Work-related road accidents: a data linkage procedure applied to assess traffic accidents at work and commuting

Luca Taiano [1], Stefania Massari [1], Tiziana Tuoto [2], Luca Valentino [2],
Silvia Bruzzone [2], Liana Veronico [1]

Abstract

Record linkage is a data integration technique whose goal is to identify the same unit represented in different data sources in different ways. Deterministic linkage and probabilistic linkage are two linkage techniques that have been already widely used. The goal of this paper is to show how a record linkage procedure based on a probabilistic approach provides an increase in linked pairs compared to the sole use of a deterministic approach and to provide a step procedure to sequentially apply multiple linkage techniques. Datasets are the Inail (Italian National Institute for Insurance against Accidents at Work) archive of work-related accidents occurring with the use of a vehicle and the Istat (Italian National Institute of Statistics) archive of road accidents resulting in death or injury. It was applied a deterministic linkage followed by probabilistic linkage on the unlinked records. Deterministic linkage undoubtedly considers two records to be a link if they agree on a selection of variables whereas probabilistic linkage assigns a probability of being a link to records. Results show that the probabilistic linkage produced an increase of 18% of the linked pairs compared to the sole use of the deterministic approach.

Keywords: Data integration, record linkage, deterministic linkage, probabilistic linkage, road accidents.

[1] Istituto Nazionale per l'Assicurazione contro gli Infortuni sul Lavoro (*Italian National Institute for Insurance against Accidents at Work*) - Inail.

[2] Istituto Nazionale di Statistica (*Italian National Institute of Statistics*) - Istat.

In the writing of this article, the authors were supported by all the members of the Inail-Istat collaborative group: Antonella Altimari, Michela Bonafede, Roberto Boscioni, Claudio Gariazzo, Alessandro Marinaccio, Stefania Massari, Antonella Pireddu, Luca Taiano and Liana Veronico (for Inail); Giordana Baldassarre and Silvia Bruzzone (for Istat).

The views and opinions expressed are those of the authors and do not necessarily reflect the official policy or position of the Italian National Institute of Statistics - Istat.

The authors would like to thank the anonymous reviewers for their comments and suggestions, which enhanced the quality of this article.

1. Introduction

Among methodologies of data integration, the techniques of records linkage are a set of methods whose goal is to identify the same unit represented in different data sources in different ways. Methodologies of data integration such as record linkage create new information assets from the already available ones, adding value to the existing data archives, allowing better insights, new conclusions and reducing the necessities to carry out new surveys.

Two approaches can be used for record linkage: deterministic and probabilistic. The deterministic method establishes whether a pair of records is a link based on a set of given conditions with a determined outcome (match or non-match). Its efficiency, measured as the number of linked records, is limited by incorrectness and incompleteness of the information to be linked. The probabilistic method assigns to each pair a probability of being a link. It uses the approach described by Fellegi and Sunter (Fellegi and Sunter 1969; Scanu 2003) and can be implemented by the software *RELAIS* (REcord Linkage At IStat). The probabilistic approach has been widely used for integrating different data sources to enlarge the analysis and the comprehension of a given phenomenon (Tuoto *et al.*, 2015; Tuoto *et al.*, 2014; Tuoto *et al.*, 2012).

The novelty of this contribution is in the process that drives the linkage activities and the roles played by the data owners. The linked data archives are the Inail (Italian National Institute for Insurance against Accidents at Work) archive of commuting to (*in itinere*) or at work road accidents and the Istat (Italian National Institute of Statistics) archive of road accidents.

Record linkage creates a new dataset where each record ideally should refer to the same unit contained in both the input datasets. The new record has a complete set of information that is, most importantly, related.

The phenomena of road accidents is an example about how complementary information collected by different Authorities can be integrated to get insights. The Italian National Institute of Statistics - Istat registered road accidents that caused persons to die or to be injured. These road accidents have also an occupational origin. Workers use vehicles both for commuting (home-work travelling routes) and for their work (*e.g.* in the transport sector). Although the Istat archive contains pieces of information about the occupational origin, it is often incomplete or unfilled by police officers, and consequently the rate of road occupational accidents cannot be assessed from this dataset. The work-related component of road accidents is instead recorded by Inail (Italian National Institute for Insurance against Accidents at Work) being it an occupational accident. However, not all work-related road accidents are associated with a request for insurance compensation, particularly for those occurring during commuting. Reasons might be due to lack of knowledge about the possibility to claim for compensation, lack of time available to manage with administrative procedures or unregistered accidents. This can underestimate the overall work-related phenomena.

The interconnection of the two road accidents archives can provide advantages, not only in adding occupational information at each accident linked, but also in assessing the efficiency of Istat dataset in registering the occupational component. It can also provide advantages in identifying the unlinked records, potential work-related accidents registered by Istat but not included in the Inail, work-related, archive, with the aim to estimate the unclaimed work-related road accident phenomena. Brusco *et al.* (2019) earlier linked the two archives using a deterministic approach for road accidents occurred in Italy in the year 2015. They found a record linkage efficiency of about 23% of the number of records contained in the Inail archive, addressing the possible unmatched accidents with inaccuracies in the registration systems. To increase the efficiency a coupled deterministic and probabilistic approach could be used.

To test and verify this hypothesis, this work describes the integration of the above road accidents archives by means of a combined deterministic and probabilistic record linkage approach applied to data collected from the year 2014 to the year 2018.

2. Materials and methods

2.1 The datasets

The Inail occupational accidents data archive covers about 80% of the Italian workforce. Inail receives compensation claims applications for occupational injuries from all over the national territory, regarding all workers except for some categories (armed forces, firefighters and police workers, air transport personnel, autonomous tradespeople and professionals with VAT registration). The archive contains information about time and location of the accident, the economic branch of the victim, the occurrence on duty or during commuting and health consequences of the accident, such as body part injured, type of injury and health effects. The Inail dataset is made of the insurance claims of workers for road accidents during work or commuting between 2014 and 2018.

Data about road accidents are routinely collected by Istat to produce statistical reports on this phenomenon on the basis of data recorded from Local Authorities ("*Carabinieri*", Motorway Police, and Local Police). Data refer to road accidents in which an injury or a fatality occurred, involving at least one vehicle, on the public roads of the national territory, occurred between 2014 and 2018. The archive contains road and vehicle-related information of the accident, road and weather conditions, road signs and crossings presence, time, location and geographical coordinates of the accident.

2.2 Pre-processing of the datasets

Unavoidably, data coming from different sources need some pre-processing before they are usable in a linkage model. Data must be recorded and stored in the same way in order for the units to be compared. The procedure required a well-structured technique and different pre-processing steps.

2.2.1 From the accident to the person in the Istat dataset

The records contained in the Inail archive refer to a single individual, while the unit recorded in the Istat dataset is an accident, involving more individuals. The road accidents Istat database is structured according to four different dimensions: Accident, Road, Traffic Unit and Person. The structure used by Istat is the same recommended and implemented by the European Commission, in the CARE database (Community database on road accidents resulting in death or injury) (European Commission 2021). Following the 93/704/EC: Council Decision of 30 November 1993 on the creation of a Community database on road accidents, Italy updates, every year, road accidents data at national level.

Since the Istat dataset represents the accident, each record of the Istat dataset was required to be transformed into *n* records, each one representing a person involved in the accident. This was possible because personal data, for each person involved in the accident, was present in the accident record. Each record contained person identification data (name and surname) and attributes (*e.g.* gender, age) for each person involved in the accident.

To provide a collection of records referred to each person involved in the accident and harmonised with the Inail data set, the authors built a mirror database containing all injured and dead, excluding the unharmed drivers, identified by a seven-digit code. The first six digits identifies the road accident in the main database and the seventh digit corresponds to the ID of the person involved in the accident. In our case, name and surname are guaranteed to refer to the same person, whereas age is not, since the form filled by authorities at the time of the accident records but does not associate ages with persons.

This introduces an error in the form of a lower recall; persons with age wrongly associated can be excluded from the linked pairs, since a difference on a single field can be enough to prevent a link. A reduction of the accuracy is less likely since a single field wrongly filled is less likely to make the difference, alone, in forming a link. There is no bias since there is no kind of records affected more than others. It is useful to point this out since it can be a source of improvement in data collection. Time and location are guaranteed to be correct, being unique for the accident.

2.2.1 Data cleaning and formatting

The second pre-processing step was to ensure that the information contained in corresponding columns in the two archives was represented in the same way. Treatment and formatting were applied where needed. The following operations were carried out:

1. name and surname were converted into uppercase.
2. blank spaces were removed.
3. non-alphabetic characters were removed.
4. letters with signs on or above them were converted to their simple A-Z counterpart.
5. age was made into an integer.
6. a key column (a unique identifier for records in the dataset) was created if not already present.
7. date was put into a unique format (*e.g.* dd/mm/yyyy) and constructed where missing.
8. hour was made into an integer.
9. it was ensured that municipality codes belonged to the same yearly classification, and they were converted to the same yearly classification where that was not the case.
10. a new column was created containing the concatenation of surname and name (in some cases surname and name were both stored in the same column).
11. equal columns were given equal names.

Tables 2.1 and 2.2 show the variables contained in each dataset in their final form.

Table 2.1 - Inail dataset

Personal data			
Case code	Name	Surname	Age
Gender	Country of birth		
Site			
Macro-region	Region	Province	Municipality
Time			
Year	Month	Day (of month)	Date
Hour	Day (of week)		
Injury			
Outcome (injured/dead)		Type of injury	Body part injured
Work			
Economic activity		Commuting (yes/no)	
Insurance			
Insurance management group	Large tariff group (a)	Compensation type	Compensated days
Assumed grade of impairment		Actual grade of impairment	

Source: Authors' processing on Inail dataset
(a) The large tariff grouping that groups the tariff items, which associate the work with the premium rate.

Table 2.2 - Istat dataset

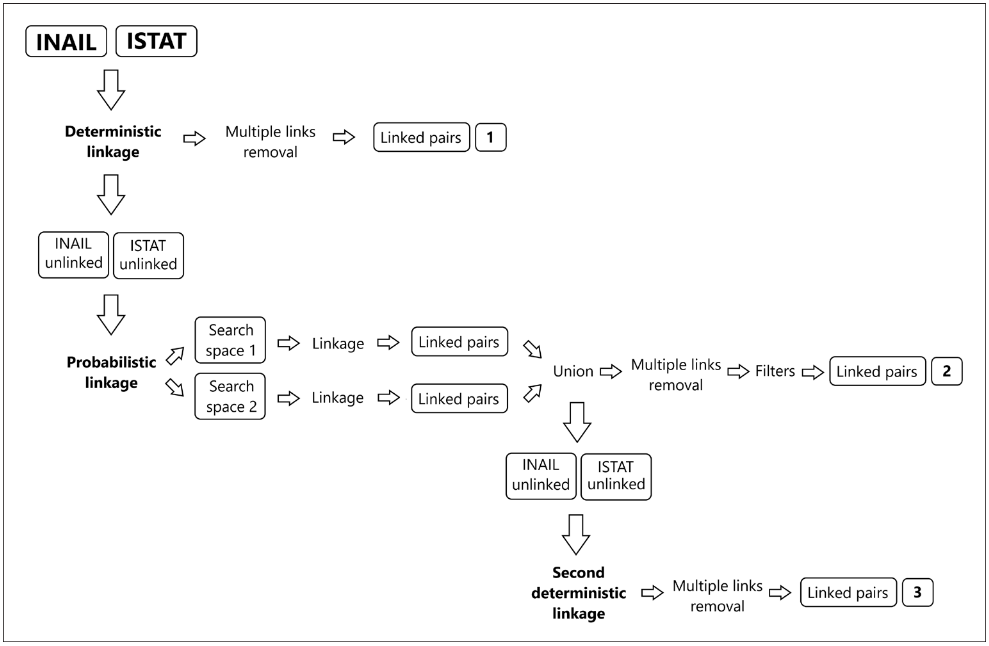
Personal data			
Name	Surname	Age	Gender
Driving license type	Role (driver/passenger/pedestrian)		
Site			
Province	Municipality	Locality	Coordinate type
Projection system	Latitude	Longitude	
Time			
Year	Month	Day (of month)	Hour
Minutes	Day (of week)	Period of day	
Injury			
Outcome (injured/dead)			
Work			
Professional condition (at work/commuting)			
Road			
Road identification code	National road or motorway section	Progressive mileage (Km)	Hectometric
Type of road	Pavement	Road-bed	Weather
Traffic signs	Junction / Non-junction	Localisation of the accident	
Hospital			
Name of the Hospital			
Vehicle			
Type of vehicle involved	Vehicle cylinder capacity	Vehicle license plate	
Accident			
Accident identification number		Road accident type	
Number of people dead in the accident		Number of people injured in the accident	

Source: Authors' processing on Istat dataset

2.3 The linkage procedure

The record linkage was carried out following a step procedure described in the following sub-sections. Figure 2.1 summarises the whole process.

Figure 2.1 - Linkage process



Source: Authors' processing on Inail and Istat datasets

2.3.1 Variables selection

In order to link records, the first step is to identify which variables can be used to perform the match. These variables must necessarily be looked for among those in common between the two datasets. The chosen variables are shown in Table 2.3.

Table 2.3 - Common variables

Personal data			
Name	Surname	Age	Gender
Site			
Province	Municipality		
Time			
Year	Month	Day (of month)	Date
Hour	Day (of week)		
Injury			
Outcome (injured/dead)			

Source: Authors' processing on Inail and Istat datasets

As privacy is concerned with such information, we applied procedures to prevent access to information by unauthorised users. After data linkage, data about individuals were removed for privacy reasons in compliance with the law.

2.3.2 Deterministic linkage

The first linkage operation consisted of a deterministic linkage. Records were matched by surname, name, date and municipality. Records reporting the same surname, name, date and municipality were undoubtedly considered representing the same accident. Clerical or data entry errors can produce a lower recall, rarely a lower accuracy, since it is more likely that two equal surnames are made different by an error than two different surnames are made equal by an error. For surname and name, their concatenation was used, so the equality condition was on their concatenation. A second check was performed in order to include accidents that were mistakenly recorded with name and surname inverted.

2.3.3 Multiple links removal

After this step, a check for multiple links was performed. Multiple links are present when a unit in one dataset is linked with more than one unit in the other. For example, it can happen if in one dataset the same accident is recorded twice, or in case of homonyms. Duplicates must then be removed and, among them, just one linked line must be kept since we want a 1:1 linkage. The key column we introduced in the previous step finds here one of its uses: if duplicates are present in a key column in the linked table, then that record has been linked to more than one record of the other dataset. Table 2.4 exemplifies this event:

Table 2.4 - Example of multiple links

Dataset A key	[other A columns …]	Dataset B key	[other B columns …]
001	[other A data …]	054	[other B data …]
001	[other A data …]	055	[other B data …]

Source: Authors' processing on Inail and Istat datasets

In this case, the element with key ‘001’ of dataset A matches two elements of dataset B. To select one link among the multiple links, equality of other common variables, like age and hour of the accident, were used, to discern which individual of dataset B represents the same accident of the record of dataset A.

2.3.4 Unlinked records archives

This step creates two sets of unlinked records. They are composed of the records remained unmatched, for each input dataset. The sets of unlinked records are the set difference between the input datasets and the linked dataset. The created key column finds here another use: the unlinked records of an input dataset are all of its elements whose key is not present in the linked dataset. A simple count check ensures that the operation has been correctly performed: the number of elements of the input dataset must be equal to the number of linked elements plus the number of the unlinked records. The datasets of unlinked records created in such a way were the input datasets for the next step.

2.3.5 Probabilistic linkage

The third step is to perform the probabilistic linkage. The input datasets are the datasets of unlinked records created in the previous step. The probabilistic linkage was performed according to the theory proposed by Fellegi and Sunter (Fellegi and Sunter 1969; Scanu 2003) and was run by the Istat software *RELAIS* (REcord Linkage At IStat). The Fellegi and Sunter theory is internationally recognised as the reference theory in record linkage (Christen, 2012*a*; Herzog, 2007), in particular its implementation due to Jaro (Jaro 1989). The size of the data processed in this paper required a reduction of the computational space for computational treatability; hence, a comparison space was created, using the Sorted Neighbourhood algorithm (Christen, 2012*b*) with a window size of 50. The Sorted Neighbourhood algorithm lists the elements of the two datasets in a single list, then sorts them according to a sorting variable. Then a fixed size window runs on the sorted list and all the pairs falling into the window are considered candidate pairs (Hernandez, 1995). This procedure was run twice, creating two sets of candidate links, one sorting on surname and name and the other sorting on the concatenation of surname and name. Then the same linkage model was applied to both of them; linked pairs were then unified in one dataset, taking the intersection and the set differences. The linkage model uses as matching variables surname, name, date and age;

for the latter a window of size 1 was considered for comparison, admitting a difference of one year in the ages of the compared records. The linkage model declared as matches the candidate pairs with a posterior linkage probability higher or equal than 0.8, and possible matches, to be reviewed by manual checks, those pairs with a posterior linkage probability in the range [0.5-0.8). At the end of this step, multiple links were removed and datasets of unlinked records calculated, as described in paragraphs 2.3.3 and 2.3.4.

2.3.6 Filtering

The dataset resulting from the probabilistic linkage is a dataset of linked pairs that we can inspect to elaborate filters (rules for clerical review and selection). The aim is to further increase the chances that the identified paired records refer to the same accident. Such a task reduces error tolerance in the matching procedure. For example, if date is different, what is the extent of the difference? If the extent is small and other fields match, can it be taken as a recording error, especially if data were collected manually?

The following filters were thus elaborated and run on the linked pairs. Filters were applied in succession; so filter 2 was applied on the pairs that did not pass filter 1, filter 3 was applied on the pairs that did not pass filter 1 and 2 and so on. Filters are expressed in the form of logical conditions that must all hold true to accept the pair.

Equal date, equal concatenation of surname and name, hour with difference not greater than one, age with a difference not greater than one.

1. On the possible matches only: equal date, equal municipality.
2. Equal date, concatenation of surname and name with a Levenshtein distance (edit difference) not greater than one, hour with a difference not greater than one, age with a difference not greater than one.
3. Equal date, equal surname, one name contained in the other, hour with a difference not greater than one, age with a difference not greater than one.
4. Equal date, equal name, one surname contained in the other, hour with a difference not greater than one, age with a difference not greater than one.

The constraint on date was then loosened: dates with a difference of one in day or month, in presence of an equality on other fields, is interpretable as a filling error (*e.g.* 07/03/2015, 07/04/2015). Having loosened the constraint on the date, the constraint on municipality was tightened, requiring the equality. It would be reasonable to assume that same age homonyms with an accident at the same time of the day are the same accident, but pairs with two inaccuracies (date and municipality) were chosen not to be included, which in any case resulted a scant minority.

5. Date with an inaccuracy of one, equal municipality, equal concatenation of surname and name, hour with a difference not greater than one, age with a difference not greater than one.
6. Date with an inaccuracy of one, equal municipality, concatenation of surname and name with a Levenshtein distance not greater than one, hour with a difference not greater than one, age with a difference not greater than one.
7. Date with an inaccuracy of one, equal municipality, equal surname, one name contained in the other, hour with a difference not greater than one, age with a difference not greater than one.
8. Date with an inaccuracy of one, equal municipality, equal name, one surname contained in the other, hour with a difference not greater than one, age with a difference not greater than one.
9. Date with an inaccuracy of one, equal municipality, equal anagram of the concatenation of surname and name (explained as typo errors, *e.g.* letter inversion), hour with a difference not greater than one, age with a difference not greater than one.
10. Date with an inaccuracy of one, equal municipality, equal concatenation of surname and name, age with a difference not greater than one.
11. Equal date, equal municipality, equal surname, equal first three letters of name.
12. Equal concatenation of surname and name, equal municipality, date with a difference of at most three days.
13. Equal date, equal municipality, concatenation of surname and name with a Levenshtein distance not greater than one, age with a difference not greater than one.

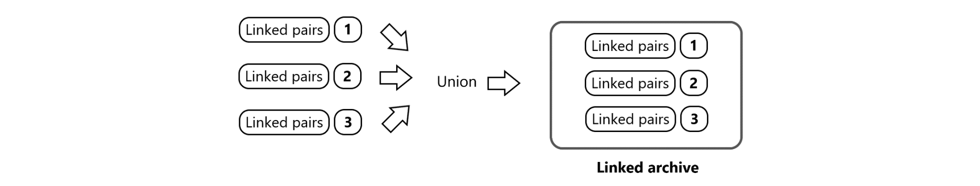
2.3.7 Second deterministic linkage

Given the filters that the probabilistic linkage allowed us to elaborate it seemed a natural extension to try to apply those filters to the unlinked records of the probabilistic linkage to try to get some other linked pairs. This step executes a deterministic linkage on the unlinked records of the probabilistic linkage applying the filters described in paragraph 2.3.6 as linking conditions. Multiple links removal was then performed.

2.3.8 Integration of the partial linked datasets

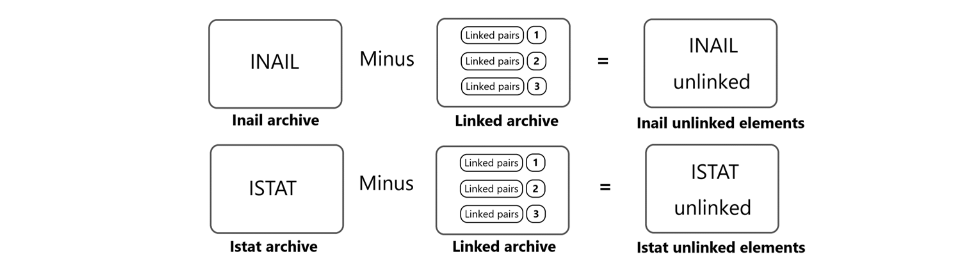
The outputs of the linkage steps described above (partial linked datasets) were then unified to form the final linked road accidents archive. Figure 2.2 sketches the partial linked sets integration. The information to identify which step linked the pair was added to the results, as a metadata, for documenting and improving future data collection and processing. Datasets of unlinked records were calculated at the end of this step by performing a set difference between the starting datasets and the linked dataset. Figure 2.3 sketches the unlinked records calculation.

Figure 2.2 - Union of linked pairs



Source: Authors' processing on Inail and Istat datasets

Figure 2.3 - Calculation of unlinked records



Source: Authors' processing on Inail and Istat datasets

Table 2.5 shows the structure of the linked dataset. The Inail occupational information integrate the Istat accident information.

Table 2.5 - Linked dataset

Personal data			
Age	Gender	Country of birth	Driving license type
Role			
Case			
Inail case code	Istat accident identification number		
Time			
Year	Month	Day (of month)	Date
Day (of week)	Hour	Minutes	
Site			
Macro-region	Region	Province	Municipality
Locality	Coordinate type	Projection system	Latitude
Longitude			
Road			
Road identification code	National road or motorway section	Progressive mileage (Km)	Hectometric
Type of road	Pavement	Road-bed	Weather
Traffic signs	Junction / Non-junction	Localisation of the accident	
Accident consequences			
Outcome	Number of people dead	Number of people injured	Road accident type
Type of injury	Body part injured	Compensation type	Compensated days
Assumed grade of impairment	Actual grade of impairment		
Work			
Commuting	Professional condition		
Vehicle			
Type of vehicle involved			
Economic area			
Economic activity	Large tariff group	Insurance management group	
Linkage info			
Linked by (algorithm step)			

Source: Authors' processing on Inail and Istat datasets

3. Results

Table 3.1 lists the number of accidents registered in each archive.

Table 3.1 - Number of records by archive and year

Archive	2014	2015	2016	2017	2018
Inail	93,056	91,418	93,243	93,673	94,553
Istat	254,528	250,348	252,458	250,128	246,253

Source: Authors' processing on Inail and Istat datasets

Table 3.2 shows the number of linked records by linkage step.

Table 3.2 - Number of linked records by linkage step

Step	Linked pairs	Overall percentage (%)
Deterministic	107,130	83.15
Probabilistic	20,169	15.65
Second deterministic	1,538	1.20
Total	**128,837**	**100.00**

Source: Authors' processing on Inail and Istat datasets

An increase of 18.83% is observed when the probabilistic linkage is applied after the deterministic approach (ratio: 20,169/107,130 = 0.1883).

Table 3.3 shows the number of linked records compared to the size of each dataset.

Table 3.3 - Number of linked records over datasets records

Year	Inail	Istat	Linked	Linked/Inail (%)	Linked/Istat (%)
2014	93,056	254,528	25,383	27.0	10.0
2015	91,418	250,348	24,824	27.0	10.0
2016	93,243	252,458	26,047	28.0	10.0
2017	93,673	250,128	25,872	28.0	10.0
2018	94,553	246,253	26,711	28.0	11.0
Total	**465,943**	**1,253,715**	**128,837**	**26.7**	**10.3**

Source: Authors' processing on Inail and Istat datasets

The number of linked pairs produced by the probabilistic algorithm before filters were applied was 31,147. Table 3.4 shows the proportions of accepted links by filter.

Table 3.4 - Proportions of accepted links by filter

Filter	Accepted links	Accepted links/Total unfiltered links (%)	Accepted links/Total accepted links (%)
Filter 1	14,651	47.04	72.64
Filter 1 on possible matches	95	0.31	0.47
Filter 2	1,660	5.33	8.23
Filter 3	1,857	5.96	9.21
Filter 4	137	0.44	0.68
Filter 5	485	1.56	2.40
Filter 6	0	0.00	0.00
Filter 7	0	0.00	0.00
Filter 8	0	0.00	0.00
Filter 9	98	0.31	0.49
Filter 10	309	0.99	1.53
Filter 11	546	1.75	2.71
Filter 12	137	0.44	0.68
Filter 13	194	0.62	0.96
Total	**20,169**	**64.75**	**100.00**

Source: Authors' processing on Inail and Istat datasets

4. Discussion

As an assessment of linkage quality, precision and recall are often used as measures of the performance of a record linkage process. Precision is the proportion of matches found by the linkage process that are correct, recall is the proportion of correct matches found by the process over all correct matches. The knowledge of correct matches is obtained either by a manual inspection of the data or from a database with known correct matches. We remind here that the goal of this article is not a performance measure, neither a comparison between different algorithms nor to reduce the computational load on probabilistic linking by first running a deterministic linkage, but to observe the improvement in linked pairs produced by a probabilistic linkage run on the unlinked pairs of a deterministic pass. Nonetheless, we are interested in the process accuracy, as we are interested that the linked pairs are correct, so that the improvement that we observe is an improvement of correct matches.

For the deterministic pass, accuracy is enforced by the nature of the variables selected. Municipalities in Italy are very small territorial entities (more than 7,900 in effect on February 2021). The surface that each one encompasses is thereby very small compared to the whole national territory. Thereby it is extremely unlikely the happening of two different accidents in the same municipality, on the same day, with persons having the same name and surname. Thereby is extremely high the chance that such records are referring to the same accident. Moreover, 98% of the linked records with a recorded hour fall in the same 2-hour class, and hour *was not used as linking variable*.

For the probabilistic pass including filters applied afterwards, we observe that filters 1-4 account for most of the linked pairs (more than 90%). More than 95% of those pairs are localised in the same province or municipalties geographically close, yet *no geographical information was used as linking variable*. This gives us enough confidence on the accuracy of the algorithm. In addition, the equality of the identifying fields makes extremely narrow the chances that they can refer to distinct accidents.

Some questions may arise. Would it have been possible to run multiple probabilistic passes using filter variables? Would it have been possible to obtain the same, or better, result with a differently designed probabilistic or deterministic process? Would it have been possible to obtain the same, or better, result with a more sophisticated deterministic strategy applying only deterministic conditions? Probably, but the goal of this article is not to examine *all possible linkage strategies* and pick the best one, the goal of this article is to examine *this strategy* and how this strategy works when applied to road accidents data. The comparison with other designs can then be an open point for further research.

Concluding remarks

This study is part of the Italian National Statistical Programme 2017-2019 (IST-02463 - Analysis of social and health aspects tied to the road accidents phenomenon through Record Linkage with other information sources). Results show that the probabilistic algorithm produced a significant increase in linked pairs compared to the sole use of a deterministic approach. Deterministic filters add further accuracy, in addition to the probability threshold set by the probabilistic algorithm.

To the best of our knowledge, this is a rare example of integrating data between different government entities that are both co-owners and co-managers of the project and, in this way, jointly pursue the common objective of enriching the comprehension of a complex phenomenon. Hence, beyond the common purpose, this joint project requires trusted and shared methodology and algorithms for data linkage. In this work, this is guaranteed by an open-source solution for record linkage, designed and maintained by Istat with the purpose of allowing Italian public entities to apply top-level linkage methodologies even without long experience in record linkage and familiarity with sophisticated statistical modelling. Open-source solution guarantees shareability and, most of all, trustworthiness of procedures and results, particularly relevant in this context, due to the double ownership of the involved data.

The application of record linkage techniques between Istat and Inail archives is useful to join pieces of information, providing added value, enhancing the potentialities of data with different origin and filling information gaps. It points out aspects that only through a joint analysis can stand out and that can be essential for injury prevention. The joint analysis can extract additional information to better characterise the accidental event and infer potential associations with risk factors (Bruzzone *et al.*, 2021; Gariazzo *et al.*, 2021; Pireddu *et al.*, 2021).

This characterisation of the accidents and the related work profiles can help to define risk prevention programmes with a resulting mitigation of the phenomenon and its impact on society and public health.

References

Brusco, A., A. Bucciarelli, M. Bugani, C. Gariazzo, C. Giliberti, M. Marinaccio, S. Massari, A. Pireddu, L. Veronico, G. Baldassarre, S. Bruzzone, M. Scortichini, M. Stafoggia, e S. Salerno. 2019. *Gli incidenti con mezzo di trasporto. Un'analisi integrata dei determinanti e dei fattori di rischio occupazionali*. Roma, Italy: Inail.

Bruzzone, S., A. Altimari, G. Baldassarre, R. Boscioni, L. Veronico. 2021. "Work-related road accidents: an in-depth statistical analysis carried out by two different integrated data sources". In this issue of the *Rivista di statistica ufficiale*, N. 3/2021. Roma, Italy: Istat.

Christen, P. 2012*a*. "The data matching process". In Christen, P. *Data matching. Concepts and Techniques for Record Linkage, Entity Resolution, and Duplicate Detection*: 23-35. Heidelberg, Germany: Springer, *Data-Centric Systems and Applications*.

Christen, P. 2012*b*. "A Survey of Indexing Techniques for Scalable Record Linkage and Deduplication". *IEEE Transactions on Knowledge and Data Engineering*, Volume 24, Issue 9: 1537-1555.

Fellegi, I.P., and A.B. Sunter. 1969. "A theory for record linkage". *Journal of the American Statistical Association*, Volume 64, Issue 328: 1183-1210.

Gariazzo, C., A. Marinaccio, S. Bruzzone, L. Taiano, and L. Veronico. 2021. "Work-related road accidents: a statistical multivariate analysis in Italy". In this issue of the *Rivista di statistica ufficiale*, N. 3/2021. Roma, Italy: Istat.

Hernandez, M.A., and S.J. Astolfo. 1995. "The merge/purge problem for large databases". *ACM Sigmod Record*, Volume 24, Issue 2: 127-138.

Herzog, T.N., F.J. Scheuren, and W.E. Winkler. 2007. *Data Quality and Record Linkage Techniques*. New York, NY, U.S.: Springer Science+Business Media, LLC.

Istituto Italiano di Statistica - Istat (*Italian National Institute of Statistics - Istat*). "RELAIS (REcord Linkage At IStat)". *Methods and Tools*. Roma, Italy: Istat. https://www.istat.it/en/methods-and-tools/methods-and-it-tools/process/processing-tools/relais.

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*, Volume 84, N. 406: 414-420.

Pireddu, A., A. Altimari, G. Baldassarre, A. Marinaccio, and L. Taiano. 2021. “An analysis on work-related road injuries by macro-economic sector, road type and Italian territorial divisions”. In this issue of the *Rivista di statistica ufficiale*, N. 3/2021. Roma, Italy: Istat.

Scanu, M. (*a cura di*). 2003. “Metodi statistici per il record linkage”. *Metodi e Norme*, N. 16. Roma, Italy: Istat.

Tuoto, T., S. Bruzzone, L. Valentino, G. Baldassarre, N. Cibella and M. Pappagallo. 2012. “Towards an integrated surveillance system of road accidents”. In Società Italiana di Statistica - SIS. *Atti della XLVI Riunione Scientifica (Roma, 20-22 giugno 2012)*. Padova, Italy: CLEUP.

Tuoto, T., A. Burgio, R. Cotroneo, C. Iaccarino, S. Prati, F. Rinesi, F. Rottino, L. Tosco, e L. Valentino. 2015. “Fecondità e maternità: un sistema integrato per la misurazione di fenomeni sanitari e socio-demografici”. *Rivista di statistica ufficiale*, N. 3/2015: 49-70. Roma, Italy: Istat. https://www.istat.it/it/archivio/187424.

Tuoto, T., L. Corallo, N. Cibella, D. Ichim, V. Mastrostefano, A. Nurra, e M. Rinaldi. 2014. “L’integrazione dei risultati delle indagini sulla tecnologia e l’innovazione nelle imprese: una sperimentazione”. *Rivista di statistica ufficiale*, N. 3/2014: 97-128. Roma, Italy: Istat. https://www.istat.it/it/archivio/152014.