger in passenger car
6 Not specified person in passenger car
7 Person in truck of bus
8 Other vehicles
9 Not specified

Not all records in LMR are relevant for the SWOV. Records with an E-code related to road accidents are relevant for research and ordered from Dutch Hospital Data. Also a set of records for unspecified accidents is included. This selection of LMR-records contains about 100.000 records each year [SWOV, 2015*c*]. This is about 1% of all LMR-records each year. Not all hospitals provide records to Dutch Hospital Data. DHD generates records for those hospitals. The records need to be removed before the record linkage, because they do not belong to an entity.

7-3-1 The quality of LMR

Paas and Veenhuizen [2002] researched in 2002 the reliability of information in LMR by reviewing 5745 health records of 55 hospitals. Their focus was on two groups of hospitalised patients; one group of general hospitalised patients and one group of hospitalised persons after a road accident. In this study, 4 types of data are distinguished; basic administrative data, hospitalisation data, medical hospitalization and road accident data. Paas and Veenhuizen [2002] asked hospitals to compare the data in LMR with the patient health records. This was used to examine the quality of LMR. The researchers did no research about the data in the health records itself. Therefore, it was especially a research about the quality of the database and the reliability of the encoding.

Of the 5745 health records, the basic administration data in LMR agreed with the health records in 99% of the cases. The same percentage was found for data about the hospitalisation, such as date, time and hospital of hospitalisation [Paas and Veenhuizen, 2002]. The information about the diagnosis was correct in 84% of the cases. The road accident data in LMR was of less quality. The circumstances of the accident are correctly reported in 91% of the cases. Paas and Veenhuizen [2002] do some recommendations to improve the quality. There is more feedback needed between the medical registrant and the medical specialist. Also the encoding method (of some types of accidents) can be unclear.

7-4 Deduplication of police and hospital data

Both the police dataset BRON and the hospital dataset LMR contain duplicate records. Duplicate records may influence the results of a linking operation. For the linking of BRON and LMR, one-to-one linking is required. Therefore, deduplication of the datasets is a good first step. The hospital dataset LMR contains duplicated records because road casualties are transported between hospitals. In each hospital, the road casualty gets a new registration. The dataset is deduplicated by the SWOV with variables not available for this thesis.

The police data in BRON is deduplicated by the supplier. In this thesis was observed that there are still duplicates in the data. Some road casualties where found multiple times in the datasets. Although it is not known for sure why this happens, it is very likely that it is the result of two police officers registering the same road accident. It should be mentioned that there are also twins among them. In this thesis, the duplicate records pairs in BRON are identified with the Fellegi and Sunter [1969] framework and thereafter removed. The probabilistic record linkage process is not extensively discussed. The linkage between the police dataset BRON and the hospital dataset LMR in Section 7-5 describes a linking operation in more detail.

7-4-1 Comparing and indexing

For the deduplication of road casualties in BRON, 6 attributes of the road casualty are used. The overview below gives a description and explains how this variable is compared. Each comparison can take the values agreement and disagreement. If an attribute is missing, the comparison disagrees. The following variables are used:

Date and time of the road accident Each record has the date and time of the road accident (according to the police officer). For each pair of records, this difference between the date and time of the the first record and the date and time of the second record may not exceed 3 hours. If the difference exceeds 3 hours, then the comparison is disagrees. If it does not, it agrees. The date of the road accident is used a blocking key (Section 2-5-1).

Sex The sex is compared as a categorical variable (See Section 2-6-3). If both records agree on the sex, the comparison agrees. Otherwise it is disagrees.

Mode of transport The mode of transport of the road accident casualty is also compared as an categorical variable (See Section 2-6-3).

Date of birth The date of birth of the road casualty is stored as a string and is compared with the exact string comparison method (Section 2-6-1). This variable is also used as blocking key (Section 2-5-1).

Coordinates of road accident Each record contains the coordinates of the road accident. The geographical distance between the coordinates in both record pairs is compared. If the distance is less than or equal to one kilometer, then the comparison agrees. If it is more than one kilometer, then the comparison disagrees.

The date of the road accident and the date of birth are used as blocking key. After blocking, 1474 record pairs are found for the police road accident data from 2007 until 2013.

7-4-2 Finding duplicates

The records pairs are classified with the Fellegi and Sunter [1969] classification framework. Two assumptions are made for this classification; the binary assumption and the conditional independence assumption. Therefore, the ECM-algorithm described in Section 4-3 is suitable for the estimation of parameters. The algorithm resulted in 893 links in the dataset. The full estimation output is given in Table 7-1. The structure of the table is similar with Table 6-1 in the simulation study. The comparison space is divided into three action sets. The 852 green comparison vectors are positive links (actionset I). The 103 orange comparison vectors are possible links (actionset II) and the red comparison vectors are positive non-links (actionset III). Note that all the record pairs with the positive links action have infinite weight. Analysis of the data showed that this is the result of comparison variable Y^{distance}. This variable is of such quality that it never occurs in the set of non-links (according to the estimation with the ECM-algorithm).

Clerical review (See Section 2-8-2) showed that many of the possible links are likely to belong to different road accident casualties. Therefore, only record pairs with action I are used to deduplicate the data. The number of records pairs with action I and action II are given in Table 7-2. The number of duplicates is related to the number of records in BRON for each year.

y^{distance}	y^{datetime}	y^{sex}	y^{mot}	$f(\boldsymbol{y})$	$w(\boldsymbol{y})$	$m(\boldsymbol{y})$	$u(\boldsymbol{y})$	λ	μ
2	2	2	2	527	∞	6.10e-01	0.00e+00	1.00e+00	0.00e+00
2	2	2	1	224	∞	2.32e-01	0.00e+00	3.90e-01	0.00e+00
2	2	1	2	84	∞	7.65e-02	0.00e+00	1.58e-01	0.00e+00
2	2	1	1	11	∞	2.91e-02	0.00e+00	8.16e-02	0.00e+00
2	1	2	2	5	∞	4.34e-03	0.00e+00	5.25e-02	0.00e+00
2	1	2	1	1	∞	1.65e-03	0.00e+00	4.82e-02	0.00e+00
2	1	1	2	0	∞	5.44e-04	0.00e+00	4.65e-02	0.00e+00
2	1	1	1	0	∞	2.07e-04	0.00e+00	4.60e-02	0.00e+00
1	2	2	2	103	-1.42	2.93e-02	1.22e-01	4.58e-02	1.22e-01
1	2	2	1	65	-2.35	1.11e-02	1.17e-01	1.65e-02	2.39e-01
1	2	1	2	57	-3.33	3.67e-03	1.02e-01	5.39e-03	3.41e-01
1	2	1	1	71	-4.26	1.40e-03	9.85e-02	1.72e-03	4.40e-01
1	1	2	2	99	-6.61	2.08e-04	1.55e-01	3.23e-04	5.95e-01
1	1	2	1	85	-7.54	7.91e-05	1.49e-01	1.15e-04	7.44e-01
1	1	1	2	68	-8.52	2.61e-05	1.30e-01	3.60e-05	8.74e-01
1	1	1	1	74	-9.45	9.92e-06	1.26e-01	9.92e-06	1.00e+00

Table 7-1: A table with the comparison space for a deduplication of the police dataset BRON. The comparison space in divided into a set of positive links (green), possible links (orange) and positive non-links (red).

	2007	2008	2009	2010	2011	2012	2013
Action I	223	181	127	73	28	36	184
Action II	31	27	21	7	0	2	15

Table 7-2: Number duplicate records pairs with action I and action II for each year.

7-5 Linking BRON and LMR

In this section, the police road accident records in BRON are linked with the hospital records in LMR. The records are linked for the years 2007 till 2013. In this section, the record linkage for year 2008 is used to show details.

For the record linkage is the Fellegi and Sunter [1969] framework described in Chapter 3 used. Because the underlying true link status between the databases is not known, the unsupervised learning methods described in this thesis are needed for the estimation of important parameters. For the linkage operation, the duplicated police database BRON and the deduplicated hospital database LMR are used.

7-5-1 Indexing and comparison

Based on previous research by Reurings and Bos [2009], the link prevalence of the record linkage between the police and hospital records is very low. The police data contains between 55000 and 10000 each year. The number of records decreased rapidly in the last years (due to under-registration). The hospital data contains about 100000 records each year. Of these records, about 25000 records are relevant for the linkage. The number of links in the other part of the data is very small. Therefore, there are $\sim 10^9$ record pairs.

In Reurings and Bos [2009] was observed that the number of links was between 2000 and 6000 each year. Therefore, the link prevalence is relatively small for a record linkage problem. Indexing was used by Reurings and Bos to reduce the number of pairs, especially for the purpose of reducing the computational time of the deterministic classification algorithm. In this thesis, indexing is used for the same purposes but also to let the unsupervised learning algorithm perform well (see Estimation of parameters with the ECM-algorithm, Section 6-4-1).

For the linkage of the police and hospital data are 5 characteristics of the road casualty used. The characteristics are described below. Some of them are also used for indexing.

Time between road accident and the hospitalisation (epoch) The police officer records the date and time of a road accident and the hospital records the date and time of the hospitalisation of road casualty. The road casualty is brought to the hospital with an ambulance or goes to the hospital on his own and enters the hospital after some time after the road accident. Reurings and Bos [2009] show that most of the road casualties that need medical help enter the hospital within 12 hours after the road accident. In may happen that the date and time of the accident are reported incorrectly. Therefore, it may happen that (according to the stored information) the road casualty enters the hospital before the road accident happened [Reurings and Bos, 2009].

In this thesis, the (registered) time between the road accident and hospitalisation is called the 'epoch'. If the epoch is between -1 and 12 hours, the comparison agrees. If it not between these times, the comparison disagrees.

The date of the road accident and the date of the hospitalisation are used for indexing. The candidate record pairs are pairs for which the date of the road accident and the date of hospitalisation do not differ more than 3 days. Record pairs for which the hospitalisation was the day before the accident are also included. Note that this a Sorted Neighbourhood indexing (see Sorted Neighbourhood Indexing, Section 2-5-2). This blocking key reduces the number of candidate record pairs to about 10^7 record pairs each year.

Sex (sex) The sex of the road casualty is found in both dataset and is compared as a categorical variable (see Comparing categorical information, Section 2-6-3). If the sex in both records agrees, then the comparison agrees. Otherwise it disagrees.

Mode of transport (mot) The mode of transport of the road accident casualty is also compared as an categorical variable (see Comparing categorical information, Section 2-6-3). If the mode of transport in both records agrees, then the comparison agrees, otherwise it disagrees. Some modes of transport are easily interchanged. For example, a moped driver and motor cyclist can be interchanged in the record generating process. In the comparison in this thesis, partial agreement is not used for these cases.

Date of birth (dob) The date of birth of the road casualty is stored as a string and is compared with the exact string comparison method (see Comparing date and time information, Section 2-6-4)).

Hospital (hosp) The police stores the name of the hospital if the road accident casualty needs to be transported to be hospital (with an ambulance). In the police dataset BRON, the name of the hospital is converted into a number. This number uniquely

identifies the hospital. This number is compared with the hospital number found in LMR. Exact comparison is used to compare this attribute. The coordinates of the road accident are found in the police dataset BRON (It is neve missing). The SWOV has a database with the coordinates of all hospitals. The geographical distance between the road accident and the hospital is computed in kilometers. If the distance is less than 25 kilometer, then the comparison agrees. Otherwise, the comparison disagrees. If one of the methods resulted in agreement, then the comparison variable for the characteristic 'hospital' agrees. If both methods disagree, the comparison disagrees.

7-5-2 Record linkage with parameters from the ECM-algorithm

In this section, the police records and hospital records are linked with the Fellegi and Sunter algorithm in combination with the ECM-algorithm. This implies that the binary assumption (see Binary assumption, Section 3-4-2) and the conditional independence assumption are applied (see Conditional independence assumption, Section 3-4-1). In the simulation study was observed that the use of the ECM-algorithm can result in a good classification. In this section, agreement/disagreement is labelled with 2/1.

The 5 comparison variables used for the record linkage imply that there are 11 parameters of interest in this model, namely 5 m-marginal probability mass functions, 5 u-marginal probability mass functions and the link prevalence. The starting values for these parameters need to be chosen such that the ECM-algorithm convergences to the 'correct' parameters. The choice of starting parameters is based on file characteristics and knowledge from previous results by the SWOV. The number of links from previous studies is used to set the starting value of the link prevalence. The number of links is set to 6000. The m-marginal probability mass functions are chosen based on assumptions about the data quality. The u-marginal probability mass functions are based on the proportions of agreement and disagreement found in all candidate record pairs. This makes use of the assumption that the number of true links is much less than the number of true non-links.

Table 7-3 shows the comparison space for the classification with the ECM-algorithm. The number of links for this classification was estimated on 9489. This table has a similar construction as Table 6-1 in the simulation study. The comparison vectors in the green rows are positive links (action I). The orange comparison vectors are possible links (action II) and the red comparison vectors are positive non-links (actionset III). The possible link comparison vector is the vector for which random decisions are needed to gain the exact link prevalence. This comparison vector is part of possible link set, because random decisions are not desired in this case. The comparison vectors in Table 7-3 are sorted based on the weight (see Computing weights, Section 3-4-3). The comparison vector with the highest weight agrees on all 5 comparison variables. The next comparison vector is the comparison vectors that agrees on all comparison variables, except the comparison of the mode of transport of the road casualty. The vector for which every comparison disagrees has the lowest weight. The comparison vectors show a reasonable ordering. Observe in Table 7-3 that all the comparison vectors in the positive link set agree on the date of birth. This variable is very important for the classification. The mode of transport is the least informative variable. The comparison vectors with the possible link action are vectors for which all comparisons of attributes agree, except the comparison of the date of birth. This comparison vector was found for 15153

record pairs.

Table 7-4 shows the classification results from 2007 to 2013. The table shows the number of record pairs with the positive link action N_I, the number of record pairs with the possible link action N_{II} and the number of records with the positive non-link action N_{III}. The estimated number of links N_M is given. Column $N_{M,SWOV}$ represents the estimated number of links found with the deterministic classification method used by the SWOV. The table shows that the number of links found in this section is higher than the number of links found by the SWOV. The trend over the years in the nearly the same. For both methods, the least number of links is found in 2012 and the largest number of links was found in 2008. Observe that the number of record pairs with the possible link action N_I is close to the number of links found by the SWOV. Each year, the number of possible links is larger than the number of positive links. In Appendix C-1 is the comparison space given for 2007 till 2013. For each year, the comparison vector for which all comparisons agree except the date of birth is classified

y^{hosp}	y^{epoch}	y^{sex}	y^{mot}	y^{dob}	$f(\boldsymbol{y})$	$w(\boldsymbol{y})$	$m(\boldsymbol{y})$	$u(\boldsymbol{y})$	λ	μ
2	2	2	2	2	4458	16.60	4.68e-01	2.90e-08	1.00e+00	2.90e-08
2	2	2	1	2	1828	13.81	1.94e-01	1.95e-07	5.32e-01	2.24e-07
2	2	1	2	2	42	12.23	5.40e-03	2.63e-08	3.39e-01	2.50e-07
2	1	2	2	2	143	11.20	1.53e-02	2.09e-07	3.33e-01	4.59e-07
1	2	2	2	2	82	10.33	9.08e-03	2.96e-07	3.18e-01	7.55e-07
2	2	1	1	2	31	9.44	2.23e-03	1.77e-07	3.09e-01	9.32e-07
2	1	2	1	2	94	8.41	6.33e-03	1.41e-06	3.07e-01	2.34e-06
1	2	2	1	2	79	7.54	3.76e-03	1.99e-06	3.00e-01	4.33e-06
2	1	1	2	2	6	6.84	1.76e-04	1.90e-07	2.97e-01	4.52e-06
1	2	1	2	2	12	5.97	1.05e-04	2.68e-07	2.96e-01	4.79e-06
2	2	2	2	1	15153	5.59	1.96e-01	7.37e-04	2.96e-01	7.41e-04
1	1	2	2	2	65	4.93	2.97e-04	2.14e-06	9.98e-02	7.43e-04
2	1	1	1	2	27	4.05	7.30e-05	1.28e-06	9.95e-02	7.45e-04
1	2	1	1	2	39	3.18	4.33e-05	1.81e-06	9.95e-02	7.47e-04
2	2	2	1	1	89784	2.80	8.13e-02	4.96e-03	9.94e-02	5.70e-03
1	1	2	1	2	235	2.15	1.23e-04	1.44e-05	1.81e-02	5.72e-03
2	2	1	2	1	13272	1.22	2.27e-03	6.68e-04	1.80e-02	6.38e-03
1	1	1	2	2	44	0.57	3.42e-06	1.94e-06	1.58e-02	6.39e-03
2	1	2	2	1	92678	0.19	6.42e-03	5.31e-03	1.58e-02	1.17e-02
1	2	2	2	1	146182	-0.68	3.81e-03	7.53e-03	9.33e-03	1.92e-02
2	2	1	1	1	82369	-1.57	9.37e-04	4.49e-03	5.52e-03	2.37e-02
1	1	1	1	2	208	-2.22	1.42e-06	1.30e-05	4.58e-03	2.37e-02
2	1	2	1	1	641212	-2.60	2.66e-03	3.57e-02	4.58e-03	5.95e-02
1	2	2	1	1	903707	-3.47	1.58e-03	5.06e-02	1.92e-03	1.10e-01
2	1	1	2	1	84077	-4.18	7.40e-05	4.82e-03	3.45e-04	1.15e-01
1	2	1	2	1	128846	-5.04	4.40e-05	6.82e-03	2.71e-04	1.22e-01
1	1	2	2	1	981977	-6.08	1.25e-04	5.43e-02	2.27e-04	1.76e-01
2	1	1	1	1	592078	-6.96	3.06e-05	3.24e-02	1.02e-04	2.08e-01
1	2	1	1	1	819194	-7.83	1.82e-05	4.59e-02	7.18e-05	2.54e-01
1	1	2	1	1	6592519	-8.87	5.16e-05	3.65e-01	5.36e-05	6.20e-01
1	1	1	2	1	873806	-10.44	1.44e-06	4.92e-02	2.03e-06	6.69e-01
1	1	1	1	1	5982938	-13.23	5.95e-07	3.31e-01	5.95e-07	1.00e+00

Table 7-3: A record linkage with the ECM-algorithm. The green rows represent positive links, the orange rows possible links and the red rows positive non-links. The estimated number of true links is 9489.

Year	N_I	N_{II}	N_{III}	N_M	$N_{M,\text{SWOV}}$	N_μ	N_λ
2007	6838	14451	19326559	9706		89	1058
2008	6775	15153	18025257	9489	6801	87	948
2009	6331	13956	15399023	8186	6155	73	678
2010	4451	9019	8911868	6786	4438	40	837
2011	2283	4617	4183878	3357	2266	17	348
2012	2145	4019	3370744	3161	2141	16	349
2013	2584	4187	5101887	4252	2566	28	760

Table 7-4: Table with the results of classifications with the ECM-algorithm. N_I is the number of record pairs classified as positive link, N_{II} the number of record pairs classified as possible link and N_{III} the number of record pairs record pairs classified as positive non-link.

as possible link. The ordering of the comparison space is for each year more or less the same.

The set of possible links is very large and contains a lot of true links according to the estimation. Clerical review showed that this does not seem to be reasonable. The ECM-algorithm does not give a reasonable estimate for the number of links in this set. Experiments with the data show that parameters in the comparison step have a large influence on this estimation result. For example, the maximum epoch (time between accident according to the police and the hospitalisation) has a large influence on the estimated number of links. In Figure 7-1 is the maximum epoch parameter varied between 1 and 36 hours in steps of 1 hour. The figure shows clearly that the number of links is heavily influenced by this parameter. Especially the years 2007, 2008 and 2009 show major differences in the estimated number of links. This appears to be related to the number of possible links, which was very large for these years.

In the simulation study was observed that the ECM-algorithm may converge to incorrect parameter estimates in case of a very small link prevalence. Table 7-4 shows that the link

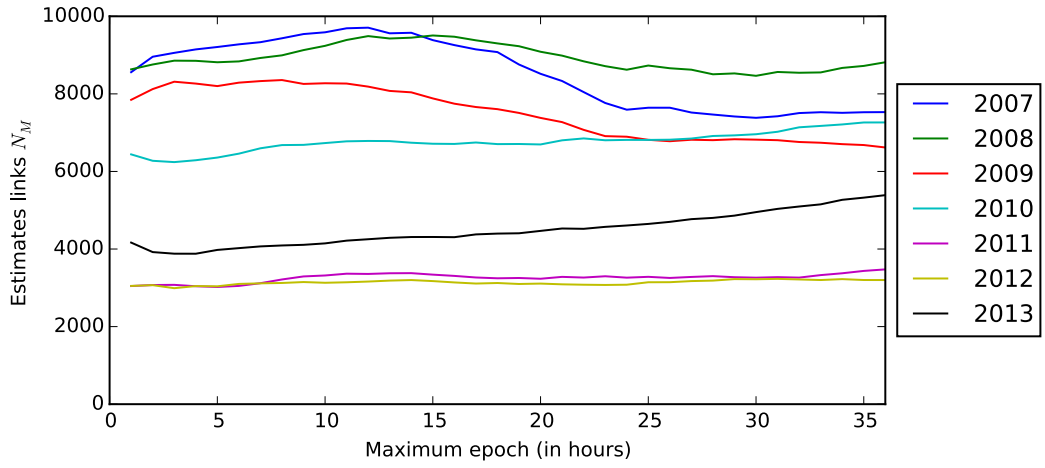


Figure 7-1: The maximum epoch is varied between 1 and 36 hours in steps of 1 hour. For each value, the data is classified and the number of links is estimated. Observe that the number of links is heavily influenced by this parameter.

y^{hosp}	y^{epoch}	y^{sex}	y^{mot}	y^{dob}	$f(\boldsymbol{y})$	$w(\boldsymbol{y})$	$m(\boldsymbol{y})$	$u(\boldsymbol{y})$	λ	μ
2	2	2	2	2	4458	12.12	6.43e-01	3.51e-06	1.00e+00	3.51e-06
2	2	2	1	2	1828	9.82	2.67e-01	1.46e-05	3.57e-01	1.81e-05
2	2	1	2	2	42	7.79	7.09e-03	2.94e-06	8.99e-02	2.10e-05
2	1	2	2	2	143	6.72	2.08e-02	2.52e-05	8.29e-02	4.62e-05
1	2	2	2	2	82	5.81	1.16e-02	3.48e-05	6.21e-02	8.10e-05
2	2	1	1	2	31	5.48	2.94e-03	1.22e-05	5.05e-02	9.32e-05
2	1	2	1	2	94	4.41	8.62e-03	1.04e-04	4.75e-02	1.98e-04
1	2	2	1	2	79	3.51	4.82e-03	1.44e-04	3.89e-02	3.42e-04
2	2	2	2	1	441	2.92	2.20e-02	1.18e-03	3.41e-02	1.52e-03
2	1	1	2	2	6	2.38	2.29e-04	2.11e-05	1.21e-02	1.55e-03
1	2	1	2	2	12	1.48	1.28e-04	2.91e-05	1.19e-02	1.57e-03
2	2	2	1	1	1232	0.62	9.12e-03	4.90e-03	1.17e-02	6.48e-03
1	1	2	2	2	65	0.41	3.75e-04	2.50e-04	2.63e-03	6.73e-03
2	1	1	1	2	27	0.08	9.50e-05	8.75e-05	2.25e-03	6.82e-03
1	2	1	1	2	39	-0.82	5.32e-05	1.21e-04	2.16e-03	6.94e-03
2	2	1	2	1	276	-1.41	2.42e-04	9.90e-04	2.11e-03	7.93e-03
1	1	2	1	2	235	-1.89	1.56e-04	1.04e-03	1.86e-03	8.96e-03
2	1	2	2	1	1971	-2.48	7.09e-04	8.47e-03	1.71e-03	1.74e-02
1	2	2	2	1	3191	-3.38	3.97e-04	1.17e-02	9.98e-04	2.92e-02
2	2	1	1	1	1005	-3.71	1.00e-04	4.11e-03	6.01e-04	3.33e-02
1	1	1	2	2	44	-3.92	4.14e-06	2.09e-04	5.01e-04	3.35e-02
2	1	2	1	1	8560	-4.78	2.94e-04	3.52e-02	4.97e-04	6.86e-02
1	2	2	1	1	11664	-5.69	1.65e-04	4.86e-02	2.02e-04	1.17e-01
1	1	1	1	2	208	-6.22	1.72e-06	8.67e-04	3.73e-05	1.18e-01
2	1	1	2	1	1638	-6.81	7.82e-06	7.10e-03	3.56e-05	1.25e-01
1	2	1	2	1	2456	-7.71	4.38e-06	9.81e-03	2.78e-05	1.35e-01
1	1	2	2	1	20473	-8.79	1.28e-05	8.40e-02	2.34e-05	2.19e-01
2	1	1	1	1	7290	-9.11	3.24e-06	2.95e-02	1.06e-05	2.48e-01
1	2	1	1	1	9580	-10.02	1.82e-06	4.07e-02	7.34e-06	2.89e-01
1	1	2	1	1	84500	-11.09	5.32e-06	3.49e-01	5.52e-06	6.38e-01
1	1	1	2	1	16748	-13.12	1.41e-07	7.03e-02	2.00e-07	7.08e-01
1	1	1	1	1	71392	-15.42	5.87e-08	2.92e-01	5.87e-08	1.00e+00

Table 7-5: A record linkage with the ECM-algorithm. The green rows represent positive links, the orange rows possible links and the red rows positive non-links. The estimated number of true links is 6901.

prevalence π is below 0.001 (the estimated number of links divided by $N_I + N_{II} + N_{III}$). The small link prevalence may influence the estimations. The solution for was found in reducing the number of candidate records pairs. To reduce the number of record pairs, an additional indexing method is applied. It seems obvious to use an additional indexing criteria based on the date of birth because this variable may cause the problems described above.

To overcome the mentioned problems, the set of comparison vectors for which all comparisons agree except the comparison of the date of birth is reduced. This is done by using different indexing criteria. Besides indexing on the epoch variable, the data is also blocked on the year of birth of the road casualty, i.e. the candidate record pairs agree on the year of birth. The candidate record pairs are classified with the ECM-algorithm. The result is given in Table 7-5. The comparison space is ordered in the same way as the comparison space in Table C-2. The estimated number of links is 6978 and the number of possible links is 441. The other years show similar results (see Estimation with the ECM-algorithm and data blocked on the year of birth, Section C-2).

Table 7-6 shows classification results for each year. These classification results differ less from the results found by the SWOV. The number of record pairs with the positive link action is almost equal or of the same order. The same indexing procedure is used with the day of birth or the month of birth instead of the year of birth. The number of estimated errors in the set

Year	Index key	N_I	N_{II}	N_{III}	N_M	$N_{M,\text{SWOV}}$	N_μ	N_λ
2007	year	6827	395	259429	6978		84	88
	month	6833	1269	1601638	7096		85	127
	day	6833	549	625142	6970		85	82
2008	year	6757	441	242612	6901		84	84
	month	6775	1315	1488959	6999	6801	87	112
	day	6763	527	581331	6839		82	59
2009	year	6331	386	207596	6464		80	81
	month	6331	1161	1271347	6447	6155	73	72
	day	6331	484	494978	6385		73	51
2010	year	4449	284	118868	4604		40	71
	month	4451	728	737095	4596	4438	40	66
	day	4451	285	288436	4489		40	28
2011	year	2283	129	54602	2353		18	30
	month	2283	390	347961	2365	2266	17	34
	day	2283	174	135307	2331		17	23
2012	year	2145	106	42246	2203		17	27
	month	2145	363	272101	2272	2141	16	50
	day	2145	139	105973	2181		16	19
2013	year	2558	88	61493	2617		19	39
	month	2584	364	416135	2731	2566	28	85
	day	2584	134	162130	2604		28	28

Table 7-6: Table with the results of classifications with the ECM-algorithm with additional indexing key. N_I is the number of record pairs classified as positive link, N_{II} the number of record pairs classified as possible link and N_{III} the number of record pairs record pairs classified as positive non-link.

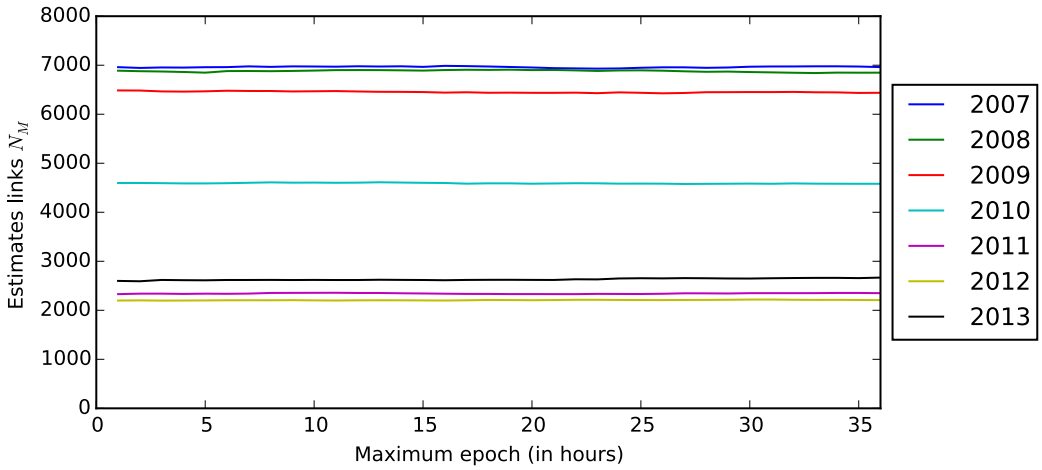


Figure 7-2: The maximum epoch is varied between 1 and 36 hours in steps of 1 hour. For each value, the data is classified and the number of links is estimated. Observe that the number of links is very constant with respect to this parameter.

of positive links N_λ is now more in balance with the number of estimated errors in the set of positive non-links N_μ. This is not a hard indication for a good linkage, but this is easier to explain the results of the classification. In Figure 7-2, the estimated number of links is given as function of the maximum epoch parameter. The parameter is varied between 1 and 36 hours in steps of 1 hour. The estimated number of links is now very constant with respect to this parameter.

Besides the binary assumption, the conditional independence assumption is applied to the record linkage when using the ECM-algorithm. In Table 7-7 is the Pearson's correlation matrix given for all comparison variables. The correlation is computed with comparison vectors of all candidate record pairs. Most of the correlations between comparison attributes are slightly positive. It is likely that this is because both the true links and true non-links are included. The true links have agreeing comparison variables for most variables while the true non-links disagree on most comparisons. Therefore, the correlation is slightly positive. The correlation of the 'date of birth' comparison variable and other comparison variables is slightly positive. In the simulation study was data simulated with a much larger (positive) correlation. The results were slightly different, but not much (see Estimation of parameters with the algorithm of Schürle, Section 6-4-3). Despite of the small correlations, the ECM-algorithm seems to be a valid algorithm for this dataset.

	y^{hosp}	y^{epoch}	y^{sex}	y^{mot}	y^{dob}
y^{hosp}	1.000000	0.005529	-0.001281	0.000035	0.057200
y^{epoch}	0.005529	1.000000	0.001412	0.011096	0.047388
y^{sex}	-0.001281	0.001412	1.000000	0.004091	0.017023
y^{mot}	0.000035	0.011096	0.004091	1.000000	0.031729
y^{dob}	0.057200	0.047388	0.017023	0.031729	1.000000

Table 7-7: A correlation matrix for the candidate comparison vectors (only indexed on the epoch) for the record linkage of road accidents in 2008.

7-5-3 Record linkage with missing values

In this section, the frequency based application of the EM-algorithm is used for classification. This algorithm was described in Section 5-4. The algorithm is applied to make use of comparisons with missing values (see The role of missing data, Section 6-5). It was shown that it can help to estimate the weight of a missing value. The comparisons agreement/disagreement/either missing are labelled 2/1/0. Not all comparison variables have missing values. The variables with missing values are the sex, mode of transport and date of birth of the road casualty.

In Table 7-9, the result of the classification with the EM-algorithm is given. The rows represent the comparison vectors of the comparison space. The record pairs are indexed on the epoch and the year of birth. The number of links N_M is 6908. Observe that the same comparison vector is classified as possible link, i.e. the comparison vector for which all comparisons agree except the comparison of the date of birth. The comparison space has a remarkable ordering. The comparison vector with the second highest weight is the vector that agrees on all comparisons except the sex of the road casualty. This variable is missing. If the variable is missing, then one of the records, or both, have a missing value on this field. It is not necessary that the comparison is incorrect. This is why the missing value comparison is expected before the disagreeing comparison in the ordering. This is observed for the sex, but not for the mode of transport. The agreeing comparison gets the most weight, but after that, the disagreeing comparison gets the most weight and not the missing comparison. In Appendix C-4, the results of the classification are given for each year. The ordering of the comparison vectors as described above is observed each year.

Table 7-8 shows the classification result given for each year. The results show minor differences with the classification in Section 7-5-2 and especially Table 7-6. Each estimation of N_M does not differ more than 10 links. The classification seem to work well but does not seem to add much distinguishing power. In Table 7-10 are the results given without using the year of birth as blocking key. The results are similar with the results achieved in Section 7-5-2 and Table 7-4. The number of links N_M is differs more than when using the year as indexing variable.

Year	N_I	N_{II}	N_{III}	N_M	$N_{M,\text{SWOV}}$	N_μ	N_λ
2007	6830	395	259426	6977		85	87
2008	6760	441	242609	6908	6801	84	85
2009	6329	386	207598	6455	6155	73	75
2010	4449	284	118868	4607	4438	41	71
2011	2283	129	54602	2352	2266	18	30
2012	2142	106	42249	2202	2141	16	27
2013	2557	88	61494	2613	2566	18	36

Table 7-8: Table with the results of classifications with the ECM-algorithm. N_I is the number of record pairs classified as positive link, N_{II} the number of record pairs classified as possible link and N_{III} the number of record pairs record pairs classified as positive non-link.

y^{hosp}	y^{epoch}	y^{sex}	y^{mot}	y^{dob}	$f(\boldsymbol{y})$	$w(\boldsymbol{y})$	$m(\boldsymbol{y})$	$u(\boldsymbol{y})$	λ	μ
2	2	2	2	2	4458	12.12	6.42e-01	3.51e-06	1.00e+00	3.51e-06
2	2	0	2	2	27	11.18	4.71e-03	6.57e-08	3.58e-01	3.58e-06
2	2	2	1	2	1339	9.97	1.95e-01	9.12e-06	3.53e-01	1.27e-05
2	2	2	0	2	489	9.48	7.15e-02	5.45e-06	1.58e-01	1.81e-05
2	2	0	1	2	13	9.04	1.43e-03	1.71e-07	8.63e-02	1.83e-05
2	2	0	0	2	5	8.55	5.24e-04	1.02e-07	8.48e-02	1.84e-05
2	2	1	2	2	15	6.76	2.49e-03	2.88e-06	8.43e-02	2.13e-05
2	1	2	2	2	143	6.71	2.06e-02	2.52e-05	8.18e-02	4.65e-05
1	2	2	2	2	82	5.82	1.17e-02	3.48e-05	6.12e-02	8.13e-05
2	1	0	2	2	0	5.77	1.51e-04	4.71e-07	4.95e-02	8.18e-05
1	2	0	2	2	2	4.88	8.59e-05	6.51e-07	4.94e-02	8.24e-05
2	2	1	1	2	12	4.62	7.56e-04	7.47e-06	4.93e-02	8.99e-05
2	1	2	1	2	60	4.56	6.25e-03	6.54e-05	4.85e-02	1.55e-04
2	2	1	0	2	1	4.13	2.76e-04	4.46e-06	4.23e-02	1.60e-04
2	1	2	0	2	34	4.07	2.29e-03	3.91e-05	4.20e-02	1.99e-04
1	2	2	1	2	49	3.67	3.56e-03	9.04e-05	3.97e-02	2.89e-04
2	1	0	1	2	1	3.62	4.59e-05	1.22e-06	3.62e-02	2.90e-04
1	2	2	0	2	30	3.18	1.30e-03	5.40e-05	3.61e-02	3.45e-04
2	1	0	0	2	0	3.13	1.68e-05	7.31e-07	3.48e-02	3.45e-04
2	2	2	2	1	441	2.95	2.26e-02	1.18e-03	3.48e-02	1.53e-03
1	2	0	1	2	3	2.74	2.61e-05	1.69e-06	1.22e-02	1.53e-03
1	2	0	0	2	0	2.25	9.55e-06	1.01e-06	1.21e-02	1.53e-03
2	2	0	2	1	9	2.02	1.66e-04	2.21e-05	1.21e-02	1.55e-03
2	1	1	2	2	6	1.35	7.96e-05	2.06e-05	1.20e-02	1.57e-03
2	2	2	1	1	840	0.81	6.89e-03	3.07e-03	1.19e-02	4.64e-03
1	2	1	2	2	10	0.46	4.53e-05	2.85e-05	4.99e-03	4.67e-03
1	1	2	2	2	65	0.41	3.75e-04	2.50e-04	4.95e-03	4.92e-03
2	2	2	0	1	392	0.32	2.52e-03	1.83e-03	4.57e-03	6.75e-03
2	2	0	1	1	14	-0.13	5.05e-05	5.74e-05	2.05e-03	6.81e-03
1	1	0	2	2	0	-0.53	2.75e-06	4.67e-06	2.00e-03	6.81e-03
2	2	0	0	1	12	-0.62	1.85e-05	3.43e-05	2.00e-03	6.85e-03
2	1	1	1	2	17	-0.79	2.42e-05	5.36e-05	1.98e-03	6.90e-03
2	1	1	0	2	9	-1.29	8.85e-06	3.20e-05	1.96e-03	6.93e-03
1	2	1	1	2	25	-1.68	1.38e-05	7.40e-05	1.95e-03	7.01e-03
1	1	2	1	2	145	-1.74	1.14e-04	6.49e-04	1.94e-03	7.66e-03
1	2	1	0	2	11	-2.17	5.04e-06	4.42e-05	1.82e-03	7.70e-03
1	1	2	0	2	90	-2.23	4.17e-05	3.87e-04	1.82e-03	8.09e-03
2	2	1	2	1	267	-2.40	8.76e-05	9.67e-04	1.77e-03	9.06e-03
2	1	2	2	1	1971	-2.46	7.25e-04	8.47e-03	1.69e-03	1.75e-02
1	1	0	1	2	2	-2.67	8.36e-07	1.21e-05	9.62e-04	1.75e-02
...	...	...	...	...	...	...	...	...	...	...
2	1	1	0	1	2823	-10.45	3.12e-07	1.08e-02	6.58e-06	2.51e-01
1	2	1	1	1	6095	-10.85	4.86e-07	2.49e-02	6.26e-06	2.76e-01
1	1	2	1	1	52625	-10.90	4.02e-06	2.18e-01	5.78e-06	4.94e-01
1	2	1	0	1	3255	-11.34	1.78e-07	1.49e-02	1.76e-06	5.09e-01
1	1	2	0	1	31875	-11.39	1.47e-06	1.30e-01	1.58e-06	6.39e-01
1	1	0	1	1	1038	-11.84	2.95e-08	4.08e-03	1.13e-07	6.43e-01
1	1	0	0	1	509	-12.33	1.08e-08	2.44e-03	8.31e-08	6.46e-01
1	1	1	2	1	16432	-14.11	5.11e-08	6.88e-02	7.24e-08	7.15e-01
1	1	1	1	1	43745	-16.26	1.55e-08	1.79e-01	2.12e-08	8.93e-01
1	1	1	0	1	26100	-16.75	5.69e-09	1.07e-01	5.69e-09	1.00e+00

Table 7-9: A record linkage with the frequency adjusted version of the EM-algorithm. The green rows represent positive links, the orange rows possible links and the red rows positive non-links. The data is indexed on the epoch and the year of birth. The estimated number of true links is 6908.

Year	N_I	N_{II}	N_{III}	N_M	$N_{M,\text{SWOV}}$	N_μ	N_λ
2007	6831	14356	19326661	9708		83	1062
2008	6769	15022	18025394	9435	6801	83	932
2009	6408	13831	15399071	7951	6155	110	588
2010	4449	8951	8911938	6938	4438	38	907
2011	2283	4592	4183903	3382	2266	17	360
2012	2145	3924	3370839	3168	2141	16	351
2013	2720	4049	5101889	4079	2566	95	675

Table 7-10: Table with the results of classifications with the ECM-algorithm. N_I is the number of record pairs classified as positive link, N_{II} the number of record pairs classified as possible link and N_{III} the number of record pairs record pairs classified as positive non-link.

7-5-4 Classification with the EM-algorithm and multiple levels of agreement

The frequency based EM-algorithm can also be used to add comparisons like "agreement and the value is ...". This type of comparison is not applicable for all fields. Especially the attributes sex and mode of transport are useful because these fields contain categorical information. In this section, these characteristics have multiple levels of agreement on a certain attribute. For the sex characteristic, there are now 3 comparison types;

- The sex in both records is 'Female'
- The sex in both records is 'Male'
- The sex disagrees

For the mode of transport, there are 8 comparisons;

- The mode of transport in both records is 'foot'
- The mode of transport in both records is 'moped'
- The mode of transport in both records is 'motorcycle'
- The mode of transport in both records is 'bicycle'
- The mode of transport in both records is 'car'
- The mode of transport in both records is 'truck or bus'
- The mode of transport in both records is another type of vehicle or unknown.
- The mode of transport disagrees

The estimation is performed with record pairs indexed on the epoch, but also with records indexed on the epoch and year of birth. The results are given in Table 7-11 and Table 7-12 respectively. The results for an index on the epoch shows again a large estimated number of links. With the year as blocking key, the results are closer related with the linking by the SWOV.

The ordering of the comparison space based on the estimated weight is given in Table 7-11 (Only the highest weights, because this comparison space is large, i.e. $2 \times 2 \times 2 \times 3 \times 8 = 192$ comparison vectors). Observe that the comparison vector with the highest weight agrees on all fields and is a female pedestrian. This is seems to be a reasonable result. Men are overrepresented in the population of serious injured road casualties. From Reurings and Bos [2009] was known that pedestrians are strongly under-registered by the police. Therefore, they do not occur often in the database. By theory, rare attributes get higher weights (see Chapter refchap:freq). The comparison vector with the second highest weight is a female

Year	N_I	N_{II}	N_{III}	N_M	$N_{M,\text{SWOV}}$	N_μ	N_λ
2007	9042	3047	19335759	10315		1437	2461
2008	9493	1340	18036352	10281	6801	1713	2446
2009	8760	1410	15409140	9070	6155	1651	1938
2010	7064	2142	8916132	7313	4438	1879	1744
2011	3173	450	4187155	3586	2266	477	880
2012	3113	976	3372819	3417	2141	672	791
2013	4226	443	5103989	4439	2566	977	1312

Table 7-11: Table with the results of classifications with the ECM-algorithm. N_I is the number of record pairs classified as positive link, N_{II} the number of record pairs classified as possible link and N_{III} the number of record pairs record pairs classified as positive non-link.

Year	N_I	N_{II}	N_{III}	N_M	$N_{M,\text{SWOV}}$	N_μ	N_λ
2007	6971	56	259624	6983		147	138
2008	6871	21	242918	6888	6801	135	153
2009	6420	27	207866	6443	6155	125	135
2010	4579	79	118943	4585	4438	119	76
2011	2342	24	54648	2358	2266	35	57
2012	2198	30	42269	2200	2141	38	36
2013	2619	23	61497	2624	2566	49	45

Table 7-12: Table with the results of classifications with the ECM-algorithm. N_I is the number of record pairs classified as positive link, N_{II} the number of record pairs classified as possible link and N_{III} the number of record pairs record pairs classified as positive non-link.

passenger of driver of a bus or truck. Women are not often truck drivers, so they do not occur often in the data.

y^{hosp}	y^{epoch}	y^{sex}	y^{mot}	y^{dob}	$f(\boldsymbol{y})$	$w(\boldsymbol{y})$	$m(\boldsymbol{y})$	$u(\boldsymbol{y})$	λ	μ
2	2	female	foot	2	135	14.83	1.66e-02	6.01e-09	1.00e+00	6.01e-09
2	2	female	truck/bus	2	3	14.78	1.12e-03	4.24e-10	9.83e-01	6.44e-09
2	2	female	motorcycle	2	29	14.43	2.21e-02	1.20e-08	9.82e-01	1.84e-08
2	2	male	foot	2	175	14.18	2.89e-02	2.01e-08	9.60e-01	3.86e-08
2	2	male	truck/bus	2	18	14.13	1.94e-03	1.42e-09	9.31e-01	4.00e-08
2	2	male	motorcycle	2	384	13.77	3.85e-02	4.02e-08	9.29e-01	8.01e-08
2	2	female	bicycle	2	651	12.67	7.23e-02	2.28e-07	8.91e-01	3.08e-07
2	2	female	car	2	513	12.40	7.69e-02	3.18e-07	8.19e-01	6.26e-07
2	2	female	moped	2	224	12.16	4.58e-02	2.40e-07	7.42e-01	8.66e-07
2	2	male	bicylce	2	719	12.01	1.26e-01	7.63e-07	6.96e-01	1.63e-06
2	2	female	else	2	0	11.91	4.83e-05	3.24e-10	5.70e-01	1.63e-06
2	2	male	car	2	977	11.74	1.34e-01	1.06e-06	5.70e-01	2.69e-06
2	2	male	moped	2	629	11.50	7.98e-02	8.05e-07	4.36e-01	3.50e-06
2	2	male	else	2	1	11.26	8.42e-05	1.09e-09	3.56e-01	3.50e-06
2	2	female	disagree	2	739	10.28	9.75e-02	3.34e-06	3.56e-01	6.84e-06
2	2	disagree	foot	2	9	10.04	5.00e-04	2.19e-08	2.58e-01	6.86e-06
2	2	disagree	truck/bus	2	0	9.99	3.37e-05	1.55e-09	2.58e-01	6.86e-06
2	2	disagree	motorcycle	2	1	9.63	6.67e-04	4.37e-08	2.58e-01	6.91e-06
2	2	male	disagree	2	1089	9.63	1.70e-01	1.12e-05	2.57e-01	1.81e-05
2	1	female	foot	2	5	9.44	5.42e-04	4.31e-08	8.72e-02	1.81e-05
2	1	female	truck/bus	2	0	9.39	3.65e-05	3.04e-09	8.67e-02	1.81e-05
2	1	female	motorcycle	2	1	9.04	7.23e-04	8.59e-08	8.66e-02	1.82e-05
2	1	female	foot	2	5	8.79	9.45e-04	1.44e-07	8.59e-02	1.84e-05
2	1	male	truck/bus	2	2	8.74	6.36e-05	1.02e-08	8.50e-02	1.84e-05
1	2	female	foot	2	2	8.54	3.03e-04	5.96e-08	8.49e-02	1.84e-05
1	2	female	truck/bus	2	0	8.49	2.04e-05	4.20e-09	8.46e-02	1.84e-05
2	1	female	motorcycle	2	21	8.38	1.26e-03	2.88e-07	8.46e-02	1.87e-05
1	2	female	motorcycle	2	1	8.13	4.05e-04	1.19e-07	8.33e-02	1.88e-05
1	2	male	foot	2	1	7.88	5.29e-04	1.99e-07	8.29e-02	1.90e-05
2	2	disagree	bicycle	2	12	7.87	2.18e-03	8.30e-07	8.24e-02	1.99e-05
1	2	male	truck/bus	2	0	7.84	3.56e-05	1.41e-08	8.02e-02	1.99e-05
2	2	disagree	car	2	6	7.60	2.32e-03	1.16e-06	8.02e-02	2.10e-05
1	2	male	motorcycle	2	10	7.48	7.05e-04	3.98e-07	7.79e-02	2.14e-05
2	2	disagree	moped	2	14	7.36	1.38e-03	8.76e-07	7.72e-02	2.23e-05
2	1	female	bicycle	2	15	7.28	2.36e-03	1.63e-06	7.58e-02	2.40e-05
...	...	...	...	...	...	...	...	...	...	...

Table 7-13: Table with the results of classifications with the ECM-algorithm. N_I is the number of record pairs classified as positive link, N_{II} the number of record pairs classified as possible link and N_{III} the number of record pairs record pairs classified as positive non-link.

Chapter 8

Horizon

8-1 Introduction

Additional resources of information are useful for the analysis of road safety (See Chapter 7 and IRTAD [2011]). In IRTAD [2011] was mentioned that data of fire services, insurance companies, ambulances services and mortality registers could be used. There are plenty of additional (open) data resources available which are interesting for road safety research. For example; court rulings can give additional information about the legal consequences of a road accident [Rechtspraak, 2015], emergency calls can be used to quantify the number of road accidents, alarms for emergency services are in some countries openly available and can be used for quantitative road accident information. Most of the mentioned data sources do not contain (personal) identifiers or are anonymised. The Fellegi and Sunter [1969] framework can be used to link anonymised records between these databases based on quasi-identifiers.

One of the largest (freely available) sources of information about road accidents may be found on the Internet, namely news articles about road accidents. In general, these articles contain basic information about the road accident such as the location and the vehicles involved. News articles may contain more information about the road accident such as pictures, eye witness reports and circumstantial information. News websites may also report accidents without police at the scene. All of this information can be used for a comprehensive analysis of road safety.

In this Chapter, the possibilities of these sources of information are explored. The goal is to retrieve articles about road accidents from news websites and link them to police accident records in BRON. This chapter provides a proof of concept for a linking operation between news articles and police records. There is a large number of challenges to link news articles to additional record based data sources. One of the largest complications is to extract information from the article that can be used to compare with information in the police records.

8-2 Collecting news articles

News articles from websites in the Netherlands are used to link to police records in BRON. The are a large number of new websites in the Netherlands. For this thesis, one news website is used to gain articles. This website, www.omroepwest.nl, is a Dutch news website that covers regional news in the north of the province Zuid-Holland (area Den Haag) [Omroep West, 2015]. The website contains hundreds of news articles about road accidents.

To collect the news articles from Omroep West, the open source web crawling software Scrapy is used (based on programming language Python) [Scrapy, 2015]. A web crawler for this website was made to collect news articles based on information in the title of the news article. Articles with road accident related words in the title are collected. In total, 2296 articles published between 2007 and 2015 are collected. From all these articles is the title, content and publication date and time collected.

The collected news articles can be compared to the police records in two ways. The information in a police record can be used to search information in a news article. Another method is to extract information from the news article and collect this into a record. This record can be compared with the police records. In this proof of concept, a combination of both methods is used.

8-3 Standardising, indexing and comparing record pairs

News articles about road accidents contain valuable information. This information is found in the headline, lead and body of the article. For a record linkage operation, the valuable information needs to be identified in the article. One can choose two approaches. First, make pairs with one article and one record out of the police database. The record in the police database can be used to search for information in the news article. If evidence is found in the news article for an attribute in the police record, then the comparison agrees. If the information is not found, then the comparison disagrees or is missing. In this thesis, another approach is chosen. The information in the news article is extracted and merged into a record. A pair of records is a pair with one police record and one news article record.

Each article is converted into a record. For example, the following interesting information for a linking can be found in news articles: the publication date and time of the article, the date and time of the road accident, the place of road accident, the street/location information, the mode of transport, the sex and so on. See Figure 8-1 for a news article with these information. Some of the information is easy to extract from the text. Some attributes are harder to extract from the text. A certain level of ‘understanding’ of the article is needed. In this thesis, the focus is not on a good extraction of the information in the article. The information is extracted in a effective manner and is subject to mistakes.

Five attributes of the road accident are used to link the police data with the news articles. For each of the variables, a global description is given about the way they are extracted from the article and how the attribute is compared with the police record. The following variables are used:

Figure 8-1: Screenshot of news article "Car and motorcyclist collide in Pijnacker" on www.omroepwest.nl [Omroep West, 2015]. The article contains information about the publication date and time, the vehicles involved, the time of the accident, the location, eye witness information, information about the casualties, the damage to the vehicles and a photo. Photo and text copyright Jonathan de Bruin.

Publication date and time Each of the articles has a publication date and time. This date and time variable is compared with the time of the road accident found in the police record. This attribute is compared in the same way as the date and time variable when linking police records with hospital records. This is because the news article is published after the road accident occurred. The article may be published 1 hour before the registered date and time of the accident to get an identical comparison. The publication date and time may not be more than 3 days after the date and time of the road accident according to the police. These values are based on experimenting.

Place name Each news article starts with the place name of the road accident (or the location of publication). This place name is compared with the place name of the road accident in police data.

Street name Extracting the location of the road accident in the news article is done by extracting the street name. A street name is not always easy to recognise. To find it in the text, a database with street names is made. This database is a collection of all street names found in the police data. Each news article is searched for these collected street names. After that, the street names are compared between the news records and the police records. Some standardisation steps were performed to make the comparison less susceptible to errors and misspellings.

Vehicles For most of the victims in police data, the mode of transport and the mode of transport of the road accident partner is available in BRON. In case of a single-vehicle conflict, the type of roadside object is documented in BRON. In each news article, all words are collected that relate to a mode of transport. All the modes of transport are collected. If the mode of transport in BRON is also found in the news article record, then the comparison is assumed to be identical.

Transport to hospital The article may contain information about the transport to the hospital. This information is often missing. This information the news article record is exactly compared with information in BRON.

A selection of records from the police database is made based on the place names found in the articles. This selection results in 14263 police records. The number of news articles is 2296. The total number of record pairs is 32747848. The number of record pairs is reduced to improve the performance of the estimation algorithm. The number of record pairs is reduced with the standard blocking method. The data is blocked on two blocking keys; the place name and the year. The place name in the article needs to be identical to the place name in the police data. The year of publication needs to be the same as the year of the road accident as reported in the police data. This reduction results in 85748 candidate record pairs (a reduction of 99.74%).

8-4 Linking police road accident records with news articles

The record pairs are classified into positive links, positive non-links and possible links with the generalised ECM-algorithm. Each comparison has 3 types of comparisons; agreement/disagreement/either missing with labels 2/1/0. This algorithm is used because the performance is good and there are many missing values in the data.

The estimation of parameters resulted in a link prevalence $P(M = 1) = 0.01725$. This estimate implies that there are 1479 links in the between the datasets. In Table 8-1, an overview is given of the comparison space and the estimated m- and u-probability mass functions. Also, the error levels μ and λ are given in the same way as in Table 6-1 (Chapter 6). If the comparison space is divided into the three action sets without random decisions, then 1149 record pairs are assigned with the positive link action, 483 record pairs with the possible link action and 84116 record pairs with the positive non-link action. The associated error levels are $\mu = 4.17e - 03$ and $\lambda = 3.76e - 01$.

The comparison vector $\boldsymbol{y} = (2, 2, 2, 2)$ has the largest weight of all vectors in the comparison space. This vector agrees on all comparison variables, so it is likely that this comparison vector indicates a large probability of being part of the true link set. Observe that transport to the hospital is not an informative variable. It is often missing and has only two possible values. The comparison variables 'date and time' and 'street name' are relatively informative variables. If the comparison 'date and time' is identical, then it is a strong indication of a positive link. This can be explained easily; the blocking keys, place name and year of the road accident, always agree. If also the date and time agree, then the place and the date and time of the road accident agree. Assuming that the number of road accidents in a small time span in a place is not large explains why this variable is so informative. For the street name, the same sort of argumentation can be used. The number of road accident in one year in the same street is (assumed to be) low. Therefore, the street name is informative for the classification.

A second result that needs to be studied is the number of links assigned as positive links, while the date and time recorded by the police differ more than 3 days from the time of publication. These news articles are not published shortly after the accident. It is strongly advisable to classify the comparison space based on a different blocking key. For example, by using the epoch for indexing. The differences between the two indexing methods are of interest.

To indicate the quality of the classification, 30 record pairs in action set I were randomly chosen and reviewed by hand. For 8 records pairs, the classification by hand resulted in the positive link action. For the other 22 record pairs, the classification by hand resulted in a positive non-link action. For most of the 22 misclassifications, there were months between the publication date and the road accident date. This indicates what was observed before; the epoch should be used for indexing. At least a casual relation should be applied to the record pairs. To indicate what happens when the epoch is restricted to 3 days, 30 record pairs with action I and this restriction are drawn randomly. A new manual classification shows that 25 record pairs are likely to be positive links, 3 record pairs are positive non-links and 2 record pairs are possible links. The classification is much better.

y^{street}	$y^{\text{m.o.t.}}$	y^{hosp}	y^{epoch}	$f(\boldsymbol{y})$	w	m	u	λ	μ
2	2	2	2	67	8.90	3.01e-02	4.13e-06	1.00e+00	4.13e-06
2	2	1	2	1	8.26	2.66e-04	6.87e-08	9.70e-01	4.20e-06
2	2	0	2	103	8.07	7.11e-02	2.22e-05	9.70e-01	2.64e-05
2	1	2	2	13	7.66	1.30e-02	6.09e-06	8.98e-01	3.25e-05
2	1	1	2	0	7.03	1.14e-04	1.01e-07	8.85e-01	3.26e-05
2	1	0	2	33	6.84	3.06e-02	3.28e-05	8.85e-01	6.54e-05
2	2	2	1	150	5.34	9.62e-02	4.61e-04	8.55e-01	5.26e-04
2	2	1	1	1	4.71	8.49e-04	7.66e-06	7.59e-01	5.34e-04
2	2	0	1	555	4.52	2.27e-01	2.48e-03	7.58e-01	3.01e-03
2	0	2	2	0	4.24	1.16e-04	1.67e-06	5.31e-01	3.01e-03
2	1	2	1	141	4.11	4.14e-02	6.79e-04	5.31e-01	3.69e-03
1	2	2	2	43	3.80	1.39e-02	3.11e-04	4.89e-01	4.00e-03
2	0	1	2	0	3.61	1.02e-06	2.77e-08	4.75e-01	4.00e-03
2	1	1	1	3	3.48	3.65e-04	1.13e-05	4.75e-01	4.02e-03
0	2	2	2	35	3.47	5.37e-03	1.68e-04	4.75e-01	4.18e-03
2	0	0	2	4	3.42	2.74e-04	8.96e-06	4.70e-01	4.19e-03
2	1	0	1	483	3.29	9.77e-02	3.65e-03	4.69e-01	7.84e-03
1	2	1	2	0	3.16	1.22e-04	5.17e-06	3.72e-01	7.85e-03
1	2	0	2	180	2.97	3.27e-02	1.67e-03	3.71e-01	9.52e-03
0	2	1	2	0	2.83	4.74e-05	2.79e-06	3.39e-01	9.53e-03
0	2	0	2	121	2.64	1.27e-02	9.02e-04	3.39e-01	1.04e-02
1	1	2	2	44	2.57	5.97e-03	4.58e-04	3.26e-01	1.09e-02
0	1	2	2	30	2.23	2.31e-03	2.47e-04	3.20e-01	1.11e-02
1	1	1	2	1	1.93	5.27e-05	7.63e-06	3.18e-01	1.11e-02
1	1	0	2	171	1.74	1.41e-02	2.47e-03	3.18e-01	1.36e-02
0	1	1	2	1	1.60	2.04e-05	4.11e-06	3.04e-01	1.36e-02
0	1	0	2	144	1.41	5.45e-03	1.33e-03	3.04e-01	1.49e-02
2	0	2	1	5	0.69	3.70e-04	1.86e-04	2.98e-01	1.51e-02
1	2	2	1	2739	0.24	4.43e-02	3.47e-02	2.98e-01	4.98e-02
2	0	1	1	0	0.06	3.27e-06	3.09e-06	2.54e-01	4.98e-02
0	2	2	1	1125	-0.09	1.71e-02	1.87e-02	2.54e-01	6.85e-02
2	0	0	1	65	-0.14	8.73e-04	1.00e-03	2.36e-01	6.95e-02
1	2	1	1	31	-0.39	3.91e-04	5.77e-04	2.35e-01	7.01e-02
1	2	0	1	17810	-0.58	1.04e-01	1.87e-01	2.35e-01	2.57e-01
0	2	1	1	1	-0.72	1.51e-04	3.11e-04	1.31e-01	2.57e-01
1	0	2	2	3	-0.85	5.34e-05	1.25e-04	1.30e-01	2.57e-01
0	2	0	1	7362	-0.91	4.04e-02	1.01e-01	1.30e-01	3.58e-01
1	1	2	1	5397	-0.99	1.90e-02	5.11e-02	9.00e-02	4.09e-01
0	0	2	2	8	-1.19	2.07e-05	6.76e-05	7.09e-02	4.09e-01
0	1	2	1	2784	-1.32	7.37e-03	2.76e-02	7.09e-02	4.37e-01
1	0	1	2	0	-1.49	4.71e-07	2.09e-06	6.35e-02	4.37e-01
1	1	1	1	144	-1.62	1.68e-04	8.50e-04	6.35e-02	4.37e-01
1	0	0	2	68	-1.68	1.26e-04	6.75e-04	6.34e-02	4.38e-01
1	1	0	1	20546	-1.81	4.49e-02	2.75e-01	6.33e-02	7.13e-01
0	0	1	2	0	-1.82	1.82e-07	1.13e-06	1.83e-02	7.13e-01
0	1	1	1	38	-1.95	6.51e-05	4.58e-04	1.83e-02	7.14e-01
0	0	0	2	32	-2.01	4.87e-05	3.64e-04	1.82e-02	7.14e-01
0	1	0	1	13641	-2.14	1.74e-02	1.48e-01	1.82e-02	8.62e-01
1	0	2	1	547	-4.41	1.70e-04	1.40e-02	7.96e-04	8.76e-01
0	0	2	1	482	-4.74	6.59e-05	7.54e-03	6.25e-04	8.84e-01
1	0	1	1	0	-5.04	1.50e-06	2.33e-04	5.59e-04	8.84e-01
1	0	0	1	6976	-5.23	4.02e-04	7.53e-02	5.58e-04	9.59e-01
0	0	1	1	2	-5.37	5.82e-07	1.25e-04	1.56e-04	9.59e-01
0	0	0	1	3625	-5.56	1.56e-04	4.06e-02	1.56e-04	1.00e+00

Table 8-1: The comparison space for a classification of news records with police records. The comparison space is monotone decreasing ordered on the weight.

Chapter 9

Conclusion and Discussion

9-1 Introduction

Record linkage is widely used for many practises where data needs to be linked between multiple sources. This thesis shows that the probabilistic record linkage framework by Fellegi and Sunter [1969] is useful for linking records between data sources. There are several methods to estimate parameters of the framework. The thesis shows that especially the ECM-algorithm is one of the effective algorithms to estimate parameters.

9-2 The Fellegi and Sunter model

The Fellegi and Sunter framework is an effective framework for classification for record linkage. The framework is built on linkage rules. Fellegi and Sunter provide an optimal and most discriminating linkage rule. This thesis showed that the framework is closely related to hypothesis testing. The formulation of the problem in terms of hypothesis testing is slightly different. With the Neyman-Pearson lemma and hypothesis testing, it was shown that the linkage rule is indeed the best linkage rule that can be achieved.

The simulation study in Section 6 showed that the Fellegi and Sunter framework is useful to link records between datasets. The number of misclassifications can be estimated accurately. It should be mentioned that this is related to the number of comparison variables and the quality of the data. Records of good quality data result in better classifications than records of poor quality. Also, the number of comparison variables is related to the accuracy of the classification. In general, more comparison variables lead to a more accurate classification.

In probabilistic record linkage, comparing record attributes is often performed under the binary assumption (see Binary assumption, Section 3-4-2). The simulation study showed that under the binary assumption, the classification is good. When the binary assumption is

not applied to the comparison vectors, multiple comparison types can be used. Comparison types such as agreement and the attribute is '...' . The simulation study showed that this makes it not directly easier to distinguish the distribution of weights of the comparison vector of the true links and the true non-links. Only a minor additional distinguishing power was observed.

9-3 Parameter estimation methods

In this thesis, several algorithms are described to estimate the parameters of the Fellegi and Sunter framework. The main focus was on unsupervised learning algorithms based on the EM-algorithm. The widely used ECM-algorithm is, according to the literature, a good method to estimate parameters. The simulation study showed that this algorithm is indeed a very good method to estimate parameters (see Estimation of parameters with the ECM-algorithm, Section 6-4-1). The algorithm was able to estimate the number of links in the dataset and the error levels very accurately. For different quality datasets, the algorithm estimated the parameters well. Also was observed that the ECM-algorithm is not very sensitive to the choice of starting parameters. For most of the starting parameters, the algorithm converges to the desired estimates. For extremely bad starting values, the algorithm classified the links as non-links and the non-links as links. This can easily be observed from the estimation result. The ECM-algorithm has properties that can lead to incorrect classifications. The simulation study shows that the link prevalence has to be larger than 0.01 to result in good estimates. If the link prevalence is less than 0.01, the estimates make no sense. The ECM-algorithm needs to train itself. Therefore, there is a reasonable proportion of true links (against candidate record pairs) needed.

An application of the EM-algorithm was the frequency-based EM-algorithm. This algorithm could handle multiple types of comparisons instead of agreement and disagreement. The simulation study (see Estimation of frequency based parameters with the EM-algorithm, Section 6-4-2), showed that the algorithm can be used for accurate classification. The classification results were similar with, or even slightly better than, the ECM-algorithm. The number of possible links is less while the error levels are nearly identical. Also, the algorithm showed good convergence results in the simulation study. The frequency based EM-algorithm has more starting parameters. Another advantage of the algorithm was the possibility to handle missing values. This thesis shows that it can be used to improve the classification if the data contain many missing values. Even if missing values are not equally distributed over the true links and non-links. This way of handling missing values was not seen before and deserves additional research. The relation between the (amount of) missing values and the classification is of interest.

If the decision maker decides to use the frequency based EM-algorithm, it is advisable to use also the ECM-algorithm. Some of the parameters need to be identical in both methods. Parameters such as the m- and u-probability mass functions for disagreement and the link prevalence π. If both algorithms return similar values for these parameters, it is a good indication that frequency based EM-algorithm converges to the right value.

The error levels μ and λ are of special interest in the frequency based EM-algorithm. The algorithm estimates the m- and u-probability mass functions for the entire comparison space.

This is because of the conditional independence assumption. Comparison vectors that do not occur in the (observed) comparison space still get a m- and u-probability mass function. Both, the frequency based EM-algorithm and the ECM-algorithm, estimate parameters for the full comparison space. For the ECM-algorithm, this is in most of the cases not of important, because all possible comparison vectors occur in the comparison space. For the frequency based EM-algorithm, not all the possible comparison vectors occur in the comparison space. This implies that the error levels are not calculated well(All m-probabilities of the comparison vectors in the comparison space do not sum up to one.). Further research is needed to study this behaviour and the accuracy of the error levels.

A last remark has to be made about the name of this algorithm. In this thesis, the algorithm is sometimes called the frequency based EM algorithm. This name comes from the formulation in Winkler [2000] (where a non-iterative estimation method was described, see also Section 5-3). In fact, this algorithm is just the EM-algorithm with some constraints to the likelihood (such as conditional independence). The frequency based EM-algorithm is the same algorithm as the ECM-algorithm, but it is more general.

The simulation study shows that the ECM-algorithm performs reasonably if the conditional independence assumption (see Conditional independence assumption, Section 3-4-1) is clearly violated. A simulation with a quite strong dependency between two comparison variables shows a small overestimation. The (Expectation-Maximisation) algorithm by Schürle could handle dependencies. This algorithm performs not better than the ECM-algorithm. In some situations even worse. Further research should reveal more about the behaviour of the ECM-algorithm on dependent data. It is interesting to know when the dependencies are small enough to result in good estimation and when they are too large for a good classification.

9-4 The role of indexing

In this thesis, it was observed that indexing of record pairs plays an important role in record linkage. It is used, among other reasons, to reduce the number of record pairs and therefore the computational resources. In Section 6-4-1, it was observed that it can be necessary for the accurate estimation of parameters. For the linking of police and hospital road accident records, indexing played a large role in the estimation of links in the set of possible links. Using a different indexing criterion reduces the estimated number of links drastically. For this thesis, the influence of indexing is studied in more detail than was described so far.

If a true link is not included in the candidate record pairs after indexing, it can no longer be linked. In general, the estimated number of links in the data is an underestimation of the number of true links in the entire set of record pairs. In some situations, the total number of links in the data is of interest, but indexing is required for some reason. It turns out that this might be possible. Consider a set of candidate record pairs $C \subset A \times B$. If the indexing method works perfectly, then all true links are included in this set.

An approach to estimate the total number of true links is to take a simple random sample of $C^c \subset A \times B$ (the complement of the set of candidate record pairs). The comparison vectors in this sample can be scaled to the number of comparison vectors in $C^c \subset A \times B$. This method seems to be a bit counter-intuitive, because the goal of indexing is to reduce

the number of comparisons. However, two datasets with 10^6 records each have 10^{12} record pairs. If an indexing method reduces the number of pairs to 10^9 pairs, it asks relatively less computational time to sample and compare 10^6 or more record pairs. This estimated distribution of the comparison space can be used to estimate the parameter derived with the ECM-algorithm.

More mathematically, consider a set of comparison vectors $\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_N$. The indices $1, \ldots, N^*$ refer to candidate comparison vectors and the incides $N^* + 1, \ldots, N$ to the non-candidate comparison vectors. The link prevalence can be expressed as a combination of both sets. Consider the link prevalence for $A \times B$

$$\begin{aligned}
\pi &= \frac{\sum_{j=1}^{N} P(M_j = 1 | \boldsymbol{Y}_j = \boldsymbol{y}_j)}{N} \\
&= \frac{\sum_{j=1}^{N^*} P(M_j = 1 | \boldsymbol{Y}_j = \boldsymbol{y}_j)}{N} + \frac{\sum_{j=N^*+1}^{N} P(M = 1 | \boldsymbol{Y}_j = \boldsymbol{y}_j)}{N} \\
&= \frac{N^* \pi^*}{N} + \frac{\sum_{j=N^*+1}^{N} P(M = 1 | \boldsymbol{Y}_j = \boldsymbol{y}_j)}{N}
\end{aligned}$$

where π^* is the link prevalence with indexed data. This first term is a scaling of the link prevalence π^* to π. The second term is the term which contains the remaining links. To estimate this term, if was considered that parameter estimates for the indexed data could be used to estimate the results. Some approaches lead to a quite accurate estimation of missed links, but the results need to be studied further.

9-5 Linking police and hospital accident records

In Chapter 7, police road accident records were linked with hospital road accident records. The record linkage was performed with the Fellegi and Sunter framework and the ECM-algorithm and frequency based EM-algorithm. The classification was made for each year between 2007 and 2013. The number of links shows a similar trend as for the record linkage by the SWOV. The ordering of the comparison vectors was well justifiable on the basis of manual review and data knowledge. Each classification had the same configuration of the comparison vector as a possible link. This comparison vector agreed on the area of the accident, the time of the accident, the sex and the mode of transport. Only the date of birth disagreed. This comparison vector occurred very often in the set of record pairs. These possible links were hard to split into two set for the ECM-algorithm. Therefore, the number of links in the data was, highly likely, overestimated. Clerical review shows that it is indeed likely that the number of links was overestimated.

Another indexing criterion was used to reduce the number of record pairs and thereby also the number of possible links. The record linkage showed significantly different results between two of the used indexing methods. In case the year of birth was used as blocking key, there were fewer possible links. Manual review showed that these results were reasonable. Also the number of estimated errors/ incorrect classifications were easier to justify. The procedure with month of birth or day of birth as blocking key was repeated. The results for the 3 types of blocking keys are slightly different, but they show the same behaviour. The estimated

number of links was slightly higher than the estimates of the SWOV. Over the years the results show the same trend.

Using additional distinguishing power, by using agreement on a certain attribute, showed not much additional distinguishing power. It also did not solve the problems with the large number of possible links (the pairs that agree on all attributes except the date of birth). The observed order of comparison vectors, based on the weight, showed that the ordering was justifiable. Therefore, it can be assumed that the classification was good. Nevertheless, the results did not show much advantage compared with the ECM-algorithm under the binary assumption or the agreement/disagreement/either missing assumption.

9-6 Linking additional data resources

Chapter 8 showed that it is possible to link other road accident related sources of information to police accident records. It was shown that it is good possible to link news articles found on the Internet with police records. The presented solution was a proof of concept. Many improvements can be made to improve the results. One of the bottlenecks is the extraction of information from the news article. For example, a better understanding of the location and the time of the accident. For example, the date and time of the road accident can be extracted from the text instead of using the publication time. The words 'yesterday evening' tell something about the date-time of the accident. Things like this need to be extracted from the text.

So far, the record linkage results showed poor results. A different indexing key might solve this problem. The record pairs are now blocked with the year and the place name of the road accident as blocking key. Using the time between the road accident and the publication can also be used for indexation, just like the indexation method to link the police and hospital data. It is recommended to apply criteria like this.

The question may arise: is the more advanced Fellegi and Sunter [1969] record linkage framework for a good classification? There are not many accidents in on one day and in particular city. Only date-time and city information might be enough for classification. If both attributes agree, then it is very likely that the record agrees with the news article. More research is needed to decide if the model of Fellegi and Sunter adds value to this linkage problem.

Glossary

binary assumption The comparison of attributes is restricted to agreement and disagreement with labels 1 and 0.

blocking Basic type of indexing for which the data are divided into mutually non-overlapping blocks that agree on a certain value.

classification The process of classifying the set of record pairs into two or three possible action sets. The positive link action set and the positive non-link action set. In some cases is a third set used; the possible link action set.

clerical review The process of classifying a pair of records as a match or non-match by a human reviewer.

comparison vector A vector for which each element represents the comparison of an attribute or field in the record pair.

conditional independence assumption The assumption that the attributes/fields of the record pairs are mutually independent given the match status.

data matching Identical procedure as record linkage, except that the pieces of information to link are not restricted to records.

deterministic record linkage An deterministic record linkage or classification method used to declare if record pairs belong to the same entity. The method is based on deterministic principles such as metrics.

distinguishing power The power of a classification method to distinguish the distribution of true links from the distribution of true non-links.

EM-algorithm An iterative statistical algorithm for maximisation of the likelihood function, while the likelihood function depends on unobserved latent variables. For probabilistic record linkage is the true match status an unobserved latent variable.

indexing A (simple) method to select a subset of record pairs such that most matches are contained in the subset. Those record pairs, candidate record pairs, are fully compared. Indexing is performed to reduce computational time or the memory usage.

linkage rule A linkage rule, or linkage function, is a function used by Fellegi and Sunter [1969] to map the comparison space into the probability of an action. The actions are the positive links action, the positive non-link action and the possible link action. The best linkage rule is the rule for which no other linkage rule performs better given the error levels. Linkage rules are closely related to decision rules.

pre-processing Process of cleaning and standardisation of datasets. This process can be useful to make the record pairs easier to compare.

privacy preserved record linkage A record linkage operation for which the privacy of the person is preserved. privacy preserved record linkage can be done by encrypting the (quasi-)identifier(s) in both datasets (by the owner of the dataset). Another method is that privacy preserved records are supplied and these records are linked.

probabilistic record linkage A probabilistic record linkage or classification method used to declare if record pairs belong to the same entity. The method is based on statistical and probabilistic principles. The framework of Fellegi and Sunter [1969] is the most popular probabilistic record linkage framework.

quasi-identifier A quasi-identifier is a non-identifying piece of information. The quasi-identifier itself is not enough to identify. A combination of quasi-identifiers can lead to identification.

randomised linkage rule A linkage rule for which the mapping from the comparison space into the probability of an action can be based on a random process.

record linkage The procedure of bringing together information from two or more records that are believed to relate to the same entity.

record pair A pair of records. A pairs of records can contain two records from the same dataset (deduplication) or two records from different datasets (record linkage).

similarity vector A vector for which each element represents the similarity between two values. The similarity is a value between 0 and 1, where 1 is used when the values are identical and 0 if they are completely different.

supervised learning The classification process is a learning process for which information is available about the true match statuses of the record pairs. This information is used to improve the classification.

unsupervised learning The classification process is a learning process for which no information is available about the true underlying true match statuses of the record pairs.

Appendix A

Upper bound for the u-probabilities mass functions

Consider the case of randomly picking a comparison vector $\boldsymbol{y}$ from the data. Define the probability as $f(\boldsymbol{y}) := P(\boldsymbol{Y} = \boldsymbol{y})$. The law of total probability implies

$$\begin{aligned} f(\boldsymbol{y}) &:= P(\boldsymbol{Y} = \boldsymbol{y}) \\ &= P(\boldsymbol{Y} = \boldsymbol{y}|M = 1)P(M = 1) + P(\boldsymbol{Y} = \boldsymbol{y}|M = 0)P(M = 0). \end{aligned} \tag{A-1}$$

In general, the probability of randomly picking a true link is much smaller than picking a true non-link. For a large dataset, the probability $P(M = 0)$ can be close to 1, while the the probability $P(M = 1)$ is close to 0. Therefore, an approximation of $f(\boldsymbol{y})$ is

$$f(\boldsymbol{y}) \approx P(\boldsymbol{Y} = \boldsymbol{y}|M = 0) = u(\boldsymbol{y}).$$

Back to Formula A-1, the difference between $f(\boldsymbol{y})$ and $u(\boldsymbol{y})$ is of interest. Rewrite Formula A-1 as

$$\begin{aligned} f(\boldsymbol{y}) &= P(\boldsymbol{Y} = \boldsymbol{y}|M = 1)P(M = 1) + P(\boldsymbol{Y} = \boldsymbol{y}|M = 0)P(M = 0) \\ &= m(\boldsymbol{y})P(M = 1) + u(\boldsymbol{y})(1 - P(M = 1)) \end{aligned}$$

Therefore,

$$f(\boldsymbol{y}) - u(\boldsymbol{y}) = P(M = 1) \cdot (u(\boldsymbol{y}) - m(\boldsymbol{y})) \tag{A-2}$$

This can be used to derive the following inequality

$$|f(\boldsymbol{y}) - u(\boldsymbol{y})| = P(M = 1) \cdot |u(\boldsymbol{y}) - m(\boldsymbol{y})| \leq P(M = 1) \tag{A-3}$$

In case of one-to-one linking, the probability that a pair of records belongs to the set of links $(M = 1)$ is bounded by the size of the datasets. Only one record of dataset A can link with

one record of dataset B and visa verse, then the number of links can not exceed the number of records of the smallest database. Therefore, Formula A-3 is bounded by

$$|f(\boldsymbol{y}) - u(\boldsymbol{y})| = \frac{\min(N_A, N_B)}{N_A \cdot N_B} \tag{A-4}$$

where N_A is the number of records in dataset A and N_B is the number of records in dataset B. If we define $f_i(y^i) = P(Y^i = y^i)$ and assume the conditional independence assumption (use Formula 3-30), then it is straightforward that

$$|\prod_{i=1}^{k} f_i(y^i) - \prod_{i=1}^{k} u_i(y^i)| \leq \frac{\min(N_A, N_B)}{N_A \cdot N_B}. \tag{A-5}$$

The goal is to find a upper bound for $|u(\boldsymbol{y}) - \prod_{i=1}^{k} u_i(y^i)|$. This bound is used to gain a difference between the u-probability mass functions and the u-marginal probability mass functions. The bound for $|u(\boldsymbol{y}) - \prod_{i=1}^{k} u_i(y^i)|$ is given by

$$|u(\boldsymbol{y}) - \prod_{i=1}^{k} u_i(y^i)| \leq |u(\boldsymbol{y}) - f(\boldsymbol{y}) + f(\boldsymbol{y}) - \prod_{i=1}^{k} f_i(y^i) + \prod_{i=1}^{k} f_i(y^i) - \prod_{i=1}^{k} u_i(y^i)| \tag{A-6}$$

$$\leq |u(\boldsymbol{y}) - f(\boldsymbol{y})| + |f(\boldsymbol{y}) - \prod_{i=1}^{k} f_i(y^i)| + |\prod_{i=1}^{k} f_i(y^i) - \prod_{i=1}^{k} u_i(y^i)|. \tag{A-7}$$

Two terms on the right hand side are bounded by inequalities (A-4) and (A-5). The term

$$\left| f(\boldsymbol{y}) - \prod_{i=1}^{k} f_i(y^i) \right|$$

can be approximated from file characteristics. The probabilities are estimated by

$$f(\boldsymbol{y}) \approx \frac{\#\text{ record pairs with } \boldsymbol{Y} = \boldsymbol{y}}{\#\text{ record pairs}} \tag{A-8}$$

and

$$f(y^i) \approx \frac{\#\text{ record pairs with } Y^i = y^i}{\#\text{ record pairs}}. \tag{A-9}$$

The bound for Formula A-6 is now

$$|u(\boldsymbol{y}) - \prod_{i=1}^{k} u_i(y^i)| \leq |f(\boldsymbol{y}) - \prod_{i=1}^{k} f_i(y^i)| + \frac{2\min(N_A, N_B)}{N_A \cdot N_B}. \tag{A-10}$$

Appendix B

Data quality versus the number of comparison variables

See the simulation study in Chapter 6 for more details, especially Section 6-3.

Dataset	K		I	II	III	μ_{sim}	μ_{est}	λ_{sim}	λ_{est}	F_{score}
Good	6	mean	440.17	167.41	999392.42	9.96e-05	9.98e-05	2.51e-01	2.50e-01	0.751
		std	42.60	106.60	109.98	4.64e-05	4.72e-05	9.20e-02	9.03e-02	0.092
		min	318.00	0.00	998469.00	1.10e-05	1.29e-05	3.40e-02	3.39e-02	0.481
		max	500.00	1035.00	999500.00	2.52e-04	2.73e-04	5.20e-01	4.76e-01	0.965
	8	mean	471.75	60.22	999468.03	7.81e-05	7.87e-05	1.76e-01	1.76e-01	0.824
		std	30.07	57.67	43.86	4.24e-05	4.29e-05	9.78e-02	9.76e-02	0.099
		min	327.00	0.00	999051.00	8.00e-06	1.01e-05	1.80e-02	2.01e-02	0.441
		max	500.00	471.00	999500.00	2.54e-04	2.68e-04	5.56e-01	5.34e-01	0.983
	10	mean	494.72	10.50	999494.79	4.65e-05	4.54e-05	9.64e-02	9.38e-02	0.904
		std	9.87	16.57	10.24	3.05e-05	3.19e-05	6.45e-02	6.59e-02	0.065
		min	411.00	0.00	999391.00	3.00e-06	9.93e-07	6.00e-03	2.48e-03	0.453
		max	500.00	129.00	999500.00	2.48e-04	2.62e-04	5.34e-01	5.37e-01	0.994

Table B-1: This table gives the result of 1000 classifications with 'good' quality comparison variables. The column K is the number of comparison variables used for classification. Observe that the quality of the classification is better when there are more variables used (see the F-score).

Dataset	K		I	II	III	μ_{sim}	μ_{est}	λ_{sim}	λ_{est}	F_{score}
Low	6	mean	407.04	221.32	999371.64	2.26e-04	2.26e-04	5.77e-01	5.78e-01	0.408
		std	68.66	127.40	125.00	6.20e-05	6.31e-05	1.24e-01	1.22e-01	0.130
		min	191.00	0.00	998658.00	8.40e-05	8.04e-05	1.94e-01	1.96e-01	0.173
		max	500.00	878.00	999500.00	4.10e-04	4.41e-04	8.20e-01	7.95e-01	0.808
	8	mean	469.69	60.87	999469.43	2.22e-04	2.23e-04	4.85e-01	4.85e-01	0.512
		std	37.70	64.43	42.81	5.84e-05	5.99e-05	1.44e-01	1.43e-01	0.148
		min	229.00	0.00	999038.00	7.50e-05	8.51e-05	1.50e-01	1.62e-01	0.124
		max	500.00	557.00	999500.00	4.07e-04	4.17e-04	8.58e-01	8.63e-01	0.850
	10	mean	494.33	11.21	999494.46	1.79e-04	1.73e-04	3.64e-01	3.62e-01	0.636
		std	8.23	13.89	8.74	6.40e-05	7.03e-05	1.34e-01	1.35e-01	0.134
		min	435.00	0.00	999387.00	4.40e-05	1.90e-05	8.80e-02	7.07e-02	0.144
		max	500.00	115.00	999500.00	4.28e-04	4.45e-04	8.36e-01	8.55e-01	0.912
Poor	6	mean	405.74	224.30	999369.96	3.08e-04	3.09e-04	7.69e-01	7.69e-01	0.215
		std	69.65	127.89	125.44	6.10e-05	6.33e-05	1.02e-01	1.01e-01	0.109
		min	174.00	0.00	998741.00	1.42e-04	1.45e-04	4.00e-01	4.16e-01	0.046
		max	500.00	854.00	999500.00	4.54e-04	4.82e-04	9.28e-01	9.08e-01	0.619
	8	mean	473.06	55.24	999471.70	3.32e-04	3.33e-04	7.07e-01	7.07e-01	0.289
		std	33.92	52.50	35.41	5.32e-05	5.59e-05	1.25e-01	1.24e-01	0.128
		min	242.00	0.00	999142.00	1.67e-04	1.73e-04	3.44e-01	3.62e-01	0.020
		max	500.00	442.00	999500.00	4.70e-04	4.87e-04	9.62e-01	9.56e-01	0.658
	10	mean	494.91	10.19	999494.90	3.03e-04	2.94e-04	6.13e-01	6.12e-01	0.386
		std	6.86	10.70	6.32	6.59e-05	7.68e-05	1.37e-01	1.37e-01	0.137
		min	415.00	0.00	999448.00	1.13e-04	7.52e-05	2.28e-01	2.32e-01	0.032
		max	500.00	108.00	999500.00	4.50e-04	4.83e-04	9.60e-01	9.61e-01	0.773
SWOV	6	mean	410.24	196.35	999393.41	1.97e-04	1.97e-04	4.89e-01	4.88e-01	0.485
		std	67.10	117.01	107.66	6.06e-05	6.09e-05	1.36e-01	1.34e-01	0.148
		min	187.00	0.00	998615.00	4.90e-05	4.64e-05	1.04e-01	1.20e-01	0.166
		max	500.00	903.00	999500.00	4.00e-04	3.95e-04	7.94e-01	7.83e-01	0.890
	8	mean	479.18	43.07	999477.75	2.00e-04	2.00e-04	4.20e-01	4.20e-01	0.575
		std	21.23	33.56	24.69	5.68e-05	5.82e-05	1.26e-01	1.26e-01	0.130
		min	294.00	0.00	999315.00	5.30e-05	4.84e-05	1.02e-01	1.04e-01	0.186
		max	500.00	234.00	999500.00	3.80e-04	3.95e-04	7.82e-01	8.05e-01	0.896
	10	mean	495.72	8.49	999495.80	1.72e-04	1.69e-04	3.48e-01	3.45e-01	0.651
		std	4.90	7.51	4.69	5.51e-05	5.85e-05	1.13e-01	1.14e-01	0.113
		min	462.00	0.00	999468.00	4.30e-05	3.68e-05	8.60e-02	6.84e-02	0.201
		max	500.00	50.00	999500.00	3.91e-04	4.05e-04	7.92e-01	7.86e-01	0.914

Table B-2: This table gives the result of 1000 classifications with 'poor' quality comparison variables, the result of 1000 classifications with 'low' quality comparison variables and the result of 1000 classifications with 'swov' related comparison variables. The column K is the number of comparison variables used for classification. Observe that the quality of the classification is better when there are more variables used (see the F-score).

Appendix C

Results of linking BRON and LMR

C-1 Estimation with the ECM-algorithm

y^{hosp}	y^{epoch}	y^{gender}	y^{mot}	y^{dob}	$f(\boldsymbol{y})$	$w(\boldsymbol{y})$	$m(\boldsymbol{y})$	$u(\boldsymbol{y})$	λ	μ
2	2	2	2	2	4348	16.69	4.48e-01	2.52e-08	1.00e+00	2.52e-08
2	2	2	1	2	2025	13.94	2.08e-01	1.83e-07	5.52e-01	2.08e-07
2	2	1	2	2	38	12.19	4.49e-03	2.28e-08	3.45e-01	2.31e-07
2	1	2	2	2	140	11.24	1.40e-02	1.84e-07	3.40e-01	4.14e-07
1	2	2	2	2	89	10.39	8.94e-03	2.74e-07	3.26e-01	6.89e-07
2	2	1	1	2	27	9.44	2.08e-03	1.65e-07	3.17e-01	8.54e-07
2	1	2	1	2	87	8.49	6.50e-03	1.33e-06	3.15e-01	2.18e-06
1	2	2	1	2	73	7.64	4.15e-03	1.99e-06	3.08e-01	4.17e-06
2	1	1	2	2	6	6.74	1.41e-04	1.66e-07	3.04e-01	4.34e-06
1	2	1	2	2	5	5.89	8.97e-05	2.48e-07	3.04e-01	4.58e-06
2	2	2	2	1	14451	5.71	1.95e-01	6.49e-04	3.04e-01	6.54e-04
1	1	2	2	2	89	4.94	2.80e-04	2.00e-06	1.09e-01	6.56e-04
2	1	1	1	2	21	4.00	6.52e-05	1.20e-06	1.09e-01	6.57e-04
1	2	1	1	2	35	3.15	4.16e-05	1.79e-06	1.09e-01	6.59e-04
2	2	2	1	1	91373	2.96	9.06e-02	4.70e-03	1.09e-01	5.36e-03
1	1	2	1	2	259	2.20	1.30e-04	1.45e-05	1.80e-02	5.37e-03
2	2	1	2	1	12810	1.21	1.96e-03	5.86e-04	1.78e-02	5.96e-03
1	1	1	2	2	49	0.44	2.81e-06	1.80e-06	1.59e-02	5.96e-03
2	1	2	2	1	87920	0.26	6.11e-03	4.73e-03	1.59e-02	1.07e-02
1	2	2	2	1	147766	-0.59	3.90e-03	7.06e-03	9.78e-03	1.77e-02
2	2	1	1	1	83831	-1.54	9.09e-04	4.24e-03	5.88e-03	2.20e-02
1	1	1	1	2	221	-2.30	1.30e-06	1.31e-05	4.97e-03	2.20e-02
2	1	2	1	1	655470	-2.49	2.83e-03	3.42e-02	4.97e-03	5.62e-02
1	2	2	1	1	975480	-3.34	1.81e-03	5.11e-02	2.14e-03	1.07e-01
2	1	1	2	1	80704	-4.24	6.13e-05	4.27e-03	3.27e-04	1.12e-01
1	2	1	2	1	131501	-5.09	3.91e-05	6.37e-03	2.66e-04	1.18e-01
1	1	2	2	1	991982	-6.04	1.22e-04	5.14e-02	2.27e-04	1.69e-01
2	1	1	1	1	605212	-6.99	2.84e-05	3.09e-02	1.05e-04	2.00e-01
1	2	1	1	1	881447	-7.84	1.82e-05	4.61e-02	7.66e-05	2.46e-01
1	1	2	1	1	7199813	-8.79	5.66e-05	3.72e-01	5.84e-05	6.18e-01
1	1	1	2	1	882939	-10.54	1.22e-06	4.64e-02	1.79e-06	6.64e-01
1	1	1	1	1	6497637	-13.29	5.68e-07	3.36e-01	5.68e-07	1.00e+00

Table C-1: The result of a classification with the binary assumption for road accidents in 2007. The comparison vectors are ordered on the weight. The green comparison vectors are classified as positive links, the orange comparison vectors as possible links and the red comparison vectors as positive non-links. The estimated number of links N_M is 9706.

y^{hosp}	y^{epoch}	y^{gender}	y^{mot}	y^{dob}	$f(\boldsymbol{y})$	$w(\boldsymbol{y})$	$m(\boldsymbol{y})$	$u(\boldsymbol{y})$	λ	μ
2	2	2	2	2	4458	16.60	4.68e-01	2.90e-08	1.00e+00	2.90e-08
2	2	2	1	2	1828	13.81	1.94e-01	1.95e-07	5.32e-01	2.24e-07
2	2	1	2	2	42	12.23	5.40e-03	2.63e-08	3.39e-01	2.50e-07
2	1	2	2	2	143	11.20	1.53e-02	2.09e-07	3.33e-01	4.59e-07
1	2	2	2	2	82	10.33	9.08e-03	2.96e-07	3.18e-01	7.55e-07
2	2	1	1	2	31	9.44	2.23e-03	1.77e-07	3.09e-01	9.32e-07
2	1	2	1	2	94	8.41	6.33e-03	1.41e-06	3.07e-01	2.34e-06
1	2	2	1	2	79	7.54	3.76e-03	1.99e-06	3.00e-01	4.33e-06
2	1	1	2	2	6	6.84	1.76e-04	1.90e-07	2.97e-01	4.52e-06
1	2	1	2	2	12	5.97	1.05e-04	2.68e-07	2.96e-01	4.79e-06
2	2	2	2	1	15153	5.59	1.96e-01	7.37e-04	2.96e-01	7.41e-04
1	1	2	2	2	65	4.93	2.97e-04	2.14e-06	9.98e-02	7.43e-04
2	1	1	1	2	27	4.05	7.30e-05	1.28e-06	9.95e-02	7.45e-04
1	2	1	1	2	39	3.18	4.33e-05	1.81e-06	9.95e-02	7.47e-04
2	2	2	1	1	89784	2.80	8.13e-02	4.96e-03	9.94e-02	5.70e-03
1	1	2	1	2	235	2.15	1.23e-04	1.44e-05	1.81e-02	5.72e-03
2	2	1	2	1	13272	1.22	2.27e-03	6.68e-04	1.80e-02	6.38e-03
1	1	1	2	2	44	0.57	3.42e-06	1.94e-06	1.58e-02	6.39e-03
2	1	2	2	1	92678	0.19	6.42e-03	5.31e-03	1.58e-02	1.17e-02
1	2	2	2	1	146182	-0.68	3.81e-03	7.53e-03	9.33e-03	1.92e-02
2	2	1	1	1	82369	-1.57	9.37e-04	4.49e-03	5.52e-03	2.37e-02
1	1	1	1	2	208	-2.22	1.42e-06	1.30e-05	4.58e-03	2.37e-02
2	1	2	1	1	641212	-2.60	2.66e-03	3.57e-02	4.58e-03	5.95e-02
1	2	2	1	1	903707	-3.47	1.58e-03	5.06e-02	1.92e-03	1.10e-01
2	1	1	2	1	84077	-4.18	7.40e-05	4.82e-03	3.45e-04	1.15e-01
1	2	1	2	1	128846	-5.04	4.40e-05	6.82e-03	2.71e-04	1.22e-01
1	1	2	2	1	981977	-6.08	1.25e-04	5.43e-02	2.27e-04	1.76e-01
2	1	1	1	1	592078	-6.96	3.06e-05	3.24e-02	1.02e-04	2.08e-01
1	2	1	1	1	819194	-7.83	1.82e-05	4.59e-02	7.18e-05	2.54e-01
1	1	2	1	1	6592519	-8.87	5.16e-05	3.65e-01	5.36e-05	6.20e-01
1	1	1	2	1	873806	-10.44	1.44e-06	4.92e-02	2.03e-06	6.69e-01
1	1	1	1	1	5982938	-13.23	5.95e-07	3.31e-01	5.95e-07	1.00e+00

Table C-2: The result of a classification with the binary assumption for road accidents in 2008. The comparison vectors are ordered on the weight. The green comparison vectors are classified as positive links, the orange comparison vectors as possible links and the red comparison vectors as positive non-links. The estimated number of links N_M is 9489.

y^{hosp}	y^{epoch}	y^{gender}	y^{mot}	y^{dob}	$f(\boldsymbol{y})$	$w(\boldsymbol{y})$	$m(\boldsymbol{y})$	$u(\boldsymbol{y})$	λ	μ
2	2	2	2	2	4091	16.58	4.99e-01	3.14e-08	1.00e+00	3.14e-08
2	2	2	1	2	1627	13.84	1.98e-01	1.93e-07	5.01e-01	2.24e-07
2	2	1	2	2	166	13.39	2.00e-02	3.06e-08	3.03e-01	2.55e-07
2	1	2	2	2	105	11.02	1.39e-02	2.27e-07	2.83e-01	4.81e-07
1	2	2	2	2	112	10.72	1.34e-02	2.97e-07	2.69e-01	7.79e-07
2	2	1	1	2	61	10.65	7.94e-03	1.88e-07	2.56e-01	9.66e-07
2	1	2	1	2	80	8.29	5.51e-03	1.39e-06	2.48e-01	2.36e-06
1	2	2	1	2	70	7.98	5.35e-03	1.82e-06	2.42e-01	4.18e-06
2	1	1	2	2	11	7.83	5.55e-04	2.21e-07	2.37e-01	4.40e-06
1	2	1	2	2	8	7.53	5.38e-04	2.90e-07	2.36e-01	4.69e-06
2	2	2	2	1	13956	5.22	1.53e-01	8.28e-04	2.36e-01	8.32e-04
1	1	2	2	2	69	5.16	3.74e-04	2.14e-06	8.28e-02	8.34e-04
2	1	1	1	2	16	5.09	2.21e-04	1.35e-06	8.24e-02	8.36e-04
1	2	1	1	2	17	4.79	2.14e-04	1.78e-06	8.22e-02	8.37e-04
2	2	2	1	1	78614	2.48	6.09e-02	5.07e-03	8.19e-02	5.91e-03
1	1	2	1	2	196	2.43	1.49e-04	1.31e-05	2.11e-02	5.92e-03
2	2	1	2	1	13532	2.03	6.13e-03	8.07e-04	2.09e-02	6.73e-03
1	1	1	2	2	51	1.97	1.50e-05	2.09e-06	1.48e-02	6.73e-03
2	1	2	2	1	88043	-0.34	4.25e-03	5.96e-03	1.48e-02	1.27e-02
1	2	2	2	1	128092	-0.64	4.13e-03	7.83e-03	1.05e-02	2.05e-02
2	2	1	1	1	77370	-0.71	2.44e-03	4.94e-03	6.41e-03	2.55e-02
1	1	1	1	2	170	-0.77	5.95e-06	1.28e-05	3.97e-03	2.55e-02
2	1	2	1	1	561438	-3.07	1.69e-03	3.66e-02	3.97e-03	6.20e-02
1	2	2	1	1	732210	-3.38	1.64e-03	4.80e-02	2.28e-03	1.10e-01
2	1	1	2	1	87096	-3.53	1.70e-04	5.81e-03	6.36e-04	1.16e-01
1	2	1	2	1	124001	-3.83	1.65e-04	7.63e-03	4.65e-04	1.23e-01
1	1	2	2	1	869056	-6.20	1.15e-04	5.64e-02	3.00e-04	1.80e-01
2	1	1	1	1	555450	-6.27	6.78e-05	3.56e-02	1.85e-04	2.15e-01
1	2	1	1	1	711697	-6.57	6.57e-05	4.68e-02	1.18e-04	2.62e-01
1	1	2	1	1	5334303	-8.93	4.56e-05	3.46e-01	5.20e-05	6.08e-01
1	1	1	2	1	839065	-9.39	4.59e-06	5.50e-02	6.42e-06	6.63e-01
1	1	1	1	1	5198537	-12.13	1.83e-06	3.37e-01	1.83e-06	1.00e+00

Table C-3: The result of a classification with the binary assumption for road accidents in 2009. The comparison vectors are ordered on the weight. The green comparison vectors are classified as positive links, the orange comparison vectors as possible links and the red comparison vectors as positive non-links. The estimated number of links N_M is 8186.

y^{hosp}	y^{epoch}	y^{gender}	y^{mot}	y^{dob}	$f(\boldsymbol{y})$	$w(\boldsymbol{y})$	$m(\boldsymbol{y})$	$u(\boldsymbol{y})$	λ	μ
2	2	2	2	2	2872	16.41	4.22e-01	3.14e-08	1.00e+00	3.14e-08
2	2	2	1	2	1293	13.80	1.90e-01	1.94e-07	5.78e-01	2.25e-07
2	2	1	2	2	16	11.63	3.20e-03	2.84e-08	3.88e-01	2.53e-07
2	1	2	2	2	104	11.06	1.45e-02	2.28e-07	3.84e-01	4.81e-07
1	2	2	2	2	61	10.29	8.76e-03	2.97e-07	3.70e-01	7.78e-07
2	2	1	1	2	16	9.01	1.44e-03	1.75e-07	3.61e-01	9.54e-07
2	1	2	1	2	48	8.45	6.54e-03	1.40e-06	3.60e-01	2.36e-06
1	2	2	1	2	39	7.67	3.95e-03	1.83e-06	3.53e-01	4.19e-06
2	1	1	2	2	2	6.28	1.10e-04	2.06e-07	3.49e-01	4.40e-06
2	2	2	2	1	9019	5.59	2.26e-01	8.47e-04	3.49e-01	8.51e-04
1	2	1	2	2	3	5.51	6.63e-05	2.70e-07	1.23e-01	8.52e-04
1	1	2	2	2	47	4.94	3.01e-04	2.16e-06	1.23e-01	8.54e-04
2	1	1	1	2	13	3.66	4.95e-05	1.27e-06	1.23e-01	8.55e-04
2	2	2	1	1	47746	2.97	1.02e-01	5.22e-03	1.23e-01	6.08e-03
1	2	1	1	2	13	2.89	2.99e-05	1.66e-06	2.11e-02	6.08e-03
1	1	2	1	2	103	2.32	1.36e-04	1.33e-05	2.10e-02	6.09e-03
2	2	1	2	1	7962	0.80	1.71e-03	7.67e-04	2.09e-02	6.86e-03
2	1	2	2	1	56343	0.23	7.75e-03	6.14e-03	1.92e-02	1.30e-02
1	1	1	2	2	26	0.15	2.28e-06	1.95e-06	1.14e-02	1.30e-02
1	2	2	2	1	74721	-0.54	4.68e-03	8.03e-03	1.14e-02	2.10e-02
2	2	1	1	1	43233	-1.81	7.72e-04	4.73e-03	6.75e-03	2.58e-02
2	1	2	1	1	333092	-2.38	3.50e-03	3.79e-02	5.98e-03	6.36e-02
1	1	1	1	2	96	-2.46	1.03e-06	1.20e-05	2.49e-03	6.36e-02
1	2	2	1	1	438204	-3.15	2.11e-03	4.95e-02	2.48e-03	1.13e-01
2	1	1	2	1	50691	-4.55	5.87e-05	5.56e-03	3.72e-04	1.19e-01
1	2	1	2	1	66297	-5.32	3.55e-05	7.27e-03	3.13e-04	1.26e-01
1	1	2	2	1	518840	-5.89	1.61e-04	5.82e-02	2.78e-04	1.84e-01
2	1	1	1	1	305436	-7.17	2.65e-05	3.43e-02	1.17e-04	2.18e-01
1	2	1	1	1	396286	-7.94	1.60e-05	4.49e-02	9.03e-05	2.63e-01
1	1	2	1	1	3203126	-8.51	7.25e-05	3.59e-01	7.43e-05	6.22e-01
1	1	1	2	1	462115	-10.68	1.22e-06	5.27e-02	1.77e-06	6.75e-01
1	1	1	1	1	2907475	-13.29	5.50e-07	3.25e-01	5.50e-07	1.00e+00

Table C-4: The result of a classification with the binary assumption for road accidents in 2010. The comparison vectors are ordered on the weight. The green comparison vectors are classified as positive links, the orange comparison vectors as possible links and the red comparison vectors as positive non-links. The estimated number of links N_M is 6786.

y^{hosp}	y^{epoch}	y^{gender}	y^{mot}	y^{dob}	$f(\boldsymbol{y})$	$w(\boldsymbol{y})$	$m(\boldsymbol{y})$	$u(\boldsymbol{y})$	λ	μ
2	2	2	2	2	1555	16.51	4.61e-01	3.11e-08	1.00e+00	3.11e-08
2	2	2	1	2	602	13.81	1.81e-01	1.81e-07	5.39e-01	2.12e-07
2	2	1	2	2	6	11.14	1.92e-03	2.80e-08	3.59e-01	2.40e-07
2	1	2	2	2	28	10.73	1.02e-02	2.23e-07	3.57e-01	4.62e-07
1	2	2	2	2	35	10.60	1.15e-02	2.89e-07	3.47e-01	7.51e-07
2	2	1	1	2	3	8.44	7.53e-04	1.63e-07	3.35e-01	9.14e-07
2	1	2	1	2	27	8.04	4.00e-03	1.29e-06	3.34e-01	2.21e-06
1	2	2	1	2	27	7.90	4.52e-03	1.68e-06	3.30e-01	3.88e-06
2	2	2	2	1	4617	5.49	2.22e-01	9.21e-04	3.26e-01	9.25e-04
2	1	1	2	2	1	5.36	4.26e-05	2.00e-07	1.04e-01	9.26e-04
1	2	1	2	2	2	5.22	4.81e-05	2.60e-07	1.04e-01	9.26e-04
1	1	2	2	2	17	4.82	2.55e-04	2.07e-06	1.04e-01	9.28e-04
2	2	2	1	1	22401	2.79	8.70e-02	5.35e-03	1.03e-01	6.28e-03
2	1	1	1	2	7	2.66	1.67e-05	1.16e-06	1.62e-02	6.28e-03
1	2	1	1	2	6	2.52	1.88e-05	1.51e-06	1.62e-02	6.28e-03
1	1	2	1	2	50	2.12	1.00e-04	1.20e-05	1.62e-02	6.29e-03
2	2	1	2	1	4138	0.11	9.26e-04	8.29e-04	1.61e-02	7.12e-03
2	1	2	2	1	28666	-0.29	4.92e-03	6.60e-03	1.52e-02	1.37e-02
1	2	2	2	1	37290	-0.43	5.56e-03	8.55e-03	1.03e-02	2.23e-02
1	1	1	2	2	6	-0.56	1.07e-06	1.86e-06	4.70e-03	2.23e-02
2	2	1	1	1	20250	-2.59	3.63e-04	4.82e-03	4.70e-03	2.71e-02
2	1	2	1	1	158465	-2.99	1.93e-03	3.83e-02	4.34e-03	6.54e-02
1	2	2	1	1	206543	-3.13	2.18e-03	4.97e-02	2.41e-03	1.15e-01
1	1	1	1	2	35	-3.25	4.17e-07	1.08e-05	2.33e-04	1.15e-01
2	1	1	2	1	25665	-5.67	2.05e-05	5.94e-03	2.33e-04	1.21e-01
1	2	1	2	1	33482	-5.81	2.32e-05	7.69e-03	2.12e-04	1.29e-01
1	1	2	2	1	254739	-6.21	1.23e-04	6.12e-02	1.89e-04	1.90e-01
2	1	1	1	1	144253	-8.36	8.04e-06	3.45e-02	6.61e-05	2.24e-01
1	2	1	1	1	185292	-8.50	9.08e-06	4.47e-02	5.80e-05	2.69e-01
1	1	2	1	1	1492517	-8.91	4.82e-05	3.56e-01	4.90e-05	6.25e-01
1	1	1	2	1	227193	-11.58	5.13e-07	5.51e-02	7.14e-07	6.80e-01
1	1	1	1	1	1342860	-14.28	2.01e-07	3.20e-01	2.01e-07	1.00e+00

Table C-5: The result of a classification with the binary assumption for road accidents in 2011. The comparison vectors are ordered on the weight. The green comparison vectors are classified as positive links, the orange comparison vectors as possible links and the red comparison vectors as positive non-links. The estimated number of links N_M is 3357.

y^{hosp}	y^{epoch}	y^{gender}	y^{mot}	y^{dob}	$f(\boldsymbol{y})$	$w(\boldsymbol{y})$	$m(\boldsymbol{y})$	$u(\boldsymbol{y})$	λ	μ
2	2	2	2	2	1422	16.25	4.47e-01	3.92e-08	1.00e+00	3.92e-08
2	2	2	1	2	605	13.74	1.93e-01	2.08e-07	5.53e-01	2.47e-07
2	2	1	2	2	5	10.71	1.57e-03	3.49e-08	3.60e-01	2.82e-07
2	1	2	2	2	40	10.70	1.29e-02	2.90e-07	3.58e-01	5.72e-07
1	2	2	2	2	23	10.08	8.51e-03	3.56e-07	3.45e-01	9.28e-07
2	2	1	1	2	2	8.20	6.76e-04	1.85e-07	3.37e-01	1.11e-06
2	1	2	1	2	24	8.19	5.55e-03	1.54e-06	3.36e-01	2.65e-06
1	2	2	1	2	24	7.57	3.67e-03	1.89e-06	3.30e-01	4.54e-06
2	2	2	2	1	4019	5.39	2.17e-01	9.93e-04	3.27e-01	9.98e-04
2	1	1	2	2	1	5.16	4.51e-05	2.58e-07	1.10e-01	9.98e-04
1	2	1	2	2	4	4.55	2.98e-05	3.17e-07	1.10e-01	9.98e-04
1	1	2	2	2	16	4.53	2.45e-04	2.63e-06	1.10e-01	1.00e-03
2	2	2	1	1	18201	2.88	9.35e-02	5.27e-03	1.10e-01	6.27e-03
2	1	1	1	2	5	2.65	1.94e-05	1.37e-06	1.63e-02	6.28e-03
1	2	1	1	2	5	2.04	1.29e-05	1.68e-06	1.62e-02	6.28e-03
1	1	2	1	2	34	2.02	1.06e-04	1.39e-05	1.62e-02	6.29e-03
2	2	1	2	1	3485	-0.15	7.59e-04	8.84e-04	1.61e-02	7.17e-03
2	1	2	2	1	25171	-0.16	6.23e-03	7.33e-03	1.54e-02	1.45e-02
1	2	2	2	1	31378	-0.78	4.12e-03	9.01e-03	9.14e-03	2.35e-02
1	1	1	2	2	11	-1.00	8.57e-07	2.34e-06	5.02e-03	2.35e-02
2	2	1	1	1	16232	-2.66	3.28e-04	4.69e-03	5.02e-03	2.82e-02
2	1	2	1	1	129421	-2.67	2.69e-03	3.89e-02	4.69e-03	6.71e-02
1	2	2	1	1	159749	-3.29	1.78e-03	4.78e-02	2.00e-03	1.15e-01
1	1	1	1	2	42	-3.51	3.70e-07	1.24e-05	2.23e-04	1.15e-01
2	1	1	2	1	22592	-5.70	2.18e-05	6.52e-03	2.22e-04	1.22e-01
1	2	1	2	1	28237	-6.32	1.44e-05	8.02e-03	2.00e-04	1.30e-01
1	1	2	2	1	222574	-6.33	1.19e-04	6.65e-02	1.86e-04	1.96e-01
2	1	1	1	1	116810	-8.21	9.42e-06	3.46e-02	6.74e-05	2.31e-01
1	2	1	1	1	142108	-8.83	6.23e-06	4.26e-02	5.80e-05	2.73e-01
1	1	2	1	1	1195882	-8.84	5.11e-05	3.53e-01	5.17e-05	6.27e-01
1	1	1	2	1	197912	-11.87	4.15e-07	5.92e-02	5.94e-07	6.86e-01
1	1	1	1	1	1060874	-14.38	1.79e-07	3.14e-01	1.79e-07	1.00e+00

Table C-6: The result of a classification with the binary assumption for road accidents in 2012. The comparison vectors are ordered on the weight. The green comparison vectors are classified as positive links, the orange comparison vectors as possible links and the red comparison vectors as positive non-links. The estimated number of links N_M is 3161.

y^{hosp}	y^{epoch}	y^{gender}	y^{mot}	y^{dob}	$f(\boldsymbol{y})$	$w(\boldsymbol{y})$	$m(\boldsymbol{y})$	$u(\boldsymbol{y})$	λ	μ
2	2	2	2	2	1430	16.59	3.33e-01	2.07e-08	1.00e+00	2.07e-08
2	2	2	1	2	835	14.09	1.96e-01	1.49e-07	6.67e-01	1.70e-07
1	2	2	2	2	103	11.81	3.07e-02	2.29e-07	4.71e-01	3.99e-07
2	2	1	2	2	5	11.20	1.39e-03	1.90e-08	4.40e-01	4.18e-07
2	1	2	2	2	43	11.18	1.10e-02	1.53e-07	4.38e-01	5.71e-07
1	2	2	1	2	108	9.30	1.81e-02	1.65e-06	4.27e-01	2.22e-06
2	2	1	1	2	4	8.69	8.17e-04	1.37e-07	4.09e-01	2.36e-06
2	1	2	1	2	30	8.68	6.47e-03	1.10e-06	4.09e-01	3.46e-06
1	2	1	2	2	1	6.41	1.28e-04	2.11e-07	4.02e-01	3.67e-06
1	1	2	2	2	25	6.39	1.01e-03	1.69e-06	4.02e-01	5.36e-06
2	2	2	2	1	4187	5.87	2.22e-01	6.29e-04	4.01e-01	6.34e-04
2	1	1	2	2	1	5.78	4.57e-05	1.41e-07	1.79e-01	6.35e-04
1	2	1	1	2	8	3.91	7.54e-05	1.51e-06	1.79e-01	6.36e-04
1	1	2	1	2	61	3.89	5.97e-04	1.22e-05	1.79e-01	6.48e-04
2	2	2	1	1	22962	3.37	1.31e-01	4.52e-03	1.78e-01	5.17e-03
2	1	1	1	2	6	3.28	2.69e-05	1.01e-06	4.70e-02	5.17e-03
1	2	2	2	1	37694	1.08	2.05e-02	6.96e-03	4.70e-02	1.21e-02
1	1	1	2	2	11	1.00	4.22e-06	1.56e-06	2.65e-02	1.21e-02
2	2	1	2	1	3911	0.47	9.26e-04	5.77e-04	2.65e-02	1.27e-02
2	1	2	2	1	26512	0.45	7.33e-03	4.65e-03	2.55e-02	1.74e-02
1	2	2	1	1	253205	-1.42	1.21e-02	5.00e-02	1.82e-02	6.74e-02
1	1	1	1	2	48	-1.50	2.49e-06	1.12e-05	6.13e-03	6.74e-02
2	2	1	1	1	21081	-2.03	5.45e-04	4.15e-03	6.12e-03	7.16e-02
2	1	2	1	1	166067	-2.05	4.32e-03	3.34e-02	5.58e-03	1.05e-01
1	2	1	2	1	34140	-4.31	8.54e-05	6.39e-03	1.26e-03	1.11e-01
1	1	2	2	1	259249	-4.33	6.76e-04	5.14e-02	1.18e-03	1.63e-01
2	1	1	2	1	24421	-4.94	3.05e-05	4.27e-03	5.01e-04	1.67e-01
1	2	1	1	1	232874	-6.82	5.03e-05	4.59e-02	4.71e-04	2.13e-01
1	1	2	1	1	1893011	-6.83	3.98e-04	3.70e-01	4.21e-04	5.83e-01
2	1	1	1	1	155602	-7.44	1.80e-05	3.07e-02	2.24e-05	6.13e-01
1	1	1	2	1	234143	-9.73	2.81e-06	4.72e-02	4.47e-06	6.61e-01
1	1	1	1	1	1736880	-12.23	1.66e-06	3.39e-01	1.66e-06	1.00e+00

Table C-7: The result of a classification with the binary assumption for road accidents in 2013. The comparison vectors are ordered on the weight. The green comparison vectors are classified as positive links, the orange comparison vectors as possible links and the red comparison vectors as positive non-links. The estimated number of links N_M is 4252.

C-2 Estimation with the ECM-algorithm and data blocked on the year of birth

y^{hosp}	y^{epoch}	y^{gender}	y^{mot}	y^{dob}	$f(\boldsymbol{y})$	$w(\boldsymbol{y})$	$m(\boldsymbol{y})$	$u(\boldsymbol{y})$	λ	μ
2	2	2	2	2	4348	12.29	6.22e-01	2.87e-06	1.00e+00	2.87e-06
2	2	2	1	2	2025	9.98	2.90e-01	1.34e-05	3.78e-01	1.63e-05
2	2	1	2	2	38	7.80	5.94e-03	2.43e-06	8.75e-02	1.87e-05
2	1	2	2	2	140	6.83	1.93e-02	2.09e-05	8.16e-02	3.96e-05
1	2	2	2	2	89	5.93	1.16e-02	3.08e-05	6.23e-02	7.04e-05
2	2	1	1	2	27	5.50	2.77e-03	1.14e-05	5.07e-02	8.18e-05
2	1	2	1	2	87	4.52	9.00e-03	9.76e-05	4.79e-02	1.79e-04
1	2	2	1	2	73	3.63	5.41e-03	1.44e-04	3.89e-02	3.23e-04
2	2	2	2	1	395	3.07	2.09e-02	9.76e-04	3.35e-02	1.30e-03
2	1	1	2	2	6	2.34	1.84e-04	1.78e-05	1.26e-02	1.32e-03
1	2	1	2	2	5	1.44	1.11e-04	2.62e-05	1.24e-02	1.34e-03
2	2	2	1	1	1279	0.76	9.76e-03	4.56e-03	1.23e-02	5.90e-03
1	1	2	2	2	89	0.47	3.60e-04	2.25e-04	2.52e-03	6.12e-03
2	1	1	1	2	21	0.04	8.59e-05	8.29e-05	2.16e-03	6.21e-03
1	2	1	1	2	35	-0.86	5.17e-05	1.22e-04	2.07e-03	6.33e-03
2	2	1	2	1	239	-1.42	2.00e-04	8.29e-04	2.02e-03	7.16e-03
1	1	2	1	2	259	-1.83	1.68e-04	1.05e-03	1.82e-03	8.21e-03
2	1	2	2	1	1753	-2.39	6.49e-04	7.12e-03	1.66e-03	1.53e-02
1	2	2	2	1	3018	-3.29	3.91e-04	1.05e-02	1.01e-03	2.58e-02
2	2	1	1	1	1048	-3.73	9.32e-05	3.87e-03	6.16e-04	2.97e-02
1	1	1	2	2	49	-4.02	3.44e-06	1.91e-04	5.22e-04	2.99e-02
2	1	2	1	1	8647	-4.70	3.03e-04	3.32e-02	5.19e-04	6.31e-02
1	2	2	1	1	12405	-5.59	1.82e-04	4.90e-02	2.16e-04	1.12e-01
1	1	1	1	2	221	-6.32	1.60e-06	8.91e-04	3.41e-05	1.13e-01
2	1	1	2	1	1508	-6.88	6.20e-06	6.04e-03	3.25e-05	1.19e-01
1	2	1	2	1	2513	-7.78	3.73e-06	8.92e-03	2.63e-05	1.28e-01
1	1	2	2	1	19716	-8.75	1.21e-05	7.66e-02	2.26e-05	2.05e-01
2	1	1	1	1	7380	-9.19	2.89e-06	2.82e-02	1.04e-05	2.33e-01
1	2	1	1	1	10552	-10.08	1.74e-06	4.16e-02	7.56e-06	2.74e-01
1	1	2	1	1	93024	-11.05	5.65e-06	3.57e-01	5.82e-06	6.32e-01
1	1	1	2	1	16673	-13.24	1.16e-07	6.50e-02	1.70e-07	6.97e-01
1	1	1	1	1	78989	-15.54	5.39e-08	3.03e-01	5.39e-08	1.00e+00

Table C-8: The result of a classification with the binary assumption for road accidents in 2007. The candidate record pairs agree on the year of birth of the road casualty. The comparison vectors are ordered on the weight. The green comparison vectors are classified as positive links, the orange comparison vectors as possible links and the red comparison vectors as positive non-links. The estimated number of links N_M is 6978.

y^{hosp}	y^{epoch}	y^{gender}	y^{mot}	y^{dob}	$f(\boldsymbol{y})$	$w(\boldsymbol{y})$	$m(\boldsymbol{y})$	$u(\boldsymbol{y})$	λ	μ
2	2	2	2	2	4458	12.12	6.43e-01	3.51e-06	1.00e+00	3.51e-06
2	2	2	1	2	1828	9.82	2.67e-01	1.46e-05	3.57e-01	1.81e-05
2	2	1	2	2	42	7.79	7.09e-03	2.94e-06	8.99e-02	2.10e-05
2	1	2	2	2	143	6.72	2.08e-02	2.52e-05	8.29e-02	4.62e-05
1	2	2	2	2	82	5.81	1.16e-02	3.48e-05	6.21e-02	8.10e-05
2	2	1	1	2	31	5.48	2.94e-03	1.22e-05	5.05e-02	9.32e-05
2	1	2	1	2	94	4.41	8.62e-03	1.04e-04	4.75e-02	1.98e-04
1	2	2	1	2	79	3.51	4.82e-03	1.44e-04	3.89e-02	3.42e-04
2	2	2	2	1	441	2.92	2.20e-02	1.18e-03	3.41e-02	1.52e-03
2	1	1	2	2	6	2.38	2.29e-04	2.11e-05	1.21e-02	1.55e-03
1	2	1	2	2	12	1.48	1.28e-04	2.91e-05	1.19e-02	1.57e-03
2	2	2	1	1	1232	0.62	9.12e-03	4.90e-03	1.17e-02	6.48e-03
1	1	2	2	2	65	0.41	3.75e-04	2.50e-04	2.63e-03	6.73e-03
2	1	1	1	2	27	0.08	9.50e-05	8.75e-05	2.25e-03	6.82e-03
1	2	1	1	2	39	-0.82	5.32e-05	1.21e-04	2.16e-03	6.94e-03
2	2	1	2	1	276	-1.41	2.42e-04	9.90e-04	2.11e-03	7.93e-03
1	1	2	1	2	235	-1.89	1.56e-04	1.04e-03	1.86e-03	8.96e-03
2	1	2	2	1	1971	-2.48	7.09e-04	8.47e-03	1.71e-03	1.74e-02
1	2	2	2	1	3191	-3.38	3.97e-04	1.17e-02	9.98e-04	2.92e-02
2	2	1	1	1	1005	-3.71	1.00e-04	4.11e-03	6.01e-04	3.33e-02
1	1	1	2	2	44	-3.92	4.14e-06	2.09e-04	5.01e-04	3.35e-02
2	1	2	1	1	8560	-4.78	2.94e-04	3.52e-02	4.97e-04	6.86e-02
1	2	2	1	1	11664	-5.69	1.65e-04	4.86e-02	2.02e-04	1.17e-01
1	1	1	1	2	208	-6.22	1.72e-06	8.67e-04	3.73e-05	1.18e-01
2	1	1	2	1	1638	-6.81	7.82e-06	7.10e-03	3.56e-05	1.25e-01
1	2	1	2	1	2456	-7.71	4.38e-06	9.81e-03	2.78e-05	1.35e-01
1	1	2	2	1	20473	-8.79	1.28e-05	8.40e-02	2.34e-05	2.19e-01
2	1	1	1	1	7290	-9.11	3.24e-06	2.95e-02	1.06e-05	2.48e-01
1	2	1	1	1	9580	-10.02	1.82e-06	4.07e-02	7.34e-06	2.89e-01
1	1	2	1	1	84500	-11.09	5.32e-06	3.49e-01	5.52e-06	6.38e-01
1	1	1	2	1	16748	-13.12	1.41e-07	7.03e-02	2.00e-07	7.08e-01
1	1	1	1	1	71392	-15.42	5.87e-08	2.92e-01	5.87e-08	1.00e+00

Table C-9: The result of a classification with the binary assumption for road accidents in 2008. The candidate record pairs agree on the year of birth of the road casualty. The comparison vectors are ordered on the weight. The green comparison vectors are classified as positive links, the orange comparison vectors as possible links and the red comparison vectors as positive non-links. The estimated number of links N_M is 6901.

y^{hosp}	y^{epoch}	y^{gender}	y^{mot}	y^{dob}	$f(\boldsymbol{y})$	$w(\boldsymbol{y})$	$m(\boldsymbol{y})$	$u(\boldsymbol{y})$	λ	μ
2	2	2	2	2	4091	12.05	6.31e-01	3.67e-06	1.00e+00	3.67e-06
2	2	2	1	2	1627	9.80	2.53e-01	1.39e-05	3.69e-01	1.76e-05
2	2	1	2	2	166	8.91	2.47e-02	3.35e-06	1.16e-01	2.10e-05
2	2	1	1	2	61	6.66	9.89e-03	1.27e-05	9.14e-02	3.37e-05
2	1	2	2	2	105	6.47	1.72e-02	2.66e-05	8.16e-02	6.03e-05
1	2	2	2	2	112	6.16	1.62e-02	3.42e-05	6.44e-02	9.44e-05
2	1	2	1	2	80	4.22	6.86e-03	1.01e-04	4.82e-02	1.95e-04
1	2	2	1	2	70	3.91	6.49e-03	1.30e-04	4.13e-02	3.25e-04
2	1	1	2	2	11	3.32	6.72e-04	2.42e-05	3.48e-02	3.49e-04
1	2	1	2	2	8	3.02	6.36e-04	3.12e-05	3.41e-02	3.81e-04
2	2	2	2	1	386	2.80	2.11e-02	1.28e-03	3.35e-02	1.66e-03
2	1	1	1	2	16	1.07	2.69e-04	9.20e-05	1.24e-02	1.75e-03
1	2	1	1	2	17	0.77	2.54e-04	1.18e-04	1.21e-02	1.87e-03
1	1	2	2	2	69	0.58	4.41e-04	2.47e-04	1.19e-02	2.12e-03
2	2	2	1	1	1150	0.55	8.44e-03	4.86e-03	1.14e-02	6.98e-03
2	2	1	2	1	301	-0.35	8.26e-04	1.17e-03	3.00e-03	8.15e-03
1	1	2	1	2	196	-1.67	1.77e-04	9.39e-04	2.18e-03	9.09e-03
1	1	1	2	2	51	-2.57	1.73e-05	2.26e-04	2.00e-03	9.31e-03
2	2	1	1	1	937	-2.60	3.31e-04	4.43e-03	1.98e-03	1.37e-02
2	1	2	2	1	1844	-2.78	5.73e-04	9.27e-03	1.65e-03	2.30e-02
1	2	2	2	1	2655	-3.09	5.42e-04	1.19e-02	1.08e-03	3.49e-02
1	1	1	1	2	170	-4.82	6.91e-06	8.56e-04	5.36e-04	3.58e-02
2	1	2	1	1	7319	-5.03	2.29e-04	3.52e-02	5.29e-04	7.10e-02
1	2	2	1	1	9130	-5.34	2.17e-04	4.52e-02	3.00e-04	1.16e-01
2	1	1	2	1	1689	-5.93	2.25e-05	8.45e-03	8.26e-05	1.25e-01
1	2	1	2	1	2383	-6.24	2.12e-05	1.09e-02	6.02e-05	1.36e-01
2	1	1	1	1	6682	-8.18	8.98e-06	3.21e-02	3.89e-05	1.68e-01
1	2	1	1	1	8423	-8.49	8.50e-06	4.12e-02	2.99e-05	2.09e-01
1	1	2	2	1	17854	-8.67	1.47e-05	8.63e-02	2.14e-05	2.95e-01
1	1	2	1	1	68265	-10.92	5.90e-06	3.28e-01	6.71e-06	6.23e-01
1	1	1	2	1	16230	-11.82	5.77e-07	7.87e-02	8.08e-07	7.01e-01
1	1	1	1	1	62215	-14.07	2.31e-07	2.99e-01	2.31e-07	1.00e+00

Table C-10: The result of a classification with the binary assumption for road accidents in 2009. The candidate record pairs agree on the year of birth of the road casualty. The comparison vectors are ordered on the weight. The green comparison vectors are classified as positive links, the orange comparison vectors as possible links and the red comparison vectors as positive non-links. The estimated number of links N_M is 6464.

y^{hosp}	y^{epoch}	y^{gender}	y^{mot}	y^{dob}	$f(\boldsymbol{y})$	$w(\boldsymbol{y})$	$m(\boldsymbol{y})$	$u(\boldsymbol{y})$	λ	μ
2	2	2	2	2	2872	11.99	6.24e-01	3.86e-06	1.00e+00	3.86e-06
2	2	2	1	2	1293	9.86	2.80e-01	1.46e-05	3.76e-01	1.85e-05
2	2	1	2	2	16	7.20	4.42e-03	3.30e-06	9.61e-02	2.18e-05
2	1	2	2	2	104	6.63	2.08e-02	2.74e-05	9.17e-02	4.92e-05
1	2	2	2	2	61	5.85	1.22e-02	3.54e-05	7.09e-02	8.46e-05
2	2	1	1	2	16	5.06	1.98e-03	1.25e-05	5.87e-02	9.71e-05
2	1	2	1	2	48	4.49	9.31e-03	1.04e-04	5.67e-02	2.01e-04
1	2	2	1	2	39	3.71	5.48e-03	1.34e-04	4.74e-02	3.35e-04
2	2	2	2	1	284	2.98	2.67e-02	1.36e-03	4.19e-02	1.70e-03
2	1	1	2	2	2	1.84	1.47e-04	2.34e-05	1.52e-02	1.72e-03
1	2	1	2	2	3	1.05	8.66e-05	3.02e-05	1.51e-02	1.75e-03
2	2	2	1	1	665	0.84	1.20e-02	5.16e-03	1.50e-02	6.91e-03
1	1	2	2	2	47	0.48	4.07e-04	2.51e-04	3.05e-03	7.16e-03
2	1	1	1	2	13	-0.30	6.59e-05	8.89e-05	2.65e-03	7.25e-03
1	2	1	1	2	13	-1.08	3.88e-05	1.15e-04	2.58e-03	7.37e-03
1	1	2	1	2	103	-1.65	1.82e-04	9.54e-04	2.54e-03	8.32e-03
2	2	1	2	1	191	-1.82	1.89e-04	1.16e-03	2.36e-03	9.48e-03
2	1	2	2	1	1181	-2.39	8.87e-04	9.66e-03	2.17e-03	1.91e-02
1	2	2	2	1	1643	-3.17	5.22e-04	1.25e-02	1.28e-03	3.16e-02
2	2	1	1	1	550	-3.95	8.46e-05	4.41e-03	7.61e-04	3.60e-02
1	1	1	2	2	26	-4.31	2.88e-06	2.15e-04	6.76e-04	3.62e-02
2	1	2	1	1	4280	-4.52	3.98e-04	3.67e-02	6.73e-04	7.29e-02
1	2	2	1	1	5450	-5.31	2.34e-04	4.74e-02	2.75e-04	1.20e-01
1	1	1	1	2	96	-6.45	1.29e-06	8.15e-04	4.11e-05	1.21e-01
2	1	1	2	1	1045	-7.18	6.28e-06	8.25e-03	3.98e-05	1.29e-01
1	2	1	2	1	1348	-7.97	3.70e-06	1.07e-02	3.35e-05	1.40e-01
1	1	2	2	1	10403	-8.54	1.74e-05	8.86e-02	2.98e-05	2.29e-01
2	1	1	1	1	3656	-9.32	2.82e-06	3.13e-02	1.24e-05	2.60e-01
1	2	1	1	1	4697	-10.10	1.66e-06	4.05e-02	9.63e-06	3.00e-01
1	1	2	1	1	40265	-10.67	7.79e-06	3.36e-01	7.97e-06	6.37e-01
1	1	1	2	1	8757	-13.33	1.23e-07	7.57e-02	1.78e-07	7.13e-01
1	1	1	1	1	34434	-15.47	5.52e-08	2.87e-01	5.52e-08	1.00e+00

Table C-11: The result of a classification with the binary assumption for road accidents in 2010. The candidate record pairs agree on the year of birth of the road casualty. The comparison vectors are ordered on the weight. The green comparison vectors are classified as positive links, the orange comparison vectors as possible links and the red comparison vectors as positive non-links. The estimated number of links N_M is 4604.

y^{hosp}	y^{epoch}	y^{gender}	y^{mot}	y^{dob}	$f(\boldsymbol{y})$	$w(\boldsymbol{y})$	$m(\boldsymbol{y})$	$u(\boldsymbol{y})$	λ	μ
2	2	2	2	2	1555	12.16	6.56e-01	3.44e-06	1.00e+00	3.44e-06
2	2	2	1	2	602	9.83	2.60e-01	1.40e-05	3.44e-01	1.74e-05
2	2	1	2	2	6	6.79	2.55e-03	2.87e-06	8.41e-02	2.03e-05
2	1	2	2	2	28	6.35	1.41e-02	2.46e-05	8.15e-02	4.49e-05
1	2	2	2	2	35	6.24	1.59e-02	3.11e-05	6.75e-02	7.60e-05
2	2	1	1	2	3	4.46	1.01e-03	1.17e-05	5.15e-02	8.77e-05
2	1	2	1	2	27	4.02	5.58e-03	1.00e-04	5.05e-02	1.88e-04
1	2	2	1	2	27	3.91	6.32e-03	1.26e-04	4.49e-02	3.14e-04
2	2	2	2	1	129	2.98	2.59e-02	1.31e-03	3.86e-02	1.63e-03
2	1	1	2	2	1	0.98	5.47e-05	2.06e-05	1.27e-02	1.65e-03
1	2	1	2	2	2	0.87	6.19e-05	2.60e-05	1.27e-02	1.67e-03
2	2	2	1	1	333	0.66	1.03e-02	5.33e-03	1.26e-02	7.00e-03
1	1	2	2	2	17	0.43	3.42e-04	2.22e-04	2.35e-03	7.22e-03
2	1	1	1	2	7	-1.35	2.17e-05	8.36e-05	2.00e-03	7.31e-03
1	2	1	1	2	6	-1.46	2.46e-05	1.06e-04	1.98e-03	7.41e-03
1	1	2	1	2	50	-1.90	1.36e-04	9.04e-04	1.96e-03	8.32e-03
2	2	1	2	1	91	-2.39	1.01e-04	1.10e-03	1.82e-03	9.41e-03
2	1	2	2	1	532	-2.83	5.55e-04	9.39e-03	1.72e-03	1.88e-02
1	2	2	2	1	726	-2.94	6.29e-04	1.18e-02	1.17e-03	3.06e-02
2	2	1	1	1	242	-4.72	3.99e-05	4.45e-03	5.37e-04	3.51e-02
1	1	1	2	2	6	-4.94	1.33e-06	1.86e-04	4.97e-04	3.53e-02
2	1	2	1	1	1992	-5.15	2.20e-04	3.81e-02	4.96e-04	7.34e-02
1	2	2	1	1	2545	-5.26	2.49e-04	4.81e-02	2.75e-04	1.22e-01
1	1	1	1	2	35	-7.27	5.27e-07	7.55e-04	2.59e-05	1.22e-01
2	1	1	2	1	420	-8.20	2.16e-06	7.84e-03	2.53e-05	1.30e-01
1	2	1	2	1	560	-8.31	2.44e-06	9.90e-03	2.32e-05	1.40e-01
1	1	2	2	1	4466	-8.75	1.35e-05	8.48e-02	2.07e-05	2.25e-01
2	1	1	1	1	1779	-10.53	8.55e-07	3.19e-02	7.25e-06	2.57e-01
1	2	1	1	1	2142	-10.63	9.69e-07	4.02e-02	6.39e-06	2.97e-01
1	1	2	1	1	19056	-11.07	5.35e-06	3.44e-01	5.42e-06	6.41e-01
1	1	1	2	1	3905	-14.12	5.24e-08	7.08e-02	7.32e-08	7.12e-01
1	1	1	1	1	15689	-16.44	2.08e-08	2.88e-01	2.08e-08	1.00e+00

Table C-12: The result of a classification with the binary assumption for road accidents in 2011. The candidate record pairs agree on the year of birth of the road casualty. The comparison vectors are ordered on the weight. The green comparison vectors are classified as positive links, the orange comparison vectors as possible links and the red comparison vectors as positive non-links. The estimated number of links N_M is 2353.

y^{hosp}	y^{epoch}	y^{gender}	y^{mot}	y^{dob}	$f(\boldsymbol{y})$	$w(\boldsymbol{y})$	$m(\boldsymbol{y})$	$u(\boldsymbol{y})$	λ	μ
2	2	2	2	2	1422	11.93	6.43e-01	4.25e-06	1.00e+00	4.25e-06
2	2	2	1	2	605	9.67	2.76e-01	1.74e-05	3.57e-01	2.17e-05
2	1	2	2	2	40	6.40	1.82e-02	3.03e-05	8.07e-02	5.20e-05
2	2	1	2	2	5	6.38	2.10e-03	3.57e-06	6.26e-02	5.55e-05
1	2	2	2	2	23	5.75	1.18e-02	3.75e-05	6.05e-02	9.31e-05
2	1	2	1	2	24	4.14	7.81e-03	1.24e-04	4.87e-02	2.17e-04
2	2	1	1	2	2	4.12	9.04e-04	1.46e-05	4.08e-02	2.32e-04
1	2	2	1	2	24	3.50	5.08e-03	1.54e-04	3.99e-02	3.86e-04
2	2	2	2	1	106	2.84	2.28e-02	1.33e-03	3.49e-02	1.71e-03
2	1	1	2	2	1	0.85	5.95e-05	2.54e-05	1.21e-02	1.74e-03
2	2	2	1	1	252	0.59	9.80e-03	5.44e-03	1.20e-02	7.18e-03
1	1	2	2	2	16	0.22	3.34e-04	2.68e-04	2.21e-03	7.45e-03
1	2	1	2	2	4	0.20	3.87e-05	3.15e-05	1.87e-03	7.48e-03
2	1	1	1	2	5	-1.41	2.55e-05	1.04e-04	1.84e-03	7.59e-03
1	1	2	1	2	34	-2.04	1.44e-04	1.10e-03	1.81e-03	8.69e-03
1	2	1	1	2	5	-2.05	1.66e-05	1.29e-04	1.67e-03	8.82e-03
2	1	2	2	1	398	-2.69	6.44e-04	9.47e-03	1.65e-03	1.83e-02
2	2	1	2	1	83	-2.70	7.46e-05	1.11e-03	1.01e-03	1.94e-02
1	2	2	2	1	540	-3.33	4.19e-04	1.17e-02	9.32e-04	3.11e-02
2	1	2	1	1	1617	-4.94	2.77e-04	3.89e-02	5.13e-04	7.00e-02
2	2	1	1	1	186	-4.96	3.20e-05	4.57e-03	2.36e-04	7.46e-02
1	1	1	2	2	11	-5.33	1.09e-06	2.25e-04	2.04e-04	7.48e-02
1	2	2	1	1	2001	-5.59	1.80e-04	4.81e-02	2.02e-04	1.23e-01
1	1	1	1	2	42	-7.58	4.70e-07	9.22e-04	2.24e-05	1.24e-01
2	1	1	2	1	341	-8.24	2.11e-06	7.95e-03	2.20e-05	1.32e-01
1	1	2	2	1	3440	-8.86	1.18e-05	8.37e-02	1.99e-05	2.15e-01
1	2	1	2	1	417	-8.88	1.37e-06	9.84e-03	8.01e-06	2.25e-01
2	1	1	1	1	1377	-10.49	9.06e-07	3.26e-02	6.64e-06	2.58e-01
1	1	2	1	1	14649	-11.12	5.09e-06	3.43e-01	5.73e-06	6.01e-01
1	2	1	1	1	1666	-11.14	5.89e-07	4.04e-02	6.44e-07	6.42e-01
1	1	1	2	1	2985	-14.41	3.87e-08	7.03e-02	5.54e-08	7.12e-01
1	1	1	1	1	12176	-16.67	1.67e-08	2.88e-01	1.67e-08	1.00e+00

Table C-13: The result of a classification with the binary assumption for road accidents in 2012. The candidate record pairs agree on the year of birth of the road casualty. The comparison vectors are ordered on the weight. The green comparison vectors are classified as positive links, the orange comparison vectors as possible links and the red comparison vectors as positive non-links. The estimated number of links N_M is 2203.

y^{hosp}	y^{epoch}	y^{gender}	y^{mot}	y^{dob}	$f(\boldsymbol{y})$	$w(\boldsymbol{y})$	$m(\boldsymbol{y})$	$u(\boldsymbol{y})$	λ	μ
2	2	2	2	2	1430	12.37	5.38e-01	2.28e-06	1.00e+00	2.28e-06
2	2	2	1	2	835	10.16	3.26e-01	1.26e-05	4.62e-01	1.49e-05
1	2	2	2	2	103	7.55	4.78e-02	2.51e-05	1.36e-01	4.00e-05
2	2	1	2	2	5	6.89	1.92e-03	1.96e-06	8.83e-02	4.19e-05
2	1	2	2	2	43	6.88	1.64e-02	1.69e-05	8.63e-02	5.88e-05
1	2	2	1	2	108	5.34	2.90e-02	1.39e-04	7.00e-02	1.98e-04
2	2	1	1	2	4	4.68	1.17e-03	1.09e-05	4.10e-02	2.09e-04
2	1	2	1	2	30	4.67	9.95e-03	9.35e-05	3.98e-02	3.02e-04
2	2	2	2	1	88	2.92	1.50e-02	8.11e-04	2.99e-02	1.11e-03
1	2	1	2	2	1	2.07	1.71e-04	2.15e-05	1.49e-02	1.13e-03
1	1	2	2	2	25	2.06	1.46e-03	1.86e-04	1.47e-02	1.32e-03
2	1	1	2	2	1	1.40	5.85e-05	1.45e-05	1.32e-02	1.34e-03
2	2	2	1	1	305	0.71	9.10e-03	4.50e-03	1.32e-02	5.83e-03
1	2	1	1	2	8	-0.14	1.04e-04	1.19e-04	4.07e-03	5.95e-03
1	1	2	1	2	61	-0.15	8.84e-04	1.03e-03	3.96e-03	6.98e-03
2	1	1	1	2	6	-0.82	3.55e-05	8.04e-05	3.08e-03	7.06e-03
1	2	2	2	1	552	-1.90	1.33e-03	8.93e-03	3.05e-03	1.60e-02
2	2	1	2	1	66	-2.57	5.36e-05	6.97e-04	1.71e-03	1.67e-02
2	1	2	2	1	407	-2.58	4.57e-04	6.01e-03	1.66e-03	2.27e-02
1	1	1	2	2	11	-3.42	5.20e-06	1.60e-04	1.20e-03	2.29e-02
1	2	2	1	1	3036	-4.11	8.09e-04	4.95e-02	1.20e-03	7.24e-02
2	2	1	1	1	259	-4.78	3.25e-05	3.87e-03	3.89e-04	7.62e-02
2	1	2	1	1	1998	-4.79	2.78e-04	3.33e-02	3.56e-04	1.10e-01
1	1	1	1	2	48	-5.64	3.16e-06	8.85e-04	7.89e-05	1.10e-01
1	2	1	2	1	503	-7.39	4.76e-06	7.68e-03	7.58e-05	1.18e-01
1	1	2	2	1	3984	-7.40	4.06e-05	6.62e-02	7.10e-05	1.84e-01
2	1	1	2	1	353	-8.06	1.63e-06	5.17e-03	3.04e-05	1.89e-01
1	2	1	1	1	2552	-9.60	2.89e-06	4.26e-02	2.88e-05	2.32e-01
1	1	2	1	1	22685	-9.61	2.46e-05	3.67e-01	2.59e-05	5.99e-01
2	1	1	1	1	1698	-10.27	9.91e-07	2.87e-02	1.22e-06	6.28e-01
1	1	1	2	1	3456	-12.88	1.45e-07	5.69e-02	2.33e-07	6.85e-01
1	1	1	1	1	19478	-15.09	8.80e-08	3.15e-01	8.80e-08	1.00e+00

Table C-14: The result of a classification with the binary assumption for road accidents in 2010. The candidate record pairs agree on the year of birth of the road casualty. The comparison vectors are ordered on the weight. The green comparison vectors are classified as positive links, the orange comparison vectors as possible links and the red comparison vectors as positive non-links. The estimated number of links N_M is 2617.

C-3 Estimation with the EM-algorithm including missing values

y^{hosp}	y^{epoch}	y^{gender}	y^{mot}	y^{dob}	$f(\boldsymbol{y})$	$w(\boldsymbol{y})$	$m(\boldsymbol{y})$	$u(\boldsymbol{y})$	λ	μ
2	2	2	2	2	4348	16.69	4.47e-01	2.53e-08	1.00e+00	2.53e-08
2	2	0	2	2	29	15.77	3.26e-03	4.62e-10	5.53e-01	2.58e-08
2	2	2	1	2	1495	14.16	1.55e-01	1.09e-07	5.50e-01	1.35e-07
2	2	2	0	2	530	13.50	5.40e-02	7.42e-08	3.95e-01	2.09e-07
2	2	0	1	2	9	13.25	1.13e-03	1.99e-09	3.41e-01	2.11e-07
2	2	0	0	2	5	12.58	3.94e-04	1.35e-09	3.40e-01	2.12e-07
2	1	2	2	2	140	11.23	1.39e-02	1.84e-07	3.40e-01	3.97e-07
2	2	1	2	2	9	10.91	1.23e-03	2.24e-08	3.26e-01	4.19e-07
1	2	2	2	2	89	10.39	8.94e-03	2.75e-07	3.25e-01	6.94e-07
2	1	0	2	2	1	10.31	1.01e-04	3.36e-09	3.16e-01	6.98e-07
1	2	0	2	2	0	9.47	6.52e-05	5.02e-09	3.16e-01	7.03e-07
2	1	2	1	2	54	8.71	4.79e-03	7.93e-07	3.16e-01	1.50e-06
2	2	1	1	2	7	8.39	4.25e-04	9.64e-08	3.11e-01	1.59e-06
2	1	2	0	2	33	8.04	1.67e-03	5.40e-07	3.11e-01	2.13e-06
1	2	2	1	2	53	7.87	3.09e-03	1.18e-06	3.09e-01	3.32e-06
2	1	0	1	2	1	7.79	3.50e-05	1.45e-08	3.06e-01	3.33e-06
2	2	1	0	2	6	7.72	1.48e-04	6.57e-08	3.06e-01	3.40e-06
1	2	2	0	2	20	7.20	1.08e-03	8.07e-07	3.06e-01	4.20e-06
2	1	0	0	2	1	7.12	1.22e-05	9.85e-09	3.04e-01	4.21e-06
1	2	0	1	2	1	6.95	2.25e-05	2.16e-08	3.04e-01	4.24e-06
1	2	0	0	2	0	6.28	7.88e-06	1.47e-08	3.04e-01	4.25e-06
2	2	2	2	1	14356	5.71	1.95e-01	6.44e-04	3.04e-01	6.48e-04
2	1	1	2	2	5	5.45	3.81e-05	1.63e-07	1.09e-01	6.49e-04
1	1	2	2	2	89	4.93	2.77e-04	2.00e-06	1.09e-01	6.51e-04
2	2	0	2	1	245	4.80	1.42e-03	1.17e-05	1.09e-01	6.62e-04
1	2	1	2	2	5	4.61	2.46e-05	2.44e-07	1.08e-01	6.63e-04
1	1	0	2	2	0	4.01	2.02e-06	3.65e-08	1.08e-01	6.63e-04
2	2	2	1	1	55555	3.19	6.74e-02	2.77e-03	1.08e-01	3.43e-03
2	1	1	1	2	7	2.93	1.32e-05	7.02e-07	4.01e-02	3.44e-03
2	2	2	0	1	35204	2.52	2.36e-02	1.89e-03	4.01e-02	5.32e-03
1	1	2	1	2	150	2.41	9.59e-05	8.62e-06	1.66e-02	5.33e-03
2	2	0	1	1	1383	2.28	4.92e-04	5.05e-05	1.65e-02	5.38e-03
2	1	1	0	2	12	2.26	4.60e-06	4.78e-07	1.60e-02	5.38e-03
1	2	1	1	2	20	2.09	8.50e-06	1.05e-06	1.60e-02	5.39e-03
2	2	2	2	0	95	1.96	3.57e-05	5.01e-06	1.60e-02	5.39e-03
1	1	2	0	2	109	1.74	3.35e-05	5.88e-06	1.59e-02	5.40e-03
2	2	0	0	1	1281	1.61	1.72e-04	3.44e-05	1.59e-02	5.43e-03
1	1	0	1	2	1	1.49	6.99e-07	1.57e-07	1.57e-02	5.43e-03
1	2	1	0	2	14	1.42	2.97e-06	7.14e-07	1.57e-02	5.43e-03
2	2	0	2	0	0	1.05	2.60e-07	9.13e-08	1.57e-02	5.43e-03
1	2	1	1	0	4116	-12.63	6.79e-10	2.07e-04	1.59e-07	6.69e-01
1	1	2	0	0	22968	-12.98	2.68e-09	1.16e-03	1.58e-07	6.71e-01
1	1	0	1	0	0	-13.23	5.59e-11	3.11e-05	1.56e-07	6.71e-01
1	2	1	0	0	2353	-13.30	2.37e-10	1.41e-04	1.56e-07	6.71e-01
1	1	0	0	0	0	-13.90	1.95e-11	2.12e-05	1.55e-07	6.71e-01
1	1	1	1	1	3721782	-14.34	1.15e-07	1.94e-01	1.55e-07	8.65e-01
1	1	1	0	1	2600172	-15.01	4.02e-08	1.32e-01	4.03e-08	9.97e-01
1	1	1	2	0	6688	-15.57	6.09e-11	3.51e-04	8.93e-11	9.97e-01
1	1	1	1	0	29256	-18.09	2.11e-11	1.51e-03	2.84e-11	9.99e-01
1	1	1	0	0	20351	-18.76	7.36e-12	1.03e-03	7.36e-12	1.00e+00

Table C-15: The result of a classification with comparisons agreement/disagreement/either missing (2/1/0) for road accidents in 2007. The comparison vectors are ordered on the weight. The green comparison vectors are classified as positive links, the orange comparison vectors as possible links and the red comparison vectors as positive non-links. The estimated number of links N_M is 9708.

y^{hosp}	y^{epoch}	y^{gender}	y^{mot}	y^{dob}	$f(\boldsymbol{y})$	$w(\boldsymbol{y})$	$m(\boldsymbol{y})$	$u(\boldsymbol{y})$	λ	μ
2	2	2	2	2	4458	16.60	4.70e-01	2.90e-08	1.00e+00	2.90e-08
2	2	0	2	2	27	15.69	3.68e-03	5.63e-10	5.30e-01	2.96e-08
2	2	2	1	2	1339	13.98	1.43e-01	1.22e-07	5.26e-01	1.51e-07
2	2	2	0	2	489	13.46	5.18e-02	7.36e-08	3.83e-01	2.25e-07
2	2	0	1	2	13	13.07	1.12e-03	2.36e-09	3.31e-01	2.27e-07
2	2	0	0	2	5	12.56	4.06e-04	1.43e-09	3.30e-01	2.29e-07
2	2	1	2	2	15	11.22	1.92e-03	2.58e-08	3.30e-01	2.54e-07
2	1	2	2	2	143	11.19	1.52e-02	2.09e-07	3.28e-01	4.64e-07
1	2	2	2	2	82	10.34	9.14e-03	2.97e-07	3.13e-01	7.60e-07
2	1	0	2	2	0	10.28	1.19e-04	4.06e-09	3.04e-01	7.65e-07
1	2	0	2	2	2	9.43	7.16e-05	5.76e-09	3.03e-01	7.70e-07
2	2	1	1	2	12	8.60	5.84e-04	1.08e-07	3.03e-01	8.78e-07
2	1	2	1	2	60	8.57	4.62e-03	8.78e-07	3.03e-01	1.76e-06
2	2	1	0	2	1	8.08	2.11e-04	6.53e-08	2.98e-01	1.82e-06
2	1	2	0	2	34	8.05	1.67e-03	5.31e-07	2.98e-01	2.35e-06
1	2	2	1	2	49	7.71	2.78e-03	1.24e-06	2.96e-01	3.60e-06
2	1	0	1	2	1	7.66	3.62e-05	1.70e-08	2.93e-01	3.61e-06
1	2	2	0	2	30	7.20	1.01e-03	7.52e-07	2.93e-01	4.36e-06
2	1	0	0	2	0	7.15	1.31e-05	1.03e-08	2.92e-01	4.38e-06
1	2	0	1	2	3	6.81	2.18e-05	2.41e-08	2.92e-01	4.40e-06
1	2	0	0	2	0	6.29	7.89e-06	1.46e-08	2.92e-01	4.41e-06
2	1	1	2	2	6	5.81	6.19e-05	1.86e-07	2.92e-01	4.60e-06
2	2	2	2	1	15022	5.58	1.94e-01	7.29e-04	2.92e-01	7.34e-04
1	2	1	2	2	10	4.95	3.73e-05	2.63e-07	9.87e-02	7.34e-04
1	1	2	2	2	65	4.93	2.95e-04	2.14e-06	9.87e-02	7.36e-04
2	2	0	2	1	299	4.67	1.52e-03	1.42e-05	9.84e-02	7.51e-04
1	1	0	2	2	0	4.02	2.31e-06	4.15e-08	9.69e-02	7.51e-04
2	1	1	1	2	17	3.19	1.88e-05	7.79e-07	9.69e-02	7.51e-04
2	2	2	1	1	56582	2.96	5.90e-02	3.06e-03	9.68e-02	3.81e-03
2	1	1	0	2	9	2.67	6.82e-06	4.71e-07	3.79e-02	3.81e-03
2	2	2	2	0	131	2.53	8.90e-05	7.09e-06	3.79e-02	3.82e-03
2	2	2	0	1	32479	2.45	2.14e-02	1.85e-03	3.78e-02	5.67e-03
1	2	1	1	2	25	2.33	1.14e-05	1.10e-06	1.64e-02	5.67e-03
1	1	2	1	2	145	2.30	8.98e-05	8.97e-06	1.64e-02	5.68e-03
2	2	0	1	1	1306	2.05	4.62e-04	5.93e-05	1.63e-02	5.74e-03
1	2	1	0	2	11	1.82	4.11e-06	6.68e-07	1.59e-02	5.74e-03
1	1	2	0	2	90	1.79	3.25e-05	5.43e-06	1.59e-02	5.74e-03
2	2	0	2	0	0	1.62	6.97e-07	1.38e-07	1.58e-02	5.74e-03
2	2	0	0	1	1127	1.54	1.67e-04	3.59e-05	1.58e-02	5.78e-03
1	1	0	1	2	2	1.40	7.03e-07	1.74e-07	1.57e-02	5.78e-03
1	1	2	1	0	43046	-11.77	1.70e-08	2.19e-03	2.30e-07	6.74e-01
1	2	1	0	0	2664	-12.25	7.78e-10	1.63e-04	2.13e-07	6.74e-01
1	1	2	0	0	25698	-12.28	6.16e-09	1.33e-03	2.12e-07	6.75e-01
1	1	0	1	0	0	-12.67	1.33e-10	4.25e-05	2.06e-07	6.75e-01
1	1	0	0	0	0	-13.19	4.82e-11	2.57e-05	2.06e-07	6.75e-01
1	1	1	1	1	3597307	-14.10	1.51e-07	2.00e-01	2.06e-07	8.75e-01
1	1	1	2	0	7447	-14.53	2.28e-10	4.64e-04	5.50e-08	8.76e-01
1	1	1	0	1	2205761	-14.61	5.47e-08	1.21e-01	5.48e-08	9.97e-01
1	1	1	1	0	34562	-17.15	6.94e-11	1.94e-03	9.45e-11	9.99e-01
1	1	1	0	0	21281	-17.66	2.51e-11	1.18e-03	2.51e-11	1.00e+00

Table C-16: The result of a classification with comparisons agreement/disagreement/either missing (2/1/0) for road accidents in 2008. The comparison vectors are ordered on the weight. The green comparison vectors are classified as positive links, the orange comparison vectors as possible links and the red comparison vectors as positive non-links. The estimated number of links N_M is 9435.

y^{hosp}	y^{epoch}	y^{gender}	y^{mot}	y^{dob}	$f(\boldsymbol{y})$	$w(\boldsymbol{y})$	$m(\boldsymbol{y})$	$u(\boldsymbol{y})$	λ	μ
2	2	2	2	2	4091	16.61	5.14e-01	3.15e-08	1.00e+00	3.15e-08
2	2	0	2	2	151	15.75	1.86e-02	2.68e-09	4.86e-01	3.41e-08
2	2	2	1	2	1241	13.97	1.54e-01	1.32e-07	4.67e-01	1.66e-07
2	2	2	0	2	386	13.63	5.02e-02	6.06e-08	3.14e-01	2.27e-07
2	2	0	1	2	34	13.11	5.56e-03	1.13e-08	2.63e-01	2.38e-07
2	2	0	0	2	19	12.77	1.82e-03	5.16e-09	2.58e-01	2.43e-07
2	2	1	2	2	15	11.09	1.84e-03	2.80e-08	2.56e-01	2.71e-07
2	1	2	2	2	105	11.04	1.42e-02	2.27e-07	2.54e-01	4.98e-07
1	2	2	2	2	112	10.75	1.39e-02	2.98e-07	2.40e-01	7.96e-07
2	1	0	2	2	6	10.19	5.13e-04	1.93e-08	2.26e-01	8.15e-07
1	2	0	2	2	3	9.89	5.02e-04	2.53e-08	2.26e-01	8.40e-07
2	2	1	1	2	5	8.45	5.50e-04	1.18e-07	2.25e-01	9.58e-07
2	1	2	1	2	55	8.40	4.24e-03	9.53e-07	2.25e-01	1.91e-06
2	2	1	0	2	3	8.11	1.80e-04	5.39e-08	2.20e-01	1.96e-06
1	2	2	1	2	53	8.11	4.15e-03	1.25e-06	2.20e-01	3.22e-06
2	1	2	0	2	25	8.06	1.39e-03	4.37e-07	2.16e-01	3.65e-06
1	2	2	0	2	17	7.77	1.36e-03	5.73e-07	2.15e-01	4.23e-06
2	1	0	1	2	1	7.54	1.53e-04	8.12e-08	2.13e-01	4.31e-06
1	2	0	1	2	4	7.25	1.50e-04	1.07e-07	2.13e-01	4.41e-06
2	1	0	0	2	1	7.21	5.01e-05	3.72e-08	2.13e-01	4.45e-06
1	2	0	0	2	2	6.91	4.91e-05	4.88e-08	2.13e-01	4.50e-06
2	1	1	2	2	5	5.53	5.08e-05	2.02e-07	2.13e-01	4.70e-06
1	2	1	2	2	5	5.24	4.97e-05	2.65e-07	2.13e-01	4.97e-06
1	1	2	2	2	69	5.18	3.83e-04	2.14e-06	2.13e-01	7.11e-06
2	2	2	2	1	13831	5.13	1.38e-01	8.19e-04	2.12e-01	8.26e-04
1	1	0	2	2	3	4.33	1.39e-05	1.83e-07	7.39e-02	8.26e-04
2	2	0	2	1	1189	4.27	5.01e-03	6.97e-05	7.39e-02	8.96e-04
2	1	1	1	2	11	2.88	1.52e-05	8.48e-07	6.89e-02	8.96e-04
1	2	1	1	2	8	2.59	1.48e-05	1.11e-06	6.89e-02	8.98e-04
2	1	1	0	2	3	2.55	4.96e-06	3.89e-07	6.89e-02	8.98e-04
1	1	2	1	2	131	2.54	1.14e-04	9.01e-06	6.89e-02	9.07e-04
2	2	2	1	1	52071	2.49	4.14e-02	3.44e-03	6.87e-02	4.35e-03
1	2	1	0	2	3	2.25	4.86e-06	5.10e-07	2.74e-02	4.35e-03
1	1	2	0	2	65	2.20	3.74e-05	4.13e-06	2.74e-02	4.35e-03
2	2	2	0	1	25742	2.15	1.35e-02	1.58e-03	2.73e-02	5.93e-03
1	1	0	1	2	10	1.69	4.14e-06	7.68e-07	1.38e-02	5.93e-03
2	2	0	1	1	4298	1.63	1.50e-03	2.93e-04	1.38e-02	6.22e-03
1	1	0	0	2	4	1.35	1.35e-06	3.52e-07	1.23e-02	6.22e-03
2	2	0	0	1	2512	1.29	4.90e-04	1.34e-04	1.23e-02	6.36e-03
1	1	1	2	2	48	-0.33	1.37e-06	1.91e-06	1.18e-02	6.36e-03
1	2	1	1	0	4999	-15.62	5.31e-11	3.23e-04	6.58e-10	9.92e-01
2	1	1	0	0	1659	-15.67	1.78e-11	1.13e-04	6.05e-10	9.92e-01
1	1	2	1	0	45455	-15.67	4.09e-10	2.62e-03	5.87e-10	9.94e-01
1	2	1	0	0	1945	-15.96	1.74e-11	1.48e-04	1.78e-10	9.95e-01
1	1	2	0	0	20447	-16.01	1.34e-10	1.20e-03	1.60e-10	9.96e-01
1	1	0	1	0	0	-16.53	1.48e-11	2.23e-04	2.65e-11	9.96e-01
1	1	0	0	0	0	-16.87	4.85e-12	1.02e-04	1.17e-11	9.96e-01
1	1	1	2	0	8650	-18.54	4.91e-12	5.55e-04	6.85e-12	9.97e-01
1	1	1	1	0	36573	-21.19	1.47e-12	2.33e-03	1.95e-12	9.99e-01
1	1	1	0	0	16249	-21.53	4.79e-13	1.07e-03	4.79e-13	1.00e+00

Table C-17: The result of a classification with comparisons agreement/disagreement/either missing (2/1/0) for road accidents in 2009. The comparison vectors are ordered on the weight. The green comparison vectors are classified as positive links, the orange comparison vectors as possible links and the red comparison vectors as positive non-links. The estimated number of links N_M is 7951.

y^{hosp}	y^{epoch}	y^{gender}	y^{mot}	y^{dob}	$f(\boldsymbol{y})$	$w(\boldsymbol{y})$	$m(\boldsymbol{y})$	$u(\boldsymbol{y})$	λ	μ
2	2	2	2	2	2872	16.38	4.10e-01	3.14e-08	1.00e+00	3.14e-08
2	2	0	2	2	8	15.99	1.38e-03	1.56e-10	5.90e-01	3.16e-08
2	2	2	1	2	1125	13.96	1.66e-01	1.44e-07	5.88e-01	1.75e-07
2	2	0	1	2	6	13.57	5.58e-04	7.11e-10	4.22e-01	1.76e-07
2	2	2	0	2	168	13.00	2.24e-02	5.04e-08	4.21e-01	2.26e-07
2	2	0	0	2	0	12.61	7.51e-05	2.50e-10	3.99e-01	2.27e-07
2	1	2	2	2	104	11.03	1.40e-02	2.28e-07	3.99e-01	4.54e-07
2	2	1	2	2	8	10.97	1.64e-03	2.83e-08	3.85e-01	4.83e-07
2	1	0	2	2	0	10.64	4.71e-05	1.13e-09	3.83e-01	4.84e-07
1	2	2	2	2	61	10.26	8.54e-03	2.98e-07	3.83e-01	7.82e-07
1	2	0	2	2	0	9.87	2.87e-05	1.48e-09	3.75e-01	7.84e-07
2	1	2	1	2	40	8.61	5.69e-03	1.04e-06	3.75e-01	1.82e-06
2	2	1	1	2	8	8.55	6.66e-04	1.29e-07	3.69e-01	1.95e-06
2	1	0	1	2	0	8.22	1.91e-05	5.16e-09	3.68e-01	1.96e-06
1	2	2	1	2	33	7.84	3.46e-03	1.36e-06	3.68e-01	3.32e-06
2	1	2	0	2	8	7.65	7.65e-04	3.66e-07	3.65e-01	3.68e-06
2	2	1	0	2	2	7.59	8.96e-05	4.54e-08	3.64e-01	3.73e-06
1	2	0	1	2	0	7.45	1.16e-05	6.74e-09	3.64e-01	3.74e-06
2	1	0	0	2	0	7.26	2.57e-06	1.81e-09	3.64e-01	3.74e-06
1	2	2	0	2	6	6.88	4.66e-04	4.78e-07	3.64e-01	4.22e-06
1	2	0	0	2	0	6.49	1.56e-06	2.37e-09	3.63e-01	4.22e-06
2	2	2	2	1	8951	5.62	2.33e-01	8.40e-04	3.63e-01	8.44e-04
2	1	1	2	2	2	5.61	5.62e-05	2.05e-07	1.31e-01	8.44e-04
2	2	0	2	1	43	5.23	7.81e-04	4.16e-06	1.31e-01	8.48e-04
2	2	2	2	0	68	4.92	9.73e-04	7.08e-06	1.30e-01	8.55e-04
1	1	2	2	2	47	4.91	2.92e-04	2.16e-06	1.29e-01	8.57e-04
1	2	1	2	2	3	4.85	3.42e-05	2.69e-07	1.29e-01	8.58e-04
2	2	0	2	0	0	4.53	3.26e-06	3.51e-08	1.29e-01	8.58e-04
1	1	0	2	2	0	4.52	9.81e-07	1.07e-08	1.29e-01	8.58e-04
2	2	2	1	1	36730	3.20	9.43e-02	3.83e-03	1.29e-01	4.69e-03
2	1	1	1	2	9	3.19	2.28e-05	9.38e-07	3.42e-02	4.69e-03
2	2	0	1	1	232	2.81	3.17e-04	1.90e-05	3.42e-02	4.71e-03
2	2	2	1	0	330	2.50	3.94e-04	3.23e-05	3.39e-02	4.74e-03
1	1	2	1	2	83	2.49	1.18e-04	9.86e-06	3.35e-02	4.75e-03
1	2	1	1	2	11	2.43	1.39e-05	1.23e-06	3.33e-02	4.75e-03
2	2	2	0	1	10606	2.24	1.27e-02	1.35e-03	3.33e-02	6.10e-03
2	1	1	0	2	4	2.23	3.07e-06	3.29e-07	2.06e-02	6.10e-03
2	2	0	1	0	0	2.11	1.32e-06	1.60e-07	2.06e-02	6.10e-03
1	1	0	1	2	1	2.10	3.97e-07	4.89e-08	2.06e-02	6.10e-03
2	2	0	0	1	48	1.85	4.26e-05	6.67e-06	2.06e-02	6.11e-03
1	1	0	0	1	4110	-9.62	3.03e-08	4.58e-04	1.05e-06	6.23e-01
1	1	2	0	0	7156	-9.94	3.78e-08	7.80e-04	1.02e-06	6.24e-01
1	2	1	0	0	735	-10.00	4.42e-09	9.70e-05	9.77e-07	6.24e-01
1	1	0	0	0	0	-10.33	1.27e-10	3.87e-06	9.73e-07	6.24e-01
1	1	1	2	1	455929	-11.27	6.64e-07	5.20e-02	9.73e-07	6.76e-01
1	1	1	2	0	3601	-11.97	2.77e-09	4.38e-04	3.09e-07	6.77e-01
1	1	1	1	1	2112101	-13.69	2.69e-07	2.37e-01	3.06e-07	9.14e-01
1	1	1	1	0	18357	-14.39	1.12e-09	2.00e-03	3.75e-08	9.16e-01
1	1	1	0	1	755088	-14.65	3.62e-08	8.34e-02	3.63e-08	9.99e-01
1	1	1	0	0	6134	-15.35	1.51e-10	7.03e-04	1.51e-10	1.00e+00

Table C-18: The result of a classification with comparisons agreement/disagreement/either missing (2/1/0) for road accidents in 2010. The comparison vectors are ordered on the weight. The green comparison vectors are classified as positive links, the orange comparison vectors as possible links and the red comparison vectors as positive non-links. The estimated number of links N_M is 6938.

y^{hosp}	y^{epoch}	y^{gender}	y^{mot}	y^{dob}	$f(\boldsymbol{y})$	$w(\boldsymbol{y})$	$m(\boldsymbol{y})$	$u(\boldsymbol{y})$	λ	μ
2	2	2	2	2	1555	16.50	4.57e-01	3.13e-08	1.00e+00	3.13e-08
2	2	0	2	2	2	15.11	5.41e-04	1.48e-10	5.43e-01	3.15e-08
2	2	2	1	2	547	13.93	1.64e-01	1.46e-07	5.43e-01	1.77e-07
2	2	2	0	2	55	13.01	1.61e-02	3.60e-08	3.79e-01	2.13e-07
2	2	0	1	2	1	12.55	1.94e-04	6.89e-10	3.63e-01	2.14e-07
2	2	0	0	2	0	11.63	1.90e-05	1.70e-10	3.62e-01	2.14e-07
2	2	1	2	2	4	10.70	1.25e-03	2.80e-08	3.62e-01	2.42e-07
2	1	2	2	2	28	10.69	9.89e-03	2.24e-07	3.61e-01	4.67e-07
1	2	2	2	2	35	10.58	1.14e-02	2.91e-07	3.51e-01	7.57e-07
2	1	0	2	2	0	9.31	1.17e-05	1.06e-09	3.40e-01	7.58e-07
1	2	0	2	2	0	9.20	1.35e-05	1.37e-09	3.40e-01	7.60e-07
2	2	1	1	2	2	8.14	4.48e-04	1.31e-07	3.40e-01	8.90e-07
2	1	2	1	2	20	8.13	3.55e-03	1.05e-06	3.39e-01	1.94e-06
1	2	2	1	2	24	8.02	4.11e-03	1.35e-06	3.36e-01	3.29e-06
2	2	1	0	2	0	7.22	4.39e-05	3.22e-08	3.32e-01	3.32e-06
2	1	2	0	2	7	7.21	3.48e-04	2.58e-07	3.32e-01	3.58e-06
1	2	2	0	2	3	7.10	4.03e-04	3.34e-07	3.31e-01	3.91e-06
2	1	0	1	2	0	6.75	4.20e-06	4.93e-09	3.31e-01	3.92e-06
1	2	0	1	2	0	6.63	4.86e-06	6.39e-09	3.31e-01	3.92e-06
2	1	0	0	2	0	5.83	4.12e-07	1.22e-09	3.31e-01	3.93e-06
1	2	0	0	2	0	5.71	4.76e-07	1.58e-09	3.31e-01	3.93e-06
2	2	2	2	1	4592	5.50	2.25e-01	9.16e-04	3.31e-01	9.20e-04
2	2	2	2	0	25	5.08	8.13e-04	5.06e-06	1.06e-01	9.25e-04
2	1	1	2	2	1	4.90	2.70e-05	2.01e-07	1.05e-01	9.25e-04
1	2	1	2	2	2	4.79	3.12e-05	2.60e-07	1.05e-01	9.26e-04
1	1	2	2	2	17	4.78	2.48e-04	2.08e-06	1.05e-01	9.28e-04
2	2	0	2	1	12	4.12	2.66e-04	4.32e-06	1.05e-01	9.32e-04
2	2	0	2	0	0	3.70	9.62e-07	2.39e-08	1.05e-01	9.32e-04
1	1	0	2	2	0	3.40	2.93e-07	9.82e-09	1.05e-01	9.32e-04
2	2	2	1	1	18318	2.94	8.07e-02	4.27e-03	1.05e-01	5.20e-03
2	2	2	1	0	90	2.52	2.92e-04	2.36e-05	2.42e-02	5.23e-03
2	1	1	1	2	6	2.34	9.70e-06	9.36e-07	2.39e-02	5.23e-03
1	2	1	1	2	6	2.23	1.12e-05	1.21e-06	2.39e-02	5.23e-03
1	1	2	1	2	44	2.22	8.90e-05	9.70e-06	2.39e-02	5.24e-03
2	2	2	0	1	3970	2.02	7.90e-03	1.05e-03	2.38e-02	6.29e-03
2	2	2	0	0	23	1.59	2.86e-05	5.81e-06	1.59e-02	6.30e-03
2	2	0	1	1	85	1.56	9.55e-05	2.02e-05	1.59e-02	6.32e-03
2	1	1	0	2	1	1.42	9.50e-07	2.31e-07	1.58e-02	6.32e-03
1	2	1	0	2	0	1.30	1.10e-06	2.99e-07	1.58e-02	6.32e-03
1	1	2	0	2	6	1.29	8.72e-06	2.39e-06	1.58e-02	6.32e-03
1	1	0	1	1	5638	-10.16	5.18e-08	1.34e-03	5.23e-07	6.26e-01
1	1	0	1	0	0	-10.58	1.87e-10	7.40e-06	4.71e-07	6.26e-01
1	1	0	0	1	1509	-11.08	5.08e-09	3.30e-04	4.71e-07	6.27e-01
1	1	0	0	0	0	-11.51	1.84e-11	1.82e-06	4.66e-07	6.27e-01
1	1	1	2	1	224976	-12.01	3.33e-07	5.45e-02	4.66e-07	6.81e-01
1	1	1	2	0	1162	-12.43	1.20e-09	3.01e-04	1.33e-07	6.82e-01
1	1	1	1	1	1060030	-14.57	1.20e-07	2.54e-01	1.32e-07	9.36e-01
1	1	1	1	0	6338	-14.99	4.32e-10	1.40e-03	1.22e-08	9.37e-01
1	1	1	0	1	267892	-15.49	1.17e-08	6.26e-02	1.17e-08	1.00e+00
1	1	1	0	0	1453	-15.91	4.24e-11	3.46e-04	4.24e-11	1.00e+00

Table C-19: The result of a classification with comparisons agreement/disagreement/either missing (2/1/0) for road accidents in 2011. The comparison vectors are ordered on the weight. The green comparison vectors are classified as positive links, the orange comparison vectors as possible links and the red comparison vectors as positive non-links. The estimated number of links N_M is 3382.

y^{hosp}	y^{epoch}	y^{gender}	y^{mot}	y^{dob}	$f(\boldsymbol{y})$	$w(\boldsymbol{y})$	$m(\boldsymbol{y})$	$u(\boldsymbol{y})$	λ	μ
2	2	2	2	2	1422	16.25	4.47e-01	3.93e-08	1.00e+00	3.93e-08
2	2	2	1	2	568	13.85	1.80e-01	1.74e-07	5.53e-01	2.13e-07
2	2	0	2	2	0	13.59	1.89e-04	2.37e-10	3.73e-01	2.14e-07
2	2	2	0	2	37	12.76	1.21e-02	3.47e-08	3.73e-01	2.48e-07
2	2	0	1	2	1	11.20	7.64e-05	1.05e-09	3.61e-01	2.49e-07
2	1	2	2	2	40	10.70	1.28e-02	2.90e-07	3.61e-01	5.40e-07
2	2	1	2	2	5	10.54	1.32e-03	3.47e-08	3.48e-01	5.74e-07
2	2	0	0	2	0	10.10	5.11e-06	2.09e-10	3.47e-01	5.75e-07
1	2	2	2	2	23	10.07	8.47e-03	3.57e-07	3.47e-01	9.31e-07
2	1	2	1	2	22	8.30	5.18e-03	1.29e-06	3.38e-01	2.22e-06
2	2	1	1	2	0	8.15	5.32e-04	1.54e-07	3.33e-01	2.37e-06
2	1	0	2	2	0	8.04	5.44e-06	1.75e-09	3.32e-01	2.37e-06
1	2	2	1	2	21	7.68	3.42e-03	1.58e-06	3.32e-01	3.95e-06
1	2	0	2	2	0	7.42	3.59e-06	2.15e-09	3.29e-01	3.95e-06
2	1	2	0	2	2	7.21	3.47e-04	2.56e-07	3.29e-01	4.21e-06
2	2	1	0	2	1	7.06	3.56e-05	3.06e-08	3.29e-01	4.24e-06
1	2	2	0	2	3	6.59	2.29e-04	3.15e-07	3.29e-01	4.56e-06
2	1	0	1	2	0	5.65	2.20e-06	7.75e-09	3.28e-01	4.56e-06
2	2	2	2	1	3924	5.42	2.18e-01	9.62e-04	3.28e-01	9.67e-04
1	2	0	1	2	0	5.02	1.45e-06	9.52e-09	1.11e-01	9.67e-04
2	1	1	2	2	1	5.00	3.79e-05	2.57e-07	1.11e-01	9.67e-04
2	1	0	0	2	0	4.56	1.47e-07	1.54e-09	1.11e-01	9.67e-04
1	1	2	2	2	16	4.53	2.44e-04	2.63e-06	1.10e-01	9.70e-04
1	2	1	2	2	4	4.37	2.50e-05	3.15e-07	1.10e-01	9.70e-04
1	2	0	0	2	0	3.93	9.68e-08	1.90e-09	1.10e-01	9.70e-04
2	2	2	1	1	14701	3.03	8.79e-02	4.26e-03	1.10e-01	5.23e-03
2	2	0	2	1	15	2.77	9.23e-05	5.80e-06	2.23e-02	5.24e-03
2	1	1	1	2	5	2.60	1.53e-05	1.14e-06	2.22e-02	5.24e-03
1	1	2	1	2	28	2.13	9.83e-05	1.17e-05	2.22e-02	5.25e-03
1	2	1	1	2	2	1.98	1.01e-05	1.40e-06	2.21e-02	5.25e-03
2	2	2	0	1	2936	1.94	5.88e-03	8.48e-04	2.21e-02	6.10e-03
1	1	0	2	2	0	1.87	1.03e-07	1.59e-08	1.62e-02	6.10e-03
2	1	1	0	2	0	1.51	1.02e-06	2.26e-07	1.62e-02	6.10e-03
2	2	2	2	0	95	1.06	8.97e-05	3.10e-05	1.62e-02	6.13e-03
1	1	2	0	2	6	1.04	6.57e-06	2.32e-06	1.61e-02	6.13e-03
1	2	1	0	2	3	0.89	6.75e-07	2.78e-07	1.61e-02	6.13e-03
2	2	0	1	1	82	0.37	3.73e-05	2.57e-05	1.61e-02	6.16e-03
2	1	2	2	1	24671	-0.13	6.27e-03	7.10e-03	1.61e-02	1.33e-02
2	2	1	2	1	3390	-0.28	6.43e-04	8.50e-04	9.80e-03	1.41e-02
1	1	0	1	2	0	-0.52	4.16e-08	7.03e-08	9.16e-03	1.41e-02
2	1	1	0	0	639	-13.67	2.06e-10	1.79e-04	1.53e-07	6.84e-01
1	1	2	0	0	5803	-14.14	1.32e-09	1.83e-03	1.53e-07	6.86e-01
1	2	1	0	0	581	-14.30	1.36e-10	2.19e-04	1.51e-07	6.86e-01
1	1	1	1	1	847411	-14.39	1.42e-07	2.52e-01	1.51e-07	9.38e-01
1	1	1	0	1	171141	-15.48	9.47e-09	5.02e-02	9.68e-09	9.88e-01
1	1	0	1	0	0	-15.71	8.37e-12	5.55e-05	2.16e-10	9.88e-01
1	1	1	2	0	5151	-16.36	1.44e-10	1.84e-03	2.07e-10	9.90e-01
1	1	0	0	0	0	-16.80	5.60e-13	1.10e-05	6.28e-11	9.90e-01
1	1	1	1	0	29136	-18.75	5.83e-11	8.14e-03	6.22e-11	9.98e-01
1	1	1	0	0	5247	-19.85	3.90e-12	1.62e-03	3.90e-12	1.00e+00

Table C-20: The result of a classification with comparisons agreement/disagreement/either missing (2/1/0) for road accidents in 2012. The comparison vectors are ordered on the weight. The green comparison vectors are classified as positive links, the orange comparison vectors as possible links and the red comparison vectors as positive non-links. The estimated number of links N_M is 3168.

C-4 Estimation with the EM-algorithm with missing values and data blocked on the year of birth

y^{hosp}	y^{epoch}	y^{gender}	y^{mot}	y^{dob}	$f(\boldsymbol{y})$	$w(\boldsymbol{y})$	$m(\boldsymbol{y})$	$u(\boldsymbol{y})$	λ	μ
2	2	2	2	2	1430	16.64	3.50e-01	2.07e-08	1.00e+00	2.07e-08
2	2	0	2	2	5	15.58	1.01e-03	1.72e-10	6.50e-01	2.09e-08
2	2	2	0	2	382	14.53	9.10e-02	4.47e-08	6.49e-01	6.56e-08
2	2	2	1	2	453	13.88	1.11e-01	1.04e-07	5.58e-01	1.70e-07
2	2	0	0	2	2	13.47	2.62e-04	3.72e-10	4.46e-01	1.70e-07
2	2	0	1	2	1	12.82	3.21e-04	8.67e-10	4.46e-01	1.71e-07
1	2	2	2	2	103	11.86	3.23e-02	2.29e-07	4.46e-01	4.00e-07
2	1	2	2	2	43	11.22	1.15e-02	1.53e-07	4.13e-01	5.54e-07
1	2	0	2	2	0	10.80	9.32e-05	1.91e-09	4.02e-01	5.56e-07
2	1	0	2	2	0	10.17	3.31e-05	1.27e-09	4.02e-01	5.57e-07
1	2	2	0	2	78	9.74	8.39e-03	4.94e-07	4.02e-01	1.05e-06
2	1	2	0	2	11	9.11	2.98e-03	3.30e-07	3.93e-01	1.38e-06
1	2	2	1	2	30	9.09	1.03e-02	1.15e-06	3.90e-01	2.54e-06
1	2	0	0	2	0	8.68	2.42e-05	4.11e-09	3.80e-01	2.54e-06
2	1	2	1	2	19	8.46	3.65e-03	7.71e-07	3.80e-01	3.31e-06
2	1	0	0	2	0	8.05	8.59e-06	2.75e-09	3.76e-01	3.31e-06
1	2	0	1	2	0	8.04	2.96e-05	9.59e-09	3.76e-01	3.32e-06
2	1	0	1	2	0	7.40	1.05e-05	6.41e-09	3.76e-01	3.33e-06
2	2	2	2	0	138	7.24	1.86e-02	1.34e-05	3.76e-01	1.68e-05
1	1	2	2	2	25	6.44	1.06e-03	1.69e-06	3.58e-01	1.84e-05
2	2	0	2	0	0	6.18	5.38e-05	1.12e-07	3.57e-01	1.86e-05
2	2	2	2	1	4049	5.74	1.91e-01	6.16e-04	3.57e-01	6.35e-04
1	1	0	2	2	1	5.38	3.05e-06	1.41e-08	1.65e-01	6.35e-04
2	2	2	0	0	119	5.12	4.84e-03	2.90e-05	1.65e-01	6.64e-04
2	2	0	2	1	24	4.68	5.52e-04	5.12e-06	1.60e-01	6.69e-04
2	2	2	1	0	486	4.47	5.93e-03	6.76e-05	1.60e-01	7.36e-04
1	1	2	0	2	15	4.32	2.75e-04	3.65e-06	1.54e-01	7.40e-04
2	2	0	0	0	0	4.06	1.40e-05	2.41e-07	1.54e-01	7.40e-04
1	1	2	1	2	46	3.68	3.37e-04	8.52e-06	1.54e-01	7.49e-04
2	2	2	0	1	5669	3.62	4.97e-02	1.33e-03	1.53e-01	2.08e-03
2	2	0	1	0	0	3.42	1.71e-05	5.62e-07	1.04e-01	2.08e-03
1	1	0	0	2	0	3.26	7.93e-07	3.04e-08	1.04e-01	2.08e-03
2	2	2	1	1	16688	2.98	6.08e-02	3.10e-03	1.04e-01	5.18e-03
1	1	0	1	2	0	2.62	9.71e-07	7.09e-08	4.28e-02	5.18e-03
2	2	0	0	1	45	2.56	1.43e-04	1.10e-05	4.28e-02	5.19e-03
1	2	2	2	0	907	2.45	1.72e-03	1.48e-04	4.27e-02	5.34e-03
2	2	0	1	1	143	1.92	1.75e-04	2.58e-05	4.10e-02	5.36e-03
2	1	2	2	0	666	1.82	6.11e-04	9.92e-05	4.08e-02	5.46e-03
1	2	0	2	0	0	1.39	4.96e-06	1.23e-06	4.02e-02	5.46e-03
1	2	2	2	1	36787	0.95	1.76e-02	6.81e-03	4.02e-02	1.23e-02
1	2	1	0	1	64814	-23.40	9.23e-13	1.34e-02	2.99e-12	5.56e-01
2	1	1	0	1	44752	-24.03	3.28e-13	8.93e-03	2.06e-12	5.65e-01
1	2	1	1	1	161174	-24.04	1.13e-12	3.12e-02	1.73e-12	5.96e-01
2	1	1	1	1	105304	-24.67	4.02e-13	2.08e-02	6.03e-13	6.17e-01
1	1	1	2	0	5118	-25.20	1.14e-14	9.98e-04	2.02e-13	6.18e-01
1	1	1	2	1	227172	-26.70	1.16e-13	4.58e-02	1.90e-13	6.64e-01
1	1	1	0	0	7479	-27.32	2.95e-15	2.15e-03	7.39e-14	6.66e-01
1	1	1	1	0	27869	-27.96	3.61e-15	5.02e-03	7.09e-14	6.71e-01
1	1	1	0	1	523514	-28.81	3.02e-14	9.87e-02	6.73e-14	7.70e-01
1	1	1	1	1	1162216	-29.46	3.71e-14	2.30e-01	3.71e-14	1.00e+00

Table C-21: The result of a classification with comparisons agreement/disagreement/either missing (2/1/0) for road accidents in 2013. The comparison vectors are ordered on the weight. The green comparison vectors are classified as positive links, the orange comparison vectors as possible links and the red comparison vectors as positive non-links. The estimated number of links N_M is 4079.

y^{hosp}	y^{epoch}	y^{gender}	y^{mot}	y^{dob}	$f(\boldsymbol{y})$	$w(\boldsymbol{y})$	$m(\boldsymbol{y})$	$u(\boldsymbol{y})$	λ	μ
2	2	2	2	2	4348	12.29	6.22e-01	2.87e-06	1.00e+00	2.87e-06
2	2	0	2	2	29	11.32	4.17e-03	5.05e-08	3.78e-01	2.92e-06
2	2	2	1	2	1495	10.20	2.14e-01	7.95e-06	3.73e-01	1.09e-05
2	2	2	0	2	530	9.55	7.65e-02	5.46e-06	1.60e-01	1.63e-05
2	2	0	1	2	9	9.24	1.43e-03	1.40e-07	8.32e-02	1.65e-05
2	2	0	0	2	5	8.58	5.13e-04	9.59e-08	8.18e-02	1.66e-05
2	1	2	2	2	140	6.82	1.92e-02	2.09e-05	8.13e-02	3.75e-05
2	2	1	2	2	9	6.52	1.62e-03	2.39e-06	6.20e-02	3.99e-05
1	2	2	2	2	89	5.93	1.17e-02	3.09e-05	6.04e-02	7.08e-05
2	1	0	2	2	1	5.86	1.29e-04	3.68e-07	4.88e-02	7.12e-05
1	2	0	2	2	0	4.97	7.82e-05	5.43e-07	4.86e-02	7.17e-05
2	1	2	1	2	54	4.74	6.61e-03	5.80e-05	4.86e-02	1.30e-04
2	2	1	1	2	7	4.44	5.58e-04	6.61e-06	4.20e-02	1.36e-04
2	1	2	0	2	33	4.08	2.36e-03	3.98e-05	4.14e-02	1.76e-04
1	2	2	1	2	53	3.85	4.00e-03	8.55e-05	3.90e-02	2.62e-04
2	2	1	0	2	6	3.78	2.00e-04	4.54e-06	3.50e-02	2.66e-04
2	1	0	1	2	1	3.77	4.43e-05	1.02e-06	3.48e-02	2.67e-04
1	2	2	0	2	20	3.19	1.43e-03	5.87e-05	3.48e-02	3.26e-04
2	1	0	0	2	1	3.12	1.58e-05	6.99e-07	3.34e-02	3.27e-04
2	2	2	2	1	395	3.07	2.10e-02	9.76e-04	3.33e-02	1.30e-03
1	2	0	1	2	1	2.88	2.68e-05	1.50e-06	1.23e-02	1.30e-03
1	2	0	0	2	0	2.23	9.60e-06	1.03e-06	1.23e-02	1.31e-03
2	2	0	2	1	3	2.11	1.41e-04	1.71e-05	1.23e-02	1.32e-03
2	1	1	2	2	5	1.06	5.02e-05	1.74e-05	1.22e-02	1.34e-03
2	2	2	1	1	786	0.98	7.21e-03	2.70e-03	1.21e-02	4.04e-03
1	1	2	2	2	89	0.47	3.60e-04	2.25e-04	4.89e-03	4.27e-03
2	2	2	0	1	493	0.33	2.58e-03	1.86e-03	4.53e-03	6.12e-03
1	2	1	2	2	5	0.17	3.04e-05	2.57e-05	1.95e-03	6.15e-03
2	2	0	1	1	16	0.02	4.84e-05	4.74e-05	1.92e-03	6.20e-03
1	1	0	2	2	0	-0.49	2.42e-06	3.96e-06	1.87e-03	6.20e-03
2	2	0	0	1	17	-0.63	1.73e-05	3.26e-05	1.87e-03	6.23e-03
2	1	1	1	2	7	-1.03	1.72e-05	4.82e-05	1.85e-03	6.28e-03
1	1	2	1	2	150	-1.62	1.24e-04	6.23e-04	1.83e-03	6.90e-03
2	1	1	0	2	12	-1.68	6.17e-06	3.31e-05	1.71e-03	6.94e-03
1	2	1	1	2	20	-1.92	1.04e-05	7.11e-05	1.70e-03	7.01e-03
1	1	2	0	2	109	-2.27	4.43e-05	4.28e-04	1.69e-03	7.44e-03
2	1	2	2	1	1753	-2.39	6.49e-04	7.12e-03	1.65e-03	1.46e-02
1	2	1	0	2	14	-2.57	3.74e-06	4.88e-05	9.98e-04	1.46e-02
1	1	0	1	2	1	-2.58	8.30e-07	1.09e-05	9.95e-04	1.46e-02
2	2	1	2	1	236	-2.70	5.48e-05	8.12e-04	9.94e-04	1.54e-02
1	1	2	1	1	54931	-10.83	4.18e-06	2.12e-01	6.44e-06	4.35e-01
2	1	1	0	1	3153	-10.90	2.08e-07	1.13e-02	2.27e-06	4.47e-01
1	2	1	1	1	6533	-11.13	3.53e-07	2.42e-02	2.06e-06	4.71e-01
1	1	2	0	1	38093	-11.49	1.49e-06	1.46e-01	1.70e-06	6.16e-01
1	2	1	0	1	3801	-11.79	1.26e-07	1.66e-02	2.11e-07	6.33e-01
1	1	0	1	1	1061	-11.80	2.80e-08	3.72e-03	8.45e-08	6.37e-01
1	1	0	0	1	570	-12.45	1.00e-08	2.56e-03	5.65e-08	6.39e-01
1	1	1	2	1	16409	-14.51	3.17e-08	6.37e-02	4.65e-08	7.03e-01
1	1	1	1	1	45717	-16.60	1.09e-08	1.76e-01	1.48e-08	8.79e-01
1	1	1	0	1	31641	-17.25	3.90e-09	1.21e-01	3.90e-09	1.00e+00

Table C-22: The result of a classification with comparisons agreement/disagreement/either missing (2/1/0) for road accidents in 2007. The candidate record pairs agree on the year of birth. The comparison vectors are ordered on the weight. The green comparison vectors are classified as positive links, the orange comparison vectors as possible links and the red comparison vectors as positive non-links. The estimated number of links N_M is 6977.

y^{hosp}	y^{epoch}	y^{gender}	y^{mot}	y^{dob}	$f(\boldsymbol{y})$	$w(\boldsymbol{y})$	$m(\boldsymbol{y})$	$u(\boldsymbol{y})$	λ	μ
2	2	2	2	2	4458	12.12	6.42e-01	3.51e-06	1.00e+00	3.51e-06
2	2	0	2	2	27	11.18	4.71e-03	6.57e-08	3.58e-01	3.58e-06
2	2	2	1	2	1339	9.97	1.95e-01	9.12e-06	3.53e-01	1.27e-05
2	2	2	0	2	489	9.48	7.15e-02	5.45e-06	1.58e-01	1.81e-05
2	2	0	1	2	13	9.04	1.43e-03	1.71e-07	8.63e-02	1.83e-05
2	2	0	0	2	5	8.55	5.24e-04	1.02e-07	8.48e-02	1.84e-05
2	2	1	2	2	15	6.76	2.49e-03	2.88e-06	8.43e-02	2.13e-05
2	1	2	2	2	143	6.71	2.06e-02	2.52e-05	8.18e-02	4.65e-05
1	2	2	2	2	82	5.82	1.17e-02	3.48e-05	6.12e-02	8.13e-05
2	1	0	2	2	0	5.77	1.51e-04	4.71e-07	4.95e-02	8.18e-05
1	2	0	2	2	2	4.88	8.59e-05	6.51e-07	4.94e-02	8.24e-05
2	2	1	1	2	12	4.62	7.56e-04	7.47e-06	4.93e-02	8.99e-05
2	1	2	1	2	60	4.56	6.25e-03	6.54e-05	4.85e-02	1.55e-04
2	2	1	0	2	1	4.13	2.76e-04	4.46e-06	4.23e-02	1.60e-04
2	1	2	0	2	34	4.07	2.29e-03	3.91e-05	4.20e-02	1.99e-04
1	2	2	1	2	49	3.67	3.56e-03	9.04e-05	3.97e-02	2.89e-04
2	1	0	1	2	1	3.62	4.59e-05	1.22e-06	3.62e-02	2.90e-04
1	2	2	0	2	30	3.18	1.30e-03	5.40e-05	3.61e-02	3.45e-04
2	1	0	0	2	0	3.13	1.68e-05	7.31e-07	3.48e-02	3.45e-04
2	2	2	2	1	441	2.95	2.26e-02	1.18e-03	3.48e-02	1.53e-03
1	2	0	1	2	3	2.74	2.61e-05	1.69e-06	1.22e-02	1.53e-03
1	2	0	0	2	0	2.25	9.55e-06	1.01e-06	1.21e-02	1.53e-03
2	2	0	2	1	9	2.02	1.66e-04	2.21e-05	1.21e-02	1.55e-03
2	1	1	2	2	6	1.35	7.96e-05	2.06e-05	1.20e-02	1.57e-03
2	2	2	1	1	840	0.81	6.89e-03	3.07e-03	1.19e-02	4.64e-03
1	2	1	2	2	10	0.46	4.53e-05	2.85e-05	4.99e-03	4.67e-03
1	1	2	2	2	65	0.41	3.75e-04	2.50e-04	4.95e-03	4.92e-03
2	2	2	0	1	392	0.32	2.52e-03	1.83e-03	4.57e-03	6.75e-03
2	2	0	1	1	14	-0.13	5.05e-05	5.74e-05	2.05e-03	6.81e-03
1	1	0	2	2	0	-0.53	2.75e-06	4.67e-06	2.00e-03	6.81e-03
2	2	0	0	1	12	-0.62	1.85e-05	3.43e-05	2.00e-03	6.85e-03
2	1	1	1	2	17	-0.79	2.42e-05	5.36e-05	1.98e-03	6.90e-03
2	1	1	0	2	9	-1.29	8.85e-06	3.20e-05	1.96e-03	6.93e-03
1	2	1	1	2	25	-1.68	1.38e-05	7.40e-05	1.95e-03	7.01e-03
1	1	2	1	2	145	-1.74	1.14e-04	6.49e-04	1.94e-03	7.66e-03
1	2	1	0	2	11	-2.17	5.04e-06	4.42e-05	1.82e-03	7.70e-03
1	1	2	0	2	90	-2.23	4.17e-05	3.87e-04	1.82e-03	8.09e-03
2	2	1	2	1	267	-2.40	8.76e-05	9.67e-04	1.77e-03	9.06e-03
2	1	2	2	1	1971	-2.46	7.25e-04	8.47e-03	1.69e-03	1.75e-02
1	1	0	1	2	2	-2.67	8.36e-07	1.21e-05	9.62e-04	1.75e-02
2	1	1	0	1	2823	-10.45	3.12e-07	1.08e-02	6.58e-06	2.51e-01
1	2	1	1	1	6095	-10.85	4.86e-07	2.49e-02	6.26e-06	2.76e-01
1	1	2	1	1	52625	-10.90	4.02e-06	2.18e-01	5.78e-06	4.94e-01
1	2	1	0	1	3255	-11.34	1.78e-07	1.49e-02	1.76e-06	5.09e-01
1	1	2	0	1	31875	-11.39	1.47e-06	1.30e-01	1.58e-06	6.39e-01
1	1	0	1	1	1038	-11.84	2.95e-08	4.08e-03	1.13e-07	6.43e-01
1	1	0	0	1	509	-12.33	1.08e-08	2.44e-03	8.31e-08	6.46e-01
1	1	1	2	1	16432	-14.11	5.11e-08	6.88e-02	7.24e-08	7.15e-01
1	1	1	1	1	43745	-16.26	1.55e-08	1.79e-01	2.12e-08	8.93e-01
1	1	1	0	1	26100	-16.75	5.69e-09	1.07e-01	5.69e-09	1.00e+00

Table C-23: The result of a classification with comparisons agreement/disagreement/either missing (2/1/0) for road accidents in 2008. The candidate record pairs agree on the year of birth. The comparison vectors are ordered on the weight. The green comparison vectors are classified as positive links, the orange comparison vectors as possible links and the red comparison vectors as positive non-links. The estimated number of links N_M is 6908.

y^{hosp}	y^{epoch}	y^{gender}	y^{mot}	y^{dob}	$f(\boldsymbol{y})$	$w(\boldsymbol{y})$	$m(\boldsymbol{y})$	$u(\boldsymbol{y})$	λ	μ
2	2	2	2	2	4091	12.06	6.32e-01	3.66e-06	1.00e+00	3.66e-06
2	2	0	2	2	151	11.28	2.25e-02	2.84e-07	3.68e-01	3.95e-06
2	2	2	1	2	1241	9.91	1.92e-01	9.51e-06	3.45e-01	1.35e-05
2	2	2	0	2	386	9.54	6.13e-02	4.40e-06	1.54e-01	1.79e-05
2	2	0	1	2	34	9.13	6.82e-03	7.37e-07	9.24e-02	1.86e-05
2	2	0	0	2	19	8.76	2.18e-03	3.41e-07	8.56e-02	1.89e-05
2	2	1	2	2	15	6.58	2.20e-03	3.06e-06	8.34e-02	2.20e-05
2	1	2	2	2	105	6.47	1.72e-02	2.65e-05	8.12e-02	4.85e-05
1	2	2	2	2	112	6.18	1.64e-02	3.41e-05	6.40e-02	8.26e-05
2	1	0	2	2	6	5.70	6.12e-04	2.05e-06	4.76e-02	8.46e-05
1	2	0	2	2	3	5.40	5.85e-04	2.64e-06	4.69e-02	8.73e-05
2	2	1	1	2	5	4.43	6.66e-04	7.93e-06	4.64e-02	9.52e-05
2	1	2	1	2	55	4.33	5.21e-03	6.88e-05	4.57e-02	1.64e-04
2	2	1	0	2	3	4.06	2.13e-04	3.67e-06	4.05e-02	1.68e-04
1	2	2	1	2	53	4.03	4.99e-03	8.84e-05	4.03e-02	2.56e-04
2	1	2	0	2	25	3.96	1.67e-03	3.18e-05	3.53e-02	2.88e-04
1	2	2	0	2	17	3.66	1.60e-03	4.09e-05	3.36e-02	3.29e-04
2	1	0	1	2	1	3.55	1.86e-04	5.33e-06	3.20e-02	3.34e-04
1	2	0	1	2	4	3.25	1.77e-04	6.85e-06	3.18e-02	3.41e-04
2	1	0	0	2	1	3.18	5.94e-05	2.47e-06	3.17e-02	3.44e-04
1	2	0	0	2	2	2.89	5.68e-05	3.17e-06	3.16e-02	3.47e-04
2	2	2	2	1	386	2.75	2.00e-02	1.28e-03	3.15e-02	1.63e-03
2	2	0	2	1	19	1.97	7.13e-04	9.93e-05	1.15e-02	1.73e-03
2	1	1	2	2	5	0.99	5.97e-05	2.21e-05	1.08e-02	1.75e-03
1	2	1	2	2	5	0.70	5.71e-05	2.84e-05	1.07e-02	1.78e-03
2	2	2	1	1	782	0.60	6.07e-03	3.33e-03	1.07e-02	5.10e-03
1	1	2	2	2	69	0.60	4.48e-04	2.47e-04	4.60e-03	5.35e-03
2	2	2	0	1	368	0.23	1.94e-03	1.54e-03	4.15e-03	6.89e-03
2	2	0	1	1	56	-0.18	2.16e-04	2.58e-04	2.21e-03	7.15e-03
1	1	0	2	2	3	-0.18	1.59e-05	1.91e-05	1.99e-03	7.17e-03
2	2	0	0	1	28	-0.54	6.92e-05	1.19e-04	1.97e-03	7.29e-03
2	1	1	1	2	11	-1.15	1.81e-05	5.74e-05	1.90e-03	7.34e-03
1	2	1	1	2	8	-1.45	1.73e-05	7.38e-05	1.89e-03	7.42e-03
2	1	1	0	2	3	-1.52	5.80e-06	2.66e-05	1.87e-03	7.44e-03
1	1	2	1	2	131	-1.55	1.36e-04	6.40e-04	1.86e-03	8.08e-03
1	2	1	0	2	3	-1.82	5.54e-06	3.41e-05	1.73e-03	8.12e-03
1	1	2	0	2	65	-1.92	4.34e-05	2.96e-04	1.72e-03	8.41e-03
1	1	0	1	2	10	-2.33	4.83e-06	4.96e-05	1.68e-03	8.46e-03
1	1	0	0	2	4	-2.70	1.54e-06	2.29e-05	1.67e-03	8.49e-03
2	2	1	2	1	282	-2.73	6.96e-05	1.07e-03	1.67e-03	9.55e-03
1	2	1	1	1	5539	-10.76	5.49e-07	2.58e-02	6.86e-06	2.81e-01
2	1	1	0	1	2240	-10.83	1.84e-07	9.29e-03	6.31e-06	2.90e-01
1	1	2	1	1	46361	-10.86	4.30e-06	2.24e-01	6.12e-06	5.14e-01
1	2	1	0	1	2183	-11.13	1.76e-07	1.19e-02	1.82e-06	5.26e-01
1	1	2	0	1	21904	-11.23	1.38e-06	1.04e-01	1.65e-06	6.29e-01
1	1	0	1	1	3755	-11.64	1.53e-07	1.74e-02	2.71e-07	6.47e-01
1	1	0	0	1	1606	-12.01	4.90e-08	8.03e-03	1.18e-07	6.55e-01
1	1	1	2	1	14913	-14.19	4.93e-08	7.20e-02	6.90e-08	7.27e-01
1	1	1	1	1	39323	-16.34	1.49e-08	1.87e-01	1.97e-08	9.14e-01
1	1	1	0	1	17531	-16.71	4.78e-09	8.64e-02	4.78e-09	1.00e+00

Table C-24: The result of a classification with comparisons agreement/disagreement/either missing (2/1/0) for road accidents in 2009. The candidate record pairs agree on the year of birth. The comparison vectors are ordered on the weight. The green comparison vectors are classified as positive links, the orange comparison vectors as possible links and the red comparison vectors as positive non-links. The estimated number of links N_M is 6455.

y^{hosp}	y^{epoch}	y^{gender}	y^{mot}	y^{dob}	$f(\boldsymbol{y})$	$w(\boldsymbol{y})$	$m(\boldsymbol{y})$	$u(\boldsymbol{y})$	λ	μ
2	2	2	2	2	2872	11.99	6.23e-01	3.86e-06	1.00e+00	3.86e-06
2	2	0	2	2	8	11.54	1.93e-03	1.88e-08	3.77e-01	3.88e-06
2	2	2	1	2	1125	10.01	2.44e-01	1.10e-05	3.75e-01	1.49e-05
2	2	0	1	2	6	9.56	7.57e-04	5.36e-08	1.30e-01	1.50e-05
2	2	2	0	2	168	9.19	3.58e-02	3.65e-06	1.30e-01	1.86e-05
2	2	0	0	2	0	8.74	1.11e-04	1.78e-08	9.38e-02	1.86e-05
2	1	2	2	2	104	6.62	2.07e-02	2.75e-05	9.37e-02	4.61e-05
2	2	1	2	2	8	6.55	2.29e-03	3.28e-06	7.30e-02	4.94e-05
2	1	0	2	2	0	6.17	6.41e-05	1.33e-07	7.07e-02	4.95e-05
1	2	2	2	2	61	5.84	1.22e-02	3.54e-05	7.07e-02	8.49e-05
1	2	0	2	2	0	5.39	3.79e-05	1.72e-07	5.85e-02	8.51e-05
2	1	2	1	2	40	4.64	8.10e-03	7.83e-05	5.84e-02	1.63e-04
2	2	1	1	2	8	4.56	8.98e-04	9.36e-06	5.03e-02	1.73e-04
2	1	0	1	2	0	4.19	2.51e-05	3.81e-07	4.94e-02	1.73e-04
1	2	2	1	2	33	3.86	4.78e-03	1.01e-04	4.94e-02	2.74e-04
2	1	2	0	2	8	3.82	1.19e-03	2.60e-05	4.46e-02	3.00e-04
2	2	1	0	2	2	3.75	1.32e-04	3.10e-06	4.34e-02	3.03e-04
1	2	0	1	2	0	3.41	1.48e-05	4.91e-07	4.33e-02	3.04e-04
2	1	0	0	2	0	3.37	3.68e-06	1.26e-07	4.33e-02	3.04e-04
1	2	2	0	2	6	3.04	7.02e-04	3.35e-05	4.33e-02	3.37e-04
2	2	2	2	1	284	3.00	2.72e-02	1.36e-03	4.26e-02	1.70e-03
1	2	0	0	2	0	2.59	2.18e-06	1.63e-07	1.54e-02	1.70e-03
2	2	0	2	1	0	2.55	8.43e-05	6.61e-06	1.54e-02	1.70e-03
2	1	1	2	2	2	1.18	7.60e-05	2.33e-05	1.53e-02	1.73e-03
2	2	2	1	1	540	1.01	1.07e-02	3.88e-03	1.52e-02	5.60e-03
2	2	0	1	1	2	0.56	3.30e-05	1.88e-05	4.56e-03	5.62e-03
1	1	2	2	2	47	0.48	4.05e-04	2.52e-04	4.52e-03	5.87e-03
1	2	1	2	2	3	0.40	4.49e-05	3.01e-05	4.12e-03	5.90e-03
2	2	2	0	1	125	0.20	1.56e-03	1.29e-03	4.07e-03	7.19e-03
1	1	0	2	2	0	0.03	1.26e-06	1.22e-06	2.51e-03	7.19e-03
2	2	0	0	1	1	-0.25	4.85e-06	6.25e-06	2.51e-03	7.20e-03
2	1	1	1	2	9	-0.80	2.98e-05	6.65e-05	2.50e-03	7.26e-03
1	1	2	1	2	83	-1.51	1.59e-04	7.18e-04	2.47e-03	7.98e-03
1	2	1	1	2	11	-1.59	1.76e-05	8.59e-05	2.31e-03	8.07e-03
2	1	1	0	2	4	-1.62	4.37e-06	2.21e-05	2.30e-03	8.09e-03
1	1	0	1	2	1	-1.96	4.92e-07	3.49e-06	2.29e-03	8.09e-03
1	1	2	0	2	20	-2.32	2.33e-05	2.38e-04	2.29e-03	8.33e-03
2	1	2	2	1	1181	-2.37	9.02e-04	9.66e-03	2.27e-03	1.80e-02
1	2	1	0	2	2	-2.40	2.58e-06	2.85e-05	1.37e-03	1.80e-02
2	2	1	2	1	191	-2.45	1.00e-04	1.16e-03	1.36e-03	1.92e-02
1	1	2	1	1	29929	-10.51	6.92e-06	2.53e-01	9.13e-06	5.05e-01
1	2	1	1	1	3620	-10.58	7.68e-07	3.02e-02	2.21e-06	5.36e-01
2	1	1	0	1	924	-10.61	1.91e-07	7.76e-03	1.44e-06	5.43e-01
1	1	0	1	1	132	-10.96	2.15e-08	1.23e-03	1.25e-06	5.45e-01
1	1	2	0	1	10336	-11.32	1.02e-06	8.38e-02	1.23e-06	6.28e-01
1	2	1	0	1	1047	-11.39	1.13e-07	1.00e-02	2.10e-07	6.38e-01
1	1	0	0	1	47	-11.77	3.15e-09	4.07e-04	9.73e-08	6.39e-01
1	1	1	2	1	8694	-13.96	6.50e-08	7.53e-02	9.42e-08	7.14e-01
1	1	1	1	1	25837	-15.95	2.55e-08	2.15e-01	2.92e-08	9.29e-01
1	1	1	0	1	8418	-16.76	3.74e-09	7.12e-02	3.74e-09	1.00e+00

Table C-25: The result of a classification with comparisons agreement/disagreement/either missing (2/1/0) for road accidents in 2010. The candidate record pairs agree on the year of birth. The comparison vectors are ordered on the weight. The green comparison vectors are classified as positive links, the orange comparison vectors as possible links and the red comparison vectors as positive non-links. The estimated number of links N_M is 4607.

y^{hosp}	y^{epoch}	y^{gender}	y^{mot}	y^{dob}	$f(\boldsymbol{y})$	$w(\boldsymbol{y})$	$m(\boldsymbol{y})$	$u(\boldsymbol{y})$	λ	μ
2	2	2	2	2	1555	12.15	6.56e-01	3.46e-06	1.00e+00	3.46e-06
2	2	0	2	2	2	10.89	8.46e-04	1.58e-08	3.44e-01	3.48e-06
2	2	2	1	2	547	9.94	2.36e-01	1.14e-05	3.43e-01	1.49e-05
2	2	2	0	2	55	9.11	2.43e-02	2.69e-06	1.07e-01	1.76e-05
2	2	0	1	2	1	8.67	3.04e-04	5.20e-08	8.29e-02	1.76e-05
2	2	0	0	2	0	7.84	3.13e-05	1.23e-08	8.26e-02	1.76e-05
2	2	1	2	2	4	6.35	1.65e-03	2.88e-06	8.26e-02	2.05e-05
2	1	2	2	2	28	6.32	1.38e-02	2.48e-05	8.09e-02	4.53e-05
1	2	2	2	2	35	6.23	1.59e-02	3.13e-05	6.72e-02	7.66e-05
2	1	0	2	2	0	5.06	1.78e-05	1.13e-07	5.12e-02	7.67e-05
1	2	0	2	2	0	4.97	2.05e-05	1.43e-07	5.12e-02	7.68e-05
2	2	1	1	2	2	4.14	5.93e-04	9.46e-06	5.12e-02	8.63e-05
2	1	2	1	2	20	4.11	4.95e-03	8.15e-05	5.06e-02	1.68e-04
1	2	2	1	2	24	4.02	5.72e-03	1.03e-04	4.56e-02	2.70e-04
2	2	1	0	2	0	3.31	6.11e-05	2.24e-06	3.99e-02	2.73e-04
2	1	2	0	2	7	3.28	5.10e-04	1.93e-05	3.99e-02	2.92e-04
1	2	2	0	2	3	3.19	5.90e-04	2.43e-05	3.93e-02	3.16e-04
2	2	2	2	1	129	2.99	2.60e-02	1.31e-03	3.88e-02	1.63e-03
2	1	0	1	2	0	2.84	6.39e-06	3.72e-07	1.27e-02	1.63e-03
1	2	0	1	2	0	2.76	7.38e-06	4.69e-07	1.27e-02	1.63e-03
2	1	0	0	2	0	2.01	6.58e-07	8.80e-08	1.27e-02	1.63e-03
1	2	0	0	2	0	1.92	7.61e-07	1.11e-07	1.27e-02	1.63e-03
2	2	0	2	1	0	1.72	3.36e-05	5.99e-06	1.27e-02	1.63e-03
2	2	2	1	1	278	0.78	9.36e-03	4.31e-03	1.27e-02	5.94e-03
2	1	1	2	2	1	0.52	3.47e-05	2.06e-05	3.31e-03	5.96e-03
1	2	1	2	2	2	0.43	4.01e-05	2.60e-05	3.28e-03	5.99e-03
1	1	2	2	2	17	0.40	3.35e-04	2.24e-04	3.24e-03	6.21e-03
2	2	2	0	1	55	-0.06	9.64e-04	1.02e-03	2.90e-03	7.23e-03
2	2	0	1	1	1	-0.49	1.21e-05	1.97e-05	1.94e-03	7.25e-03
1	1	0	2	2	0	-0.86	4.32e-07	1.02e-06	1.93e-03	7.25e-03
2	2	0	0	1	0	-1.32	1.24e-06	4.66e-06	1.93e-03	7.26e-03
2	1	1	1	2	6	-1.69	1.25e-05	6.77e-05	1.93e-03	7.33e-03
1	2	1	1	2	6	-1.78	1.44e-05	8.54e-05	1.91e-03	7.41e-03
1	1	2	1	2	44	-1.81	1.20e-04	7.36e-04	1.90e-03	8.15e-03
2	1	1	0	2	1	-2.52	1.28e-06	1.60e-05	1.78e-03	8.16e-03
1	2	1	0	2	0	-2.61	1.48e-06	2.02e-05	1.78e-03	8.18e-03
1	1	2	0	2	6	-2.64	1.24e-05	1.74e-04	1.78e-03	8.36e-03
2	2	1	2	1	91	-2.81	6.55e-05	1.09e-03	1.76e-03	9.45e-03
2	1	2	2	1	532	-2.84	5.47e-04	9.39e-03	1.70e-03	1.88e-02
1	2	2	2	1	726	-2.93	6.32e-04	1.18e-02	1.15e-03	3.07e-02
1	2	1	1	1	1767	-10.94	5.71e-07	3.23e-02	6.00e-06	2.84e-01
1	1	2	1	1	15403	-10.97	4.77e-06	2.79e-01	5.43e-06	5.62e-01
2	1	1	0	1	351	-11.69	5.10e-08	6.07e-03	6.55e-07	5.68e-01
1	2	1	0	1	364	-11.77	5.89e-08	7.65e-03	6.04e-07	5.76e-01
1	1	2	0	1	3653	-11.81	4.92e-07	6.59e-02	5.45e-07	6.42e-01
1	1	0	1	1	76	-12.24	6.16e-09	1.27e-03	5.34e-08	6.43e-01
1	1	0	0	1	18	-13.07	6.34e-10	3.01e-04	4.73e-08	6.43e-01
1	1	1	2	1	3885	-14.56	3.34e-08	7.04e-02	4.67e-08	7.14e-01
1	1	1	1	1	12534	-16.77	1.20e-08	2.31e-01	1.32e-08	9.45e-01
1	1	1	0	1	3061	-17.61	1.24e-09	5.48e-02	1.24e-09	1.00e+00

Table C-26: The result of a classification with comparisons agreement/disagreement/either missing (2/1/0) for road accidents in 2011. The candidate record pairs agree on the year of birth. The comparison vectors are ordered on the weight. The green comparison vectors are classified as positive links, the orange comparison vectors as possible links and the red comparison vectors as positive non-links. The estimated number of links N_M is 2352.

y^{hosp}	y^{epoch}	y^{gender}	y^{mot}	y^{dob}	$f(\boldsymbol{y})$	$w(\boldsymbol{y})$	$m(\boldsymbol{y})$	$u(\boldsymbol{y})$	λ	μ
2	2	2	2	2	1422	11.92	6.44e-01	4.26e-06	1.00e+00	4.26e-06
2	2	2	1	2	568	9.78	2.59e-01	1.46e-05	3.56e-01	1.89e-05
2	2	0	2	2	0	9.40	2.95e-04	2.45e-08	9.73e-02	1.89e-05
2	2	2	0	2	37	8.70	1.72e-02	2.88e-06	9.70e-02	2.18e-05
2	2	0	1	2	1	7.26	1.19e-04	8.38e-08	7.98e-02	2.19e-05
2	1	2	2	2	40	6.39	1.82e-02	3.04e-05	7.97e-02	5.23e-05
2	2	1	2	2	5	6.20	1.76e-03	3.55e-06	6.15e-02	5.58e-05
2	2	0	0	2	0	6.17	7.89e-06	1.65e-08	5.97e-02	5.58e-05
1	2	2	2	2	23	5.74	1.17e-02	3.76e-05	5.97e-02	9.35e-05
2	1	2	1	2	22	4.25	7.32e-03	1.04e-04	4.80e-02	1.98e-04
2	2	1	1	2	0	4.06	7.08e-04	1.22e-05	4.07e-02	2.10e-04
2	1	0	2	2	0	3.86	8.33e-06	1.75e-07	4.00e-02	2.10e-04
1	2	2	1	2	21	3.60	4.73e-03	1.29e-04	3.99e-02	3.39e-04
1	2	0	2	2	0	3.21	5.38e-06	2.16e-07	3.52e-02	3.39e-04
2	1	2	0	2	2	3.17	4.87e-04	2.05e-05	3.52e-02	3.60e-04
2	2	1	0	2	1	2.98	4.71e-05	2.40e-06	3.47e-02	3.62e-04
2	2	2	2	1	106	2.83	2.25e-02	1.33e-03	3.47e-02	1.69e-03
1	2	2	0	2	3	2.52	3.14e-04	2.54e-05	1.22e-02	1.72e-03
2	1	0	1	2	0	1.72	3.35e-06	5.98e-07	1.19e-02	1.72e-03
1	2	0	1	2	0	1.07	2.16e-06	7.40e-07	1.19e-02	1.72e-03
2	2	2	1	1	208	0.69	9.06e-03	4.55e-03	1.19e-02	6.27e-03
2	1	1	2	2	1	0.67	4.97e-05	2.53e-05	2.81e-03	6.29e-03
2	1	0	0	2	0	0.64	2.23e-07	1.18e-07	2.76e-03	6.29e-03
2	2	0	2	1	0	0.30	1.03e-05	7.62e-06	2.76e-03	6.30e-03
1	1	2	2	2	16	0.21	3.31e-04	2.69e-04	2.75e-03	6.57e-03
1	2	1	2	2	4	0.02	3.21e-05	3.14e-05	2.42e-03	6.60e-03
1	2	0	0	2	0	-0.01	1.44e-07	1.46e-07	2.39e-03	6.60e-03
2	2	2	0	1	44	-0.40	6.02e-04	8.96e-04	2.39e-03	7.50e-03
2	1	1	1	2	5	-1.47	2.00e-05	8.69e-05	1.79e-03	7.58e-03
2	2	0	1	1	2	-1.84	4.15e-06	2.61e-05	1.77e-03	7.61e-03
1	1	2	1	2	28	-1.93	1.33e-04	9.20e-04	1.76e-03	8.53e-03
1	2	1	1	2	2	-2.12	1.29e-05	1.08e-04	1.63e-03	8.64e-03
1	1	0	2	2	0	-2.32	1.52e-07	1.54e-06	1.62e-03	8.64e-03
2	1	1	0	2	0	-2.55	1.33e-06	1.71e-05	1.62e-03	8.66e-03
2	1	2	2	1	398	-2.70	6.35e-04	9.48e-03	1.62e-03	1.81e-02
2	2	1	2	1	83	-2.89	6.15e-05	1.11e-03	9.80e-04	1.92e-02
2	2	0	0	1	0	-2.93	2.76e-07	5.14e-06	9.18e-04	1.92e-02
1	1	2	0	2	6	-3.02	8.88e-06	1.81e-04	9.18e-04	1.94e-02
1	2	1	0	2	3	-3.21	8.59e-07	2.12e-05	9.09e-04	1.94e-02
1	2	2	2	1	540	-3.35	4.10e-04	1.17e-02	9.08e-04	3.12e-02
1	2	1	1	1	1405	-11.22	4.51e-07	3.35e-02	8.91e-07	5.73e-01
1	1	0	2	1	17	-11.41	5.31e-09	4.80e-04	4.40e-07	5.74e-01
2	1	1	0	1	276	-11.65	4.65e-08	5.33e-03	4.34e-07	5.79e-01
1	1	2	0	1	2344	-12.11	3.10e-07	5.65e-02	3.88e-07	6.36e-01
1	2	1	0	1	247	-12.30	3.00e-08	6.60e-03	7.75e-08	6.42e-01
1	1	0	1	1	72	-13.55	2.14e-09	1.65e-03	4.75e-08	6.44e-01
1	1	1	2	1	2968	-14.61	3.17e-08	6.98e-02	4.54e-08	7.14e-01
1	1	0	0	1	14	-14.64	1.42e-10	3.24e-04	1.37e-08	7.14e-01
1	1	1	1	1	10085	-16.75	1.27e-08	2.39e-01	1.36e-08	9.53e-01
1	1	1	0	1	2005	-17.83	8.48e-10	4.71e-02	8.48e-10	1.00e+00

Table C-27: The result of a classification with comparisons agreement/disagreement/either missing (2/1/0) for road accidents in 2012. The candidate record pairs agree on the year of birth. The comparison vectors are ordered on the weight. The green comparison vectors are classified as positive links, the orange comparison vectors as possible links and the red comparison vectors as positive non-links. The estimated number of links N_M is 2202.

y^{hosp}	y^{epoch}	y^{gender}	y^{mot}	y^{dob}	$f(\boldsymbol{y})$	$w(\boldsymbol{y})$	$m(\boldsymbol{y})$	$u(\boldsymbol{y})$	λ	μ
2	2	2	2	2	1430	12.38	5.39e-01	2.27e-06	1.00e+00	2.27e-06
2	2	0	2	2	5	11.44	1.74e-03	1.88e-08	4.61e-01	2.29e-06
2	2	2	0	2	382	10.61	1.58e-01	3.89e-06	4.59e-01	6.18e-06
2	2	2	1	2	453	9.87	1.69e-01	8.71e-06	3.01e-01	1.49e-05
2	2	0	0	2	2	9.67	5.10e-04	3.22e-08	1.33e-01	1.49e-05
2	2	0	1	2	1	8.93	5.45e-04	7.20e-08	1.32e-01	1.50e-05
1	2	2	2	2	103	7.56	4.82e-02	2.50e-05	1.31e-01	4.00e-05
2	1	2	2	2	43	6.88	1.63e-02	1.68e-05	8.33e-02	5.68e-05
1	2	0	2	2	0	6.62	1.56e-04	2.07e-07	6.70e-02	5.70e-05
2	1	0	2	2	0	5.93	5.26e-05	1.39e-07	6.69e-02	5.72e-05
1	2	2	0	2	78	5.80	1.41e-02	4.28e-05	6.68e-02	9.99e-05
2	1	2	0	2	11	5.11	4.77e-03	2.88e-05	5.27e-02	1.29e-04
1	2	2	1	2	30	5.06	1.51e-02	9.58e-05	4.79e-02	2.25e-04
1	2	0	0	2	0	4.86	4.56e-05	3.54e-07	3.28e-02	2.25e-04
2	1	2	1	2	19	4.37	5.10e-03	6.45e-05	3.28e-02	2.89e-04
2	1	0	0	2	0	4.17	1.54e-05	2.38e-07	2.77e-02	2.90e-04
1	2	0	1	2	0	4.12	4.87e-05	7.92e-07	2.77e-02	2.90e-04
2	1	0	1	2	0	3.43	1.65e-05	5.33e-07	2.76e-02	2.91e-04
2	2	2	2	1	88	2.84	1.40e-02	8.12e-04	2.76e-02	1.10e-03
1	1	2	2	2	25	2.06	1.46e-03	1.85e-04	1.36e-02	1.29e-03
2	2	0	2	1	1	1.90	4.51e-05	6.72e-06	1.22e-02	1.30e-03
1	1	0	2	2	1	1.12	4.70e-06	1.53e-06	1.21e-02	1.30e-03
2	2	2	0	1	77	1.08	4.09e-03	1.39e-03	1.21e-02	2.69e-03
2	2	2	1	1	228	0.34	4.38e-03	3.11e-03	8.04e-03	5.80e-03
1	1	2	0	2	15	0.30	4.27e-04	3.17e-04	3.66e-03	6.12e-03
2	2	0	0	1	0	0.14	1.32e-05	1.15e-05	3.24e-03	6.13e-03
1	1	2	1	2	46	-0.44	4.56e-04	7.09e-04	3.23e-03	6.84e-03
2	2	0	1	1	3	-0.60	1.41e-05	2.57e-05	2.77e-03	6.86e-03
1	1	0	0	2	0	-0.64	1.38e-06	2.62e-06	2.75e-03	6.87e-03
1	1	0	1	2	0	-1.38	1.47e-06	5.86e-06	2.75e-03	6.87e-03
1	2	2	2	1	552	-1.97	1.25e-03	8.93e-03	2.75e-03	1.58e-02
2	1	2	2	1	407	-2.66	4.22e-04	6.02e-03	1.50e-03	2.18e-02
1	2	0	2	1	2	-2.91	4.03e-06	7.39e-05	1.08e-03	2.19e-02
2	1	0	2	1	1	-3.60	1.36e-06	4.98e-05	1.08e-03	2.19e-02
1	2	2	0	1	821	-3.73	3.66e-04	1.53e-02	1.08e-03	3.72e-02
2	1	2	0	1	578	-4.42	1.24e-04	1.03e-02	7.11e-04	4.75e-02
1	2	2	1	1	2215	-4.47	3.91e-04	3.42e-02	5.87e-04	8.18e-02
1	2	0	0	1	4	-4.67	1.18e-06	1.26e-04	1.96e-04	8.19e-02
2	1	2	1	1	1420	-5.16	1.32e-04	2.31e-02	1.95e-04	1.05e-01
2	1	0	0	1	6	-5.36	3.99e-07	8.52e-05	6.29e-05	1.05e-01
1	1	1	1	2	35	-106.80	2.51e-50	6.04e-04	1.76e-49	5.48e-01
1	2	1	2	1	501	-108.32	6.88e-50	7.61e-03	1.51e-49	5.56e-01
2	1	1	2	1	352	-109.01	2.33e-50	5.12e-03	8.24e-50	5.61e-01
1	2	1	0	1	757	-110.09	2.02e-50	1.30e-02	5.91e-50	5.74e-01
2	1	1	0	1	488	-110.78	6.82e-51	8.77e-03	3.90e-50	5.83e-01
1	2	1	1	1	1772	-110.83	2.15e-50	2.91e-02	3.22e-50	6.12e-01
2	1	1	1	1	1183	-111.52	7.29e-51	1.96e-02	1.06e-50	6.31e-01
1	1	1	2	1	3434	-113.82	2.08e-51	5.63e-02	3.34e-51	6.88e-01
1	1	1	0	1	6231	-115.59	6.09e-52	9.65e-02	1.26e-51	7.84e-01
1	1	1	1	1	13053	-116.33	6.51e-52	2.16e-01	6.51e-52	1.00e+00

Table C-28: The result of a classification with comparisons agreement/disagreement/either missing (2/1/0) for road accidents in 2013. The candidate record pairs agree on the year of birth. The comparison vectors are ordered on the weight. The green comparison vectors are classified as positive links, the orange comparison vectors as possible links and the red comparison vectors as positive non-links. The estimated number of links N_M is 2613.

Appendix D

Python code for record linkage

In the beginning of this research, the code of the Freely Extensible Biomedical Record Linkage (Febrl) package was used to experiment. This python package has a user interface and is extensible [Christen, 2008]. The package was not sufficient for this thesis and is has no (un-)supervised learning possibilities. The essential code used for this thesis is the code needed for the ECM-algorithm and the frequency based EM-algorithm. The code for this Python class is given in this appendix.

D-1 Class for estimating parameters with the EM-algorithm

```
1  # estimation.py
2
3  # import for Python 2.7
4  from __future__ import division
5
6  import time
7  import copy
8
9  import pandas as pd
10 import numpy as np
11
12 from sklearn.utils.extmath import cartesian
13
14 class EMEstimate(object):
15
16     def __init__(self, comparison_vectors, start_m, start_u, start_p):
17
18 #         Set comparison vectors
19         self.comparison_vectors = comparison_vectors
20
21 #         Set first iteration
```

```
22          self.m = start_m
23          self.u = start_u
24          self.p = start_p
25
26          # Count the number of iterations
27          self.iteration = 0
28
29          self.comparison_space = pd.DataFrame({'count' : self.
                comparison_vectors.groupby(list(self.comparison_vectors)).size
                ()}).reset_index()
30
31      def estimate(self, max_iter=100, log=False):
32
33          self.max_iter = max_iter
34
35          while self.iteration < self.max_iter:
36
37              # Compute expectation
38              self.g = self._expectation(self.comparison_space[list(self.
                    comparison_vectors)])
39
40              # Maximize
41              self.m, self.u, self.p = self._maximization(self.
                    comparison_space[list(self.comparison_vectors)], self.
                    comparison_space['count'], self.g)
42
43              # Increment counter
44              self.iteration = self.iteration+1
45
46
47      def _maximizion(self):
48
49          """ To be overwritten """
50
51          pass
52
53      def _expectation(self):
54
55          """ To be overwritten """
56
57          pass
58
59      def m_prob(self, y):
60
61          """ To be overwritten """
62
63          pass
64
65      def u_prob(self, y):
66
67          """ To be overwritten """
68
69          pass
```

```
70
71      def weights(self, y):
72
73          return np.log(np.divide(self.m_prob(y), self.u_prob(y)))
74
75      def summary(self):
76
77          summary = pd.merge(self.cartesian(), self.comparison_space, on=
                list(self.comparison_vectors), how='left').fillna(0)
78
79          summary['m'] = self.m_prob(summary[list(self.comparison_vectors)
                ])
80          summary['u'] = self.u_prob(summary[list(self.comparison_vectors)
                ])
81          summary['w'] = self.weights(summary[list(self.comparison_vectors)
                ])
82          summary['g'] = self._expectation(summary[list(self.
                comparison_vectors)])
83
84          summary.sort('w', ascending=True, inplace=True)
85          summary['lambda'] = summary['m'].cumsum()
86
87          summary.sort('w', ascending=False, inplace=True)
88          summary['mu'] = summary['u'].cumsum()
89
90          return summary
91
92      def cartesian(self): # aanpassen max. Moet unique worden..
93
94          # Cartesian product of all possible options
95
96          max_tuple = []
97
98          for col in list(self.comparison_vectors):
99
100             max_tuple.append(self.comparison_vectors[col].unique())
101
102         y_cart = pd.DataFrame(cartesian(max_tuple)) #
                ([0,1],[0,1],[0,1],[0,1],[0,1],[0,1],[0,1],[0,1])
103         y_cart.columns = list(self.comparison_vectors)
104
105         return y_cart
106
107 class ECMEstimate(EMEstimate):
108
109     def _maximization(self, y, f, g):
110
111         for col in y.columns:
112
113             for level in y[col].unique():
114
115                 # Maximization of m
116                 self.m[col][level] = sum((g*f)[y[col] == level])/sum(g*f)
```

```
117
118                 # Maximization of u
119                 self.u[col][level] = sum((((1-g)*f)[y[col] == level]))/
                        sum((1-g)*f)
120
121         # Maximization of p
122         self.p = sum(g*f)/sum(f)
123
124         return self.m, self.u, self.p
125
126     def _expectation(self, y):
127
128         return self.p*self.m_prob(y)/(self.p*self.m_prob(y)+(1-self.p)*
                self.u_prob(y))
129
130     def m_prob(self, y):
131
132         return y.replace(self.m).prod(axis=1)
133
134     def u_prob(self, y):
135
136         return y.replace(self.u).prod(axis=1)
```

D-2 Using the estimation class

An example to use the ECM-algorithm and frequency based EM-algorithm are given in this section.

```
import pandas as pd

import estimation

# --------------------------------------
# EXAMPLE BINARY ASSUMPTION
# --------------------------------------

# Comparison vectors with label 2 for agreement and 1 for disagreement
y = "YOUR PANDAS DATAFRAME"

# Define a dict with the starting values for the m marginal probability
    mass functions
m_start = {'y_1': {1: 0.1, 2: 0.9},
           'y_2': {1: 0.1, 2: 0.9},
           'y_3': {1: 0.1, 2: 0.9},
           'y_4': {1: 0.1, 2: 0.9},
           'y_5': {1: 0.1, 2: 0.9}}

# Define a dict with the starting values for the u marginal probability
    mass functions
u_start = {'y_1': {1: 0.9, 2: 0.1},
           'y_2': {1: 0.9, 2: 0.1},
           'y_3': {1: 0.9, 2: 0.1},
           'y_4': {1: 0.9, 2: 0.1},
           'y_5': {1: 0.9, 2: 0.1}}

# set the match prevalence
p_start = 0.1

# Start an estimation with 150 iterations
est = estimation.ECMEstimate(y, m_start, u_start, p_start, max_iter =
    150)
est.estimate()

# Print the summary (similar with the output in this thesis)
est_sum = est.summary()

# Print the parameters m, u and p
print est.m
print est.u
print est.p

# --------------------------------------
# EXAMPLE FREQUENCY BASED ESTIMATION
# --------------------------------------

```

```
45 # Comparison vectors
46 y = "YOUR PANDAS DATAFRAME"
47
48 # Define a dict with the starting values for the m marginal probability
       mass functions
49 m_start = {'y_1': {1: 0.1, 2: 0.9},
50            'y_2': {1: 0.1, 2: 0.1, 3: 0.1, 4: 0.15, 5: 0.1, 6: 0.1, 7:
                  0.15, 8: 0.1, 9: 0.1},
51            'y_3': {0: 0.2, 1: 0.1, 2: 0.7},
52            'y_4': {1: 0.1, 2: 0.9},
53            'y_5': {1: 0.2, 2: 0.1, 3: 0.2, 4: 0.1, 5: 0.4}}
54
55 # Define a dict with the starting values for the u marginal probability
       mass functions
56 u_start = {'y_1': {1: 0.9, 2: 0.1},
57            'y_2': {1: 0.3, 2: 0.3, 3: 0.05, 4: 0.15, 5: 0.05, 6: 0.05, 7:
                  0.05, 8: 0.05, 9: 0},
58            'y_3': {0: 0.2, 1: 0.7, 2: 0.1},
59            'y_4': {1: 0.9, 2: 0.1},
60            'y_5': {1: 0.4, 2: 0.2, 3: 0.1, 4: 0.1, 5: 0.2}}
61
62 # set the match prevalence
63 p_start = 0.1
64
65 # Start an estimation with 150 iterations
66 est = estimation.ECMEstimate(y, m_start, u_start, p_start, max_iter =
       150)
67 est.estimate()
68
69 # Print the summary (similar with the output in this thesis)
70 est_sum = est.summary()
71
72 # Print the parameters m, u and p
73 print est.m
74 print est.u
75 print est.p
```

Bibliography

Amoros, E., International Traffic Safety Data, Analysis Group, B. Noble and International Transport Forum. 2011. *Reporting on Serious Road Traffic Casualties: Combining and Using Different Data Sources to Improve Understanding of Non-fatal Road Traffic Crashes*. International Transport Forum.

Augsten, N. and M. Bohlen. 2013. *Similarity Joins in Relational Database Systems*. Synthesis Lectures on Data Management Morgan & Claypool Publishers.

Baxter, Rohan, Peter Christen and Tim Churches. 2003. "A comparison of fast blocking methods for record linkage." 3:25–27.

Centraal Bureau voor de Statistiek. 2015. "Documentatierapport Landelijke Medische Registratie (LMR) 2011.".

CHeReL. 2015. "Centre for Health Record Linkage.".
URL: *http://www.cherel.org.au/*

Christen, Peter. 2005. Probabilistic data generation for deduplication and data linkage. In *Intelligent Data Engineering and Automated Learning-IDEAL 2005*. Springer pp. 109–116.

Christen, Peter. 2008. "Febrl: a freely available record linkage system with a graphical user interface." pp. 17–25.

Christen, Peter. 2012*a*. *Data matching: concepts and techniques for record linkage, entity resolution, and duplicate detection*. Springer Science & Business Media.

Christen, Peter. 2012*b*. "A survey of indexing techniques for scalable record linkage and deduplication." *Knowledge and Data Engineering, IEEE Transactions on* 24(9):1537–1555.

Christen, Peter and Karl Goiser. 2007. "Quality and complexity measures for data linkage and deduplication." pp. 127–151.

Daggy, Joanne K, Huiping Xu, Siu L Hui, Roland E Gamache and Shaun J Grannis. 2013. "A practical approach for incorporating dependence among fields in probabilistic record linkage." *BMC Med. Inf. & Decision Making* 13:97.

Dempster, Arthur P, Nan M Laird and Donald B Rubin. 1977. "Maximum likelihood from incomplete data via the EM algorithm." *Journal of the Royal Statistical Society. Series B (Methodological)* pp. 1–38.

Dunn, Halbert L. 1946. "Record linkage." *American Journal of Public Health and the Nations Health* 36(12):1412–1416.

Dutch Medical Data. 2015. "Landelijke Medische Registratie (LMR).".
URL: *www.dutchhospitaldata.nl/registraties/lmrlazr/*

Elfeky, Mohamed G, Vassilios S Verykios and Ahmed K Elmagarmid. 2002. "TAILOR: A record linkage toolbox." pp. 17–28.

Fellegi, Ivan P and Alan B Sunter. 1969. "A theory for record linkage." *Journal of the American Statistical Association* 64(328):1183–1210.

Gill. 2001. *Methods for automatic record matching and linkage and their use in national statistics*. London : National Statistics.

Gomatam, Shanti and Michael D Larsen. 2004. "Record linkage and counterterrorism." *Chance* 17(1):25–29.

Hernández, Mauricio A and Salvatore J Stolfo. 1995. "The merge/purge problem for large databases." 24(2):127–138.

Herzog, Thomas N, Fritz J Scheuren and William E Winkler. 2007. *Data quality and record linkage techniques*. Vol. 1 Springer.

Hilbert, Martin and Priscila López. 2011. "The world's technological capacity to store, communicate, and compute information." *science* 332(6025):60–65.

IRTAD. 2011. *Reporting on serious road traffic casualties : combining and using different data sources to improve understanding of non-fatal road traffic crashes*.

Jaro, Matthew A. 1989. "Advances in record-linkage methodology as applied to matching the 1985 census of Tampa, Florida." *Journal of the American Statistical Association* 84(406):414–420.

Levenshtein, Vladimir I. 1966. "Binary codes capable of correcting deletions, insertions, and reversals." 10(8):707–710.

McLachlan, Geoffrey and Thriyambakam Krishnan. 2007. *The EM algorithm and extensions*. Vol. 382 John Wiley & Sons.

Meng, Xiao-Li and Donald B Rubin. 1993. "Maximum likelihood estimation via the ECM algorithm: A general framework." *Biometrika* 80(2):267–278.

Ministerie van Infrastructuur en Milieu. 2015. "Bestand geRegistreerde Ongevallen Nederland.".
URL: *https://data.overheid.nl/data/dataset/bestand-geregistreerde-ongevallen-nederland*

Newcombe, HB, JM Kennedy, SJ Axford and AP James. 1986. "Automatic linkage of vital records." 1299:7.

Newcombe, Howard B and James M Kennedy. 1962. "Record linkage: making maximum use of the discriminating power of identifying information." *Communications of the ACM* 5(11):563–566.

Newcombe, Howard B, James M Kennedy, SJ Axford and Allison P James. 1959. "Automatic Linkage of Vital Records Computers can be used to extract " follow-up" statistics of families from files of routine records." *Science* 130(3381):954–959.

Neyman, Jerzy and Egon S Pearson. 1933. "The testing of statistical hypotheses in relation to probabilities a priori." 29(04):492–510.

Omroep West. 2015. "Omroep West.".
URL: *http://www.omroepwest.nl*

Paas, G. R. A. and K. C. W. Veenhuizen. 2002. *Onderzoek naar de betrouwbaarheid van de Landelijke Medische Registratie LMR.*

Parmigiani, Giovanni and Lurdes Inoue. 2009. *Decision theory: principles and approaches.* Vol. 812 John Wiley & Sons.

Politie. 2015. "Aanrijding.".
URL: *https://www.politie.nl/themas/aanrijding.html*

Puckett, Carolyn. 2009. "The Story of the Social Security Number." *Soc. Sec. Bull.* 69:55.

Pyle, Dorian. 1999. *Data preparation for data mining.* Vol. 1 Morgan Kaufmann.

Rechtspraak. 2015. "Rechtspraak.nl - Zoeken in uitspraken.".
URL: *http://uitspraken.rechtspraak.nl/*

Reurings, M. C. B. and N. M. Bos. 2009. *Ernstig gewonde verkeersslachtoffers in Nederland in 1993-2008 : het werkelijke aantal in ziekenhuizen opgenomen verkeersslachtoffers met een MAIS van ten minste 2.*

Rice, John. 2006. *Mathematical statistics and data analysis.* Cengage Learning.

Rosman, Diana L. 2001. "The Western Australian Road Injury Database (1987âĂŞ1996):: ten years of linked police, hospital and death records of road crashes and injuries." *Accident Analysis & Prevention* 33(1):81 – 88.

SafetyNet. 2008. "Building the European Road Safety Observatory.".

Schrijver, Alexander. 2003. *Combinatorial optimization: polyhedra and efficiency.* Vol. 24 Springer Science & Business Media.

Schürle, Josef. 2005. "A method for consideration of conditional dependencies in the Fellegi and Sunter model of record linkage." *Statistical Papers* 46(3):433–449.

Scrapy. 2015. "Scrapy - A Fast and Powerful Scraping and Web Crawling Framework.".
URL: *http://scrapy.org/*

Smith, M.E. and H. B. Newcombe. 1979. "Accuracies of computer versus manual linkages of routine health records." *Methods of information in medicine* 2:89–97.

Southwood, Thomas Richard Edmund and Peter A Henderson. 2009. *Ecological methods.* John Wiley & Sons.

SWOV. 2015*a*. "Bestand geRegistreerde Ongevallen in Nederland, BRON.".
URL: *http://www.swov.nl/nl/Research/cijfers/Toelichting-gegevensbronnen/BRON.html*

SWOV. 2015*b*. "Definitions of terms that are used in road safety analysis.".
URL: *https://www.swov.nl/UK/Research/cijfers/definitielijst-UK.html*

SWOV. 2015*c*. "Landelijke Medische Registratie (LMR).".
URL: *http://www.swov.nl/NL/Research/cijfers/Toelichting-gegevensbronnen/LMR.html*

Thibaudeau, Yves. 1992. "Identifying discriminatory models in record-linkage." pp. 835–840.

Thibaudeau, Yves. 1993. "The discrimination power of dependency structures in record linkage.".

US National archieves. 2007. "The Soundex Indexing System.".
URL: *http://www.archives.gov/research/census/soundex.html*

Winkler, William E. 1988. "Using the EM algorithm for weight computation in the Fellegi-Sunter model of record linkage." 667:671.

Winkler, William E. 1999. "The state of record linkage and current research problems.".

Winkler, William E. 2000. "Frequency-based matching in the Fellegi-Sunter model of record linkage." *Bureau of the Census Statistical Research Division* 14.

Winkler, William E, United States. Bureau of the Census et al. 1993. *Improved decision rules in the fellegi-sunter model of record linkage*.

Winkler, William E, William E Yancey and EH Porter. 2010. "Fast record linkage of very large files in support of decennial and administrative records projects.".

World Health Organization. 2004. *International statistical classification of diseases and related health problems*. Vol. 1 World Health Organization.

Wu, CF Jeff. 1983. "On the convergence properties of the EM algorithm." *The Annals of statistics* pp. 95–103.

Yancey, William E. 2002. "Improving EM algorithm estimates for record linkage parameters." *Proceedings of the Section on Survey Research Methods, American Statisitcal Association* .