Available online at www.sciencedirect.com

ScienceDirect

Transportation Research Procedia 3 (2014) 224 – 233

Transportation Research Procedia

www.elsevier.com/locate/procedia

17th Meeting of the EURO Working Group on Transportation, EWGT2014, 2-4 July 2014, Sevilla, Spain

A Conceptual Algorithm to Link Police and Hospital Records Based on Occurrence of Values

Marco Amorim[a,]*, Sara Ferreira[a], António Couto[a]

[a] *Research Centre for Territory, Transports and Environment, Faculty of Engineering of University of Porto, Rua Dr Roberto Frias, s/n, 4200-465 Porto, Portugal*

Abstract

Road safety research, in particular road and traffic safety evaluation research, is highly applied and carried out mostly to help reducing the number of road accidents and the injuries resulting from them. This subject has been continuously studied, and in developed countries road safety is improved in a way that, more and more, new measures have less visible impact. Although measures are usually taken directly in the source, which makes all the sense, it is possible to reduce the accident impact if improvements are made "a posteriori"; this is, improving the emergency system to minimize the socioeconomic impact of the accident.

In order to study accidents impact it is necessary to merge two different datasets – police and hospital. This process is known as data linkage and besides a manual linkage process there are three main numerical methodologies: deterministic record linkage, fuzzy matching and probabilistic record linkage. Because these types of datasets are usually protected by anonymity, unique identifiers are not possible to achieve, thus the probabilistic record linkage is usually the chosen method.

This paper presents a concept for an algorithm based on the databases' demographics. By analyzing the various demographic fields it is possible for the algorithm to calculate individual weights that depend on the occurrence of each fields' values among a specific dataset. The demographics are based on the case of Gaia's city road record accidents.

Selection and peer-review under responsibility of the Scientific Committee of EWGT2014

Keywords: linkage algorithm; numerical methods; traffic injury; police data; helthcare data.

* Corresponding author. Tel.: +351-22-508-2161.
E-mail address: mraul@fe.up.pt

Selection and peer-review under responsibility of the Scientific Committee of EWGT2014
doi:10.1016/j.trpro.2014.10.108

1. Introduction

1.1. Why linking databases?

No single database provides enough information to give a complete picture of road traffic injuries and to fully understand the underlying injury mechanisms (Lujic et al., 2008, Clark, 2004). Most common practice for motorized road accidents lays hope in police report, still there is a high portion of underreporting, mainly concerning elderly causalities, urban crashes, slightly injured, users of two-wheeled vehicles and car occupants (Amoros et al., 2008, Amoros et al., 2006, Alsop and Langley, 2001, Pérez et al., 2006).

The other side of the coin shows that medical reports usually lack information on the accident circumstances, but have proved to be very useful to complement police data by capturing missing cases, and also provide detailed information on injury diagnosis (Kilss et al., 1986, Winkler, 1999, Cirera et al., 2001, Jaro, 1995, Newcombe, 1988).

Data linkage with the intend of connect information from road traffic accident, police and medical reports, is not a common area of study although some publishing and studies on this behalf have been reported worldwide – Australia (Ferrante et al., 1993, Rosman, 2001, Boufous and Williamson, 2006), England (Bull and Roberts, 1973, Cryer et al., 2001), New Zealand (Alsop and Langley, 2001) and in United States (Sandra W. Johnson and Walker, 1996, Singleton et al., 2004). The use of these sources allows: international comparisons; estimation of the actual costs to society posed by accidents; and the planning of health care resources (International Traffic Safety Data and Analysis Group, 2011).

1.2. Placement of this paper

Since the 50's that the idea of linking records from different databases has been studied by many researchers either in theoretical aspects of it (Du Bois, 1969, Nathan, 1967, Fellegi and Sunter, 1969) or already envisioning the new technologies by defining computer-oriented record linkages methods (Newcombe and Kennedy, 1962, Newcombe et al., 1959, Newcombe and Rhynas, 1962, Phillips Jr and Bahn, 1963, Tepping, 1955, Tepping, 1968, Tepping and Chu, 1958).

Recently data association through classical probabilistic record linkage has been improved using expectation-maximization algorithms for better parameter estimation in record pair classification (Winkler and Census, 1993, Winkler, 2000). Also by using approximate string comparison to calculate partial agreement weights when attribute values have typographic variations (Winkler, 2006, Christen, 2006).

Moreover, it is reported that data linkage between two or more sources is much more dependent on data quality rather than the record linkage methodology used (Clark, 2004). Usually the linkage process is confronted with the lack of a common unique entity identifier (in this case by the confidentiality of the individuals' identity), and thus becomes non-trivial (Christen et al., 2004). Although Churches and Christen (2004) already proposed blind data linkage technics (BDL), it would require access to the full databases in order to implement such technic. Because requesting confidential data is always a hard and long process, sometimes even impossible, a BDL process is not as interesting as technics that could allow a linkage methodology based on variables with lower variety of values but of easier access to the researcher – e.g. while age comprises values between 0 and 125 (as per Weon and Je (2009) theoretical estimation of maximum human lifespan) and it usually does not risks anonymity, address values can vary up to as many as records existing in the database however likely to risk anonymity (Hundepool and Willenborg, 1996, Sweeney, 1997).

Furthermore the usual focus of the literature is on the specifications for typographical variations matching. This creates a hole in the investigation of records linkage using solely numerical variables. Not many studies explore linkage methods or technics that focus only in non-personal variables such as age, records date and time, gender, and any other variables that alone or as a whole are not enough to identify their "owner" but together might be enough to be identified within a database with a certain probability.

This study is developed under the research project titled LIVE "Tools to Injury Prevention", granted by the European Commission (EC). The major objective of LIVE is to standardize the severity classification among the European countries, in order to improve the actual comparison between them and to increase the richness of data to assist in the prioritization of injury prevention activity. For this a linkage methodology was created to connect police and hospitals databases with intention of replication in other countries and with use of only non-personal variables. Also it is required that the linkage process can take place without supervision, thus disposing of any clerical review. This methodology proves to be efficient after the use of emergency ambulance data to verify the linkage. However there are some situations where the linked data is ambiguous or that would require a better parameter to evaluate the linkage accuracy.

This paper is a conceptual idea of a new solution to weight compared records in a linkage process using the distribution of the characteristics of the population. The solution will be idealized in an algorithm form, opening doors for future research to convert it into a software and optimizing the computational algorithm. Furthermore to help understanding the concept here presented we use the LIVE project study case with focus on Gaia city data.

2. Linkage concepts

The basics of a linkage process consists in a two steps methodology: firstly a search operation to identify potential linkable records; secondly a detailed comparison between the grouped records resulted from earlier (Newcombe and Kennedy, 1962). Thus, the first step goal is to reduce the number of comparisons between unlikely matching records – e.g. matching an accident victim record with a hospital entry record that occurred several weeks after – the second step consist on the use of a classification system that will allow ranking each compared records accordingly to a method to assess similarity.

Usually the process of linking records starts with the data collection followed by the standardization of each field to be compared; afterwards an indexing (blocking) definition is made to search and block records together; records in the same block are then compared against each other using field comparison functions; finally the record pairs matching status are analysed and measured with a weight vector classification in order to assess the set of matched records (M), unmatched records (U) and possible matched records (P) (Alvey and Jamerson, 1997, Elfeky et al., 2002, Christen et al., 2004).

Generally the major challenges in linkage are its computational complexity and linkage accuracy. The last one is the core of this paper, as per the former one the problem relies on the number of comparisons needed to go through all records. A linkage process between database X and database Y will require N_x*N_y comparisons, where N_i is the number of records in database i. For problems where for each record of database X correspond one and only one record of database Y, and vice-versa, the complexity of the process goes as N^2. Thus the number of record pairs' comparison grows quadratically with the number of records to be matched.

To reduce the complexity of the linkage process, indexing (blocking) is used in order to decrease the number of comparisons. In this process the database is subdivided into a set of mutually exclusive subsets (blocks), hence assuming that no match can occur across different blocks (Jaro, 1989). Choosing the blocking variables is a cost-benefit trade-off. If the blocks will contain a high number of records – e.g. blocking by gender - there will be an inefficient large number of comparisons to be made, while blocks that have a low number of records – e.g. blocking by ID – might reduce linkage accuracy (Baxter et al., 2003). The second attention to take when choosing the blocking variable is the error-prone of each variable/field. While in a updated database informatics system the date and time would be automatic fields thus with low probability of error, age and gender would require human intervention to be add in the database, thus raising the probability of misreporting.

Several indexing systems have been developed. The standard and more widely spread method was presented by Jaro (1989) and consist in clustering records into blocks with the same identical blocking key. This method also allows combination of blocking keys – e.g. First 2 letters of the surname and the postal code. However this method, and depending on the available variables, might still require a high amount of comparisons. Hernández and Stolfo (1998) present a Sorted Neighbourhood method that overcomes this problem. The method consists of sorting the records based on a sorting key and then moves a fixed size window through the sorted list comparing only the records inside the window. When one wants to index through a string, Christen et al. (2004) present a Bigram Indexing method. Furthermore McCallum et al. (2000) present a more complex method called Canopy Clustering with TFIDF which focus in choosing a random record and clustering all records within a certain loose threshold distance.

The linkage accuracy depends on the technique used to compare records and the matching evaluation process. The most common models are based on probabilistic record linkage which weight the comparison fields through a binary or categorical score system. Probabilistic record linkage models (PRLM) differ from each other depending on how they calculate the score thresholds used to assess the matched (M), unmatched (U) and probable matches (P) set of records. These are usually Error-based PRLM which calculates the thresholds by minimizing the probability of making an incorrect decision for a record pair (Fellegi and Sunter, 1969). Also the existence of EM-Based PRLM, that focus on incomplete data where each iteration consists of an expectation step followed by a maximization step (Dempster et al., 1977), and a Cost-Based PRLM, that minimize the cost of making a decision rather than the probability of error when making a decision (Verykios et al., 2003), must be pointed out.

However, PRLMs have the disadvantage of handling only binary or categorical comparison vectors. In a case where compared records are restricted to a few set of comparable fields, such is the case of what this paper tries to explore, Cook et al. (2001) feasibility test indicates low wiggle room to use categorical comparison vectors and even more to define thresholds for M, U and P sets of records. The problem escalates when there is no way to implement a clerical review.

It is then important to define a method that allows overcome the issues reported above. The method presented in this paper focus on the linkage of police and hospital car accident victims, thus some specifications and assumptions might not be applied and replicable in other linkage situations.

3. Linking Police and Hospital road accident records

3.1. General specifications

Although linkage methods are able to suit a wide range of linking records problems, most of them focus their power in strings comparisons such as addresses and names. Here, robust algorithm and lists are compiled to compare strings and approximated strings that might have orthographic errors, misspelling, names variations, acronyms and others.

However, as presented in the previous chapter, when studying accident data and due mainly to privacy policies, obtaining victims' individual information other than age and gender proves to be hard. Therefore we focus our linkage algorithm in the most common type of information that is usually accessible when studying road accidents. These are victims' records with gender and age information, the time and date of the record, and the type of victims (e.g. driver, passenger, pedestrian). Moreover in some countries, which is the case of Portugal, hospitals have an area of influence, usually by parish, thus the location of the accident (parish) might be valuable linking information when more than one hospital exists in the study region. In some cases, using the emergency ambulance records can help to connect victims to the destination hospital. All other information in police and hospital records are usually unrelated thus providing no extra linkage variable.

The most important specificity on linking police and hospital road accident records is that the police record timestamp must date before the hospital one, this is, the event that produces the two records works as an "assembly line": Road accident – Police and emergency ambulance service arrival – victim's transportation – arrival at the hospital – treatment - discharge. This brings an important feature of the algorithm that is: the source database is the police database and the reference database is the hospital database thus, the source record has always to precede the reference record.

With the analysis of the available variables we conclude, according to chapter 2, that the *Date* is the optimal cost-benefit blocking variable. *Gender* would simply divide the data set in two groups thus requiring high computational resources; *Time* is not suitable as per last paragraph, matching record will never have the same timestamp – assembly line specificity; *Age* would be a reasonable choice it is more error-prone when compared to *Date*. Nevertheless it is necessary to be careful with "assembly line" property. The time frame between the accident and the arrival at the hospital can be enough to lag police and hospital records by one day. This lag can even be longer than a day if a victim reports to the hospital only few days after the accident as some symptoms might only manifest by then. However these are slight injury victims with minor injuries with small reflection on the type of studies that would require linking police and hospital data.

3.2. Special specification: Population demographics - Gaia city case

In a first look over a linkage system we would conclude that the population to be linked has a random distribution of its characteristics. This would be true if the databases were not dependent of a specific occurrence. For a moment, if we take in consideration two different databases, nursing home and a gymnasium, it stands to reason that in the former, we will have a population much older than the last one. Obviously if we have now a pharmacy that supplies individuals from the both previous services with a database with two records, one of an individual of young age and other of much older age the odds are that the first one most likely belongs to the gymnasium and the last one to the nursing center. This proves that although in a probabilistic record linkage method variables have different weights, even inside each variable values should also have different weights depending on the database population.

To better understand the former statements, and test the algorithm assumptions and data specifications we used part of the database of the project LIVE. The focus is in the city of Gaia with a total number of 6195 victims' records on the police database and 10963 victims' records on the hospital database. One of the reasons for the difference between both databases is that Hospital of Gaia receives victims from accident occurred in the city of Espinho.

Let us focus on the police database to simplify our analysis. Looking at the variable *Date* we found that road accidents have an average of 3 injuries per day with a standard deviation of 2. This means that usually in a day police records from 1 to 5 victims of road accidents. Clearly there is not much variation in this spectrum. Looking further in this variable we can assess its variation during the year and through the week days. The statistical analysis shows that the number of victims varies from around 800, on Sundays, to up to 1000 on Fridays, fig 1(a). The monthly variation goes from around 450 victims, on August, to up to 600 victims during November. Similarly December shows the same figures as November, fig 1(b).

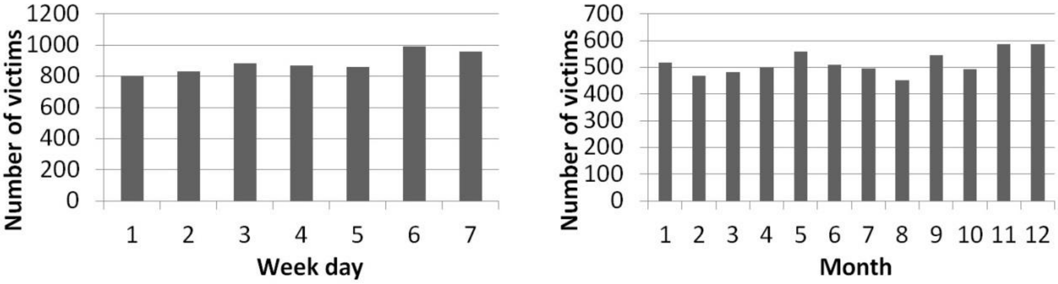


Fig 1: (a) demographics of police record victims by week day, 1 being Sunday; (b) demographics of police record victims by month.

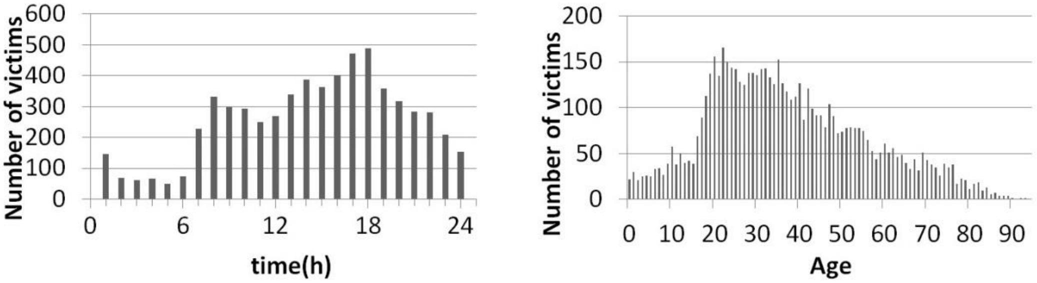


Fig 2: (a) demographics of police record victims by time of the day (hour); (b) demographics of police record victims by Age.

Although *Date* variable shows an equilibrate occurrence of values, the same does not happen with *Time* and *Age*. Fig 2(a) shoes a clear concentration of the number of victims per hour during the working hour period (above 200 victims and up to 500 victims), while during dawn the figures are between 50 and 75 victims per hour. Moreover the focus of victims occurs during the afternoon peak hour. Evidently, when studying the victims population we can conclude that time statistic has a nonlinear variation, being the 18th hour of the day the one with more occurrences (489 victims) and the 5th hour of the day the one with the less occurrences (51 victims).

The population age statistic presents a wide range of values, some much more common than others, fig 2(b). The victims' *Age* variable values concentrate between the 20s and 30s years range, having this range a total of around 45% of all occurrences.

Evidently age and time have a strong correlation with the event of a road accident and its outcomes. Although the day of the accident do not show a strong relation, winter and weekends (mainly Friday and Saturday) apparently have an higher number of victims, however not as evident as time and age. These conclusions provide the base of the algorithm presented in the 5th chapter. Additionally we point out the advantages of its application as a mean to implement a more detailed score system for linked records pair.

4. Occurrence of values as a way to weight linked records

In chapter 3 we pointed that some of the variables (fields) have inconstant demographics where some values appear much more times than others. Age demographic indicates a concentration of the victims in between 20s and 30s year's old range, similar to what other studies indicated (Lourens et al., 1999, Lam, 2002). For time, the database indicates a higher number of accidents during the peak hours and the periods in between. Many other variables have similar behavior.

Let us assume that we have a source database *X* with *i* variables, and *j* records. Being *I* the set of variables *i* and *J* de set of records *j*, *X* is:

$$X = \begin{pmatrix} x_1^1 & & \cdots & & x_I^1 \\ & \ddots & & \iddots & \\ \vdots & & x_i^j & & \vdots \\ & \iddots & & \ddots & \\ x_1^J & & \cdots & & x_I^J \end{pmatrix} \tag{1}$$

The probability of drawing a record with specific record value of a variable *i* from *X* is:

$$\text{Prob}(x_i = x_i^j \big| X) = \frac{\text{Nº of times the variable i assumes the record j value}}{\text{Total of records of variable i (I)}} \tag{2}$$

Thus the probability of drawing a record with a specific combination of values of a record of the set *I* of *X* is:

$$\text{Prob}(x = x^j \big| X) = \prod_{i=1}^{I} \text{Prob}(x_i = x_i^j \big| X) \tag{3}$$

Let us add a reference database *Y* with *k* records of the set *K,* and the same variables as per equation 1. The probability of drawing a record *j* from *X* that is equal to the record *k* is:

$$\text{Prob}(x = \text{y}^k \mid X) \tag{4}$$

Assuming now that a linkage exists between a record *j* from *X* and a record *k* from *Y*, if a certain gap exists within a variable between both records, such that:

$$gap = \left| x_i^j - y_i^k \right| \tag{5}$$

We can say that in terms of variable *i*, record *j* is within the neighborhood of the record *k* with radius:

$$R^{(j-k)} = y_i^k \pm \left| x_i^j - y_i^k \right| \tag{6}$$

We can then assume that *j* is similar to *k* within $R^{(j-k)}$. Therefore, the probability of drawing a record of *X* and of it being similar to *k* as *j* is, comes as:

$$Weakness^{(j-k)} = \prod_{i=1}^{I} \text{Prob}\left(y_i^k - \left| x_i^j - y_i^k \right| \le x_i \le y_i^k + \left| x_i^j - y_i^k \right| \right) \tag{7}$$

This is what we can call as the weakness of the link $x^j \rightarrow y^k$. Then, the weight of the link of record *j* with record *k* is:

$$Weight^{(j-k)} = 1 - \prod_{i=1}^{I} \text{Prob}\left(y_i^k - \left| x_i^j - y_i^k \right| \le x_i \le y_i^k + \left| x_i^j - y_i^k \right| \right) \tag{8}$$

Fig 3 shows an example of application of this mathematical model.

5. Conceptual algorithm structure

5.1. Generalities

In chapter 3 we presented several specifications for the linkage of police and hospital accident records. The general idea of this paper algorithm is to take advantage of these specifications and develop a more detailed method to identify and classify possible pairs in the two databases.

Earlier in chapter 2, indications on the usual design of a linkage method were made. For this algorithm similar structure is used.

Each record is considered to be a vector with several fields that are called variables after standardization. The suffix PLC indicates that the origin of the record is the police database and the EMR suffix indicates that the origin is the hospital database. The representation is as following: SOURCE_i(*Age, Date, Time, Gender, Type*).

5.2. Standardization

The data standardization is the most important process as all other steps are depending on it. The fields to be used have to have either a numerical value or the possibility to be converted as such.

Some fields' standardization will be presented now, however many others can be used if similar structure is used.

Concept for the standardization part of the algorithm:

SOURCE_i(age, date, time, gender, type) → SOURCE_i(*Age, Date, Time, Gender, Type*)

Age – age converted to an integer, 1 unit = 1 year of life.

Date – date converted to an integer, 100 units = 1 day. $Date_i$ = count_number_of_days (date_i - date_0)

time – time converted to an integer, 100 units = 24H. *time* = (hours + minutes/60) * 100/24;

Time – time dependent on date, 100 units = 24H. *Time* = *time* + *Date*;

Gender – gender converted to an integer, 1 = male, 2 = "no information", and 3 = female;

Type – type converted to an integer, 1 = driver, 2 = passenger, 3 = "no information", and 5 = pedestrian.

Where:

Italic distinguish a field from a *variable*;

count_number_of_days – functions that counts the number of days between two dates;

hours – hour of the field time

minutes – minutes of the field time

i – record ID, here 0 is the earliest record

The algorithm accounts for fields where no information is available. In the conversion to variables the important aspect is that the distance between the numerical value for the missing field and the other variable's values has be constant.

5.3. Indexing (blocking)

The indexing process reduces the computational effort, as demonstrated in chapter 2. The most suitable solution here is to use a mix of the clustering and sorted neighborhood systems proposed by Jaro (1989) and Hernández and Stolfo (1998) respectively. As explained in chapter 3 the preferable variable for indexing is Date. Moreover, when more than one hospital may exist, the indexing can be complemented by comparing records of accidents occurred only on the hospital's influence area.

Concept for the indexing part of the algorithm:

1st. Define blocking variable range (maximum difference allowed between source and reference records date to be linked)

2nd. Sort records by date

3rd. Clustering together PLC records of day X with EMR records of day X to X + range

4th. For all possible combinations of records from PLC with records from EMR within cluster in analysis:

   a. execute field comparison (fig 3);

   b. Weight classification (fig 3);

5th. Repeat step 4 for all PLC days.

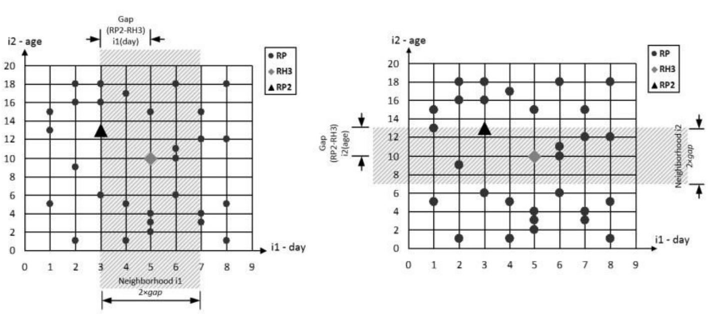
Example of the weighting process

Table with the records to compare, based on RH3 (5;10):

Variables	Police	Hospital	GAP (eq 5)	R (Neighborhood, eq 6)
Records	RP2	RH3	RP2-RP3	RP2-RP3
Day (i1)	3	5	2	3 - 7
Age (i2)	13	10	3	7 - 13

Table with all the possible police records that will shape the demographics of the police population:

Record (j)	day(i1)	age(i2)	Record (j)	day(i1)	age(i2)
RP1	6	18	RP17	6	10
RP2	3	13	RP18	1	15
RP3	7	3	RP19	1	15
RP4	2	18	RP20	3	16
RP5	8	5	RP21	2	16
RP6	5	2	RP22	4	1
RP7	5	15	RP23	8	1
RP8	6	11	RP24	7	12
RP9	2	9	RP25	1	5
RP10	6	6	RP26	2	1
RP11	5	4	RP27	3	18
RP12	4	17	RP28	7	15
RP13	8	18	RP29	8	12
RP14	5	3	RP30	4	5
RP15	7	4	RP31	1	13
RP16	3	6			


Graphic 1:

Graphic 2:

Table with the final results of the comparison between police record RP2 and hospital record RH3:

	Day (graphic 1)	Age (graphic 2)	Weakness (eq 7)	Weight (eq 8)
Records	29	29	0.15	0.85
Neighborhood Records	18	7		
Probability	0.62	0.24		

Fig 3: Example of an application of the mathematical model to measure the weight between the link of a record from the police and the a record from a hospital. The variables were converted to integers.

5.4. Field comparison function and weight classification

The field comparison results from equation 6 of chapter 4. Here, contrary to what is usually made, we do not define a categorical score system as presented in chapter 2. The goal is to let the algorithm measures the existent gap between two records and assess how many more records exist in the block with an equal or lower gap to the reference record.

The algorithm is implemented in step 4th the indexing algorithm. The two steps, a. and b., are executed together using the equation 8 of chapter 4. Moreover it is important to add that for variables such as *time*, where a real gap is present, equation 5 is of no use. Here it is preferable to add a probability function time, $f(time_{k-j})$, transforming equation 7 and the final weight is:

$$\begin{aligned} Weight^{(j-k)} = 1 - \prod_{i=1}^{g} \text{Prob}\left(y_i^k - \left| x_i^j - y_i^k \right| \le x_i \le y_i^k + \left| x_i^j - y_i^k \right| \right) \times f(time_{k-j}) \\ \times \prod_{i=g+2}^{I} \text{Prob}\left(y_i^k - \left| x_i^j - y_i^k \right| \le x_i \le y_i^k + \left| x_i^j - y_i^k \right| \right) \end{aligned} \tag{9}$$

Where *g* is the variable that precedes the time variable.

Finally, after finishing the indexing and weighting, the algorithm sorts the list of links by weights and eliminates the repeated PLC records (the ones with the lower weights). The final links are considered to be true matches if we assume that all victims reported from police were transported to the respective hospital, and in the hospital database only exists records from road accidents.

6. Future developments

As future lines of research the probability function time should be studied and assessed through real cases. The best way would be to utilize ambulance records with the full history of the emergency ambulance system – emergency call; arrival at the accident scene; accommodation of the victims in the ambulance; transportation to the hospital.

Moreover a fully detailed description of the algorithm and its implementation in computer software should be made in order to automate the process. This would also allow that comparisons can be made with already linked databases in order to analyze thresholds for the weight system. These thresholds are important to understand the potentiality of the weight measured in cases where both databases only have partial connection – e.g. linking a database of national records of road accidents with the hospital emergency records with no information of the victims' origins.

Acknowledgements

The authors would like to thank to the European Commission which funded this study through the project LIVE, and also to the institutions which granted the access to the information required by the study, namely: Autoridade Nacional de Segurança Rodoviária e Centro Hospitalar de Gaia. This project has also been partially supported by FCT (Fundação para a Ciência e a Tecnologia), the Portuguese Agency for R&D, under SFRH/BD/52355/2013.

References

Alsop, J. & Langley, J. 2001. Under-reporting of motor vehicle traffic crash victims in New Zealand. *Accident Analysis & Prevention,* 33**,** 353-359.

Alvey, W. & Jamerson, B. Record linkage techniques–1997. Proceedings of an International Workshop and Exposition. March, 1997. 20-21.

Amoros, E., Martin, J.-L., Lafont, S. & Laumon, B. 2008. Actual incidences of road casualties, and their injury severity, modelled from police and hospital data, France. *The European Journal of Public Health,* 18**,** 360-365.

Amoros, E., Martin, J.-L. & Laumon, B. 2006. Under-reporting of road crash casualties in France. *Accident Analysis & Prevention,* 38**,** 627-635.

Baxter, R., Christen, P. & Churches, T. A comparison of fast blocking methods for record linkage. 2003 2003. Citeseer, 25-27.

Boufous, S. & Williamson, A. 2006. Work-related traffic crashes: A record linkage study. *Accident Analysis and Prevention,* 38**,** 14-21.

Bull, J. P. & Roberts, B. J. 1973. Road accident statistics—A comparison of police and hospital information. *Accident Analysis & Prevention,* 5**,** 45-53.

Christen, P. A Comparison of Personal Name Matching: Techniques and Practical Issues. Data Mining Workshops, 2006. ICDM Workshops 2006. Sixth IEEE International Conference on, Dec. 2006 2006. 290-294.

Christen, P., Churches, T. & Hegland, M. 2004. Febrl – A Parallel Open Source Data Linkage System. *In:* Dai, H., Srikant, R. & Zhang, C. (eds.) *Advances in Knowledge Discovery and Data Mining.* Springer Berlin Heidelberg.

Churches, T. & Christen, P. 2004. Blind Data Linkage Using n-gram Similarity Comparisons. *In:* Dai, H., Srikant, R. & Zhang, C. (eds.) *Advances in Knowledge Discovery and Data Mining.* Springer Berlin Heidelberg.

Cirera, E. V. A., Plasència, A., Ferrando, J. & Arribas, P. 2001. Probabilistic Linkage of Police and Emergency Department Sources of Information on Motor-Vehicle Injury Cases: a Proposal for Improvement. *Journal of Crash Prevention and Injury Control,* 2**,** 229-237.

Clark, D. E. 2004. Practical introduction to record linkage for injury research. *Injury Prevention,* 10**,** 186-191.

Cook, L. J., Olson, L. M. & Dean, J. M. 2001. Probabilistic Record Linkage: Relationships between File Sizes, Identifiers, and Match Weights. *Methods of Information in Medicine,* 40**,** 196-203.

Cryer, P. C., Westrup, S., Cook, A. C., Ashwell, V., Bridger, P. & Clarke, C. 2001. Investigation of bias after data linkage of hospital admissions data to police road traffic crash reports. *Injury Prevention,* 7**,** 234-241.

Dempster, A., Laird, N. & Rdin, D. Maximum Likelihood from Incomplete Data via the EM Algorithm. JOURNAL OF THE ROYAL STATISTICAL SOCIETY, SERIES B, 1977. 1-38.

Du Bois, N. S. D. A. 1969. A Solution to the Problem of Linking Multivariate Documents. *Journal of the American Statistical Association,* 64**,** 163-174.

Elfeky, M. G., Verykios, V. S. & Elmagarmid, A. K. TAILOR: a record linkage toolbox. Data Engineering, 2002. Proceedings. 18th International Conference on, 2002 2002. 17-28.

Fellegi, I. P. & Sunter, A. B. 1969. A Theory for Record Linkage. *Journal of the American Statistical Association,* 64**,** 1183-1210.

Ferrante, A. M., Rosman, D. L. & Knuiman, M. W. 1993. The construction of a road injury database. *Accident Analysis & Prevention,* 25**,** 659-665.

Hernández, M. & Stolfo, S. 1998. Real-world Data is Dirty: Data Cleansing and The Merge/Purge Problem. *Data Mining and Knowledge Discovery,* 2**,** 9-37.

Hundepool, A. & Willenborg, L. μ-and τ-argus: Software for statistical disclosure control. Third International Seminar on Statistical Confidentiality, 1996.

Jaro, M. A. 1989. Advances in Record-Linkage Methodology as Applied to Matching the 1985 Census of Tampa, Florida. *Journal of the American Statistical Association,* 84**,** 414-420.

Jaro, M. A. 1995. Probabilistic linkage of large public health data files. *Statistics in Medicine,* 14**,** 491-498.

Kilss, B., Alvey, W., Methodology, U. S. F. C. O. S., Division, U. S. I. R. S. S. O. I. & Society, W. S. 1986. *Record linkage techniques, 1985: proceedings of the Workshop on Exact Matching Methodologies, Arlington, Virginia, May 9-10, 1985 : co-sponsored with the Washington Statistical Society and the Federal Committee on Statistical Methodology*, Dept. of the Treasury, Internal Revenue Service, Statistics of Income Division.

Lam, L. T. 2002. Distractions and the risk of car crash injury: The effect of drivers' age. *Journal of Safety Research,* 33**,** 411-419.

Lourens, P. F., Vissers, J. a. M. M. & Jessurun, M. 1999. Annual mileage, driving violations, and accident involvement in relation to drivers' sex, age, and level of education. *Accident Analysis & Prevention,* 31**,** 593-597.

Lujic, S., Finch, C., Boufous, S., Hayen, A. & Dunsmuir, W. 2008. How comparable are road traffic crash cases in hospital admissions data and police records? An examination of data linkage rates. *Australian and New Zealand Journal of Public Health,* 32**,** 28-33.

Mccallum, A., Nigam, K. & Ungar, L. H. 2000. Efficient clustering of high-dimensional data sets with application to reference matching. *Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining.* Boston, Massachusetts, USA: ACM.

Nathan, G. 1967. Outcome Probabilities for a Record Matching Process with Complete Invariant Information. *Journal of the American Statistical Association,* 62**,** 454-469.

Newcombe, H., Kennedy, J., Axford, S. & James, A. 1959. *Automatic linkage of vital records*.

Newcombe, H. & Rhynas, P. 1962. Family linkage of population records. *The Use of Vital and Health Statistics for Genetic and Radiation Studies***,** 135-154.

Newcombe, H. B. 1988. *Handbook of record linkage: methods for health and statistical studies, administration, and business*, Oxford University Press, Inc.

Newcombe, H. B. & Kennedy, J. M. 1962. Record linkage: making maximum use of the discriminating power of identifying information. *Commun. ACM,* 5**,** 563-566.

Pérez, C., Cirera, E., Borrell, C. & Plasència, A. 2006. Motor vehicle crash fatalities at 30 days in Spain. *Gaceta Sanitaria,* 20**,** 108-115.

Phillips Jr, W. & Bahn, A. K. 1963. Experience with computer matching of names. *American Statistical Association Proceedings, Social Statistics Sec***,** 26-38.

Rosman, D. L. 2001. The Western Australian Road Injury Database (1987–1996):: ten years of linked police, hospital and death records of road crashes and injuries. *Accident Analysis & Prevention,* 33**,** 81-88.

Sandra W. Johnson & Walker, J. 1996. The Crash Outcome Data Evaluation System (CODE). Washington DC: National Highway Traffic Safety Administration.

Singleton, M., Qin, H. & Luan, J. 2004. Factors Associated with Higher Levels of Injury Severity in Occupants of Motor Vehicles That Were Severely Damaged in Traffic Crashes in Kentucky, 2000-2001. *Traffic Injury Prevention,* 5**,** 144-150.

Sweeney, L. 1997. Guaranteeing anonymity when sharing medical data, the Datafly System. *Proc AMIA Annu Fall Symp***,** 51-5.

Tepping, B. J. 1955. Study of matching techniques for subscriptions fulfillment. *National Analysts Inc., Philadelphia*.

Tepping, B. J. 1968. A Model for Optimum Linkage of Records. *Journal of the American Statistical Association,* 63**,** 1321-1332.

Tepping, B. J. & Chu, J. T. 1958. A report on matching rules applied to readers digest data. *National Analysts Inc., Philadelphia, August*.

Verykios, V. S., Moustakides, G. V. & Elfeky, M. G. 2003. A Bayesian decision model for cost optimal record matching. *The VLDB Journal,* 12**,** 28-40.

Weon, B. & Je, J. 2009. Theoretical estimation of maximum human lifespan. *Biogerontology,* 10**,** 65-71.

Winkler, W. 1999. The state of record linkage and current research problems. Statistics of Income Division, Internal Revenue Service Publication R99/04.

Winkler, W. E. Machine learning, information retrieval and record linkage. Proc Section on Survey Research Methods, American Statistical Association, 2000. 20-29.

Winkler, W. E. Overview of record linkage and current research directions. Bureau of the Census, 2006. Citeseer.

Winkler, W. E. & Census, U. S. B. O. T. 1993. *Improved decision rules in the fellegi-sunter model of record linkage*, Citeseer.