

Probabilistic Record Linkage for Integrated Surveillance of Road Accidents

Tecniche di record linkage probabilistico per la sorveglianza integrata degli incidenti stradali

F. Chini¹, S. Farchi¹, M. Fortini², V. Greco², T. Tuoto²

¹Agenzia di Sanità Pubblica della Regione Lazio, chini@asplazio.it

²ISTAT, tuoto@istat.it

Riassunto: Le tecniche di record linkage suscitano un interesse crescente poiché permettono il migliore sfruttamento di informazioni già disponibili. L'efficienza delle diverse tecniche dipende dal potere di identificazione delle variabili chiave disponibili. In questo lavoro sono presentati i primi risultati di un record linkage probabilistico realizzato in assenza delle variabili identificatrici comunemente considerate più discriminanti per l'abbinamento di individui (nome, cognome e indirizzo). Scopo del lavoro è quello di integrare le informazioni raccolte dalle autorità di polizia in occasione di incidenti stradali con esito mortale con i dati di carattere sanitario riguardanti le vittime di tali incidenti, così come raccolti nei certificati di morte. I risultati della sperimentazione sono incoraggianti.

Keywords: Administrative Data, Matching, Matching Errors.

1. Introduction

Road accidents (RA) are the leading cause of death for young adults in industrialised countries (Peden et al. 2004); RA cause 8000 deaths and 20000 permanent disabled persons per year in Italy (Taggi et al. 2001). The available nationwide data about RA come from police reports, which list the number of accidents, deaths, and “injured people” (ISTAT-ACI, 2004). Unfortunately, police reports provide few personal and health information about people involved in RA; indeed, their role is legal rather than medical. A further source of personal information is available from death causes that are collected by means of the regional hospital information systems; later such causes are also gathered by Istat. Moreover, few Regions have recently enriched their systems with information coming from the emergency department (ED) admissions, both public and private. Though these different data sources, when independently considered, are important for the monitoring of RA, the integration of medical and non medical data is essential for the construction of a monitoring system, in order to engage preventive and repressive actions. An integrated monitoring allows to obtain a very sensitive and accurate control on two principal aspects of the problem: the circumstances of the accident, including relevant risk factors, and the health consequences. Nevertheless, so far such integration has been prevented by the absence of common variables capable to permit the personal identification through the two lists.

This paper reports the preliminary results of an experiment of probabilistic record linkage (RL) between RA and sanitary information, when personal identifiers (names and addresses) are missing. The considered data concern the Lazio region, about 5,3 million inhabitants, which is the region of central Italy that includes Rome (3 million inhabitants).

2. Data sources and study context

The first considered data source is the Mortality Registry (MR) related the 2000 year, which collects all the death certificates registered inside the territory of the Lazio region. It reports: name, date and place of birth, date and place of death and the IDC 9 cause of death. This source was integrated with information collected in two further sanitary databases managed by the regional health authority, namely:

- the Emergency Information System (EIS) that collects all emergency ward admissions in Lazio;
- the Hospital Information System (HIS) that collects all the hospital discharges occurred in the Lazio region;

Specifically, we used the dates of ED admissions and discharges if the person did not die immediately but survived to the accident up to a maximum of thirty days.

The further considered source is the RTA Information System, managed by Istat, which collects information coming from police reports. It reports age and gender of the victims and of the other people involved in the road accidents; moreover it includes information about date, time, nature of accident, type of vehicles, weather and road conditions. Instead, names and surnames of people involved in RTA are not reported.

A first step of the research considers the easier problem of recognising the 663 victims of road accidents registered in the RA information system for the year 2000, among the approximately 49000 deaths included in MR for the same year. A second step, not discussed here, will take into account also the linkage of the people injured in RA.

Besides the unfeasibility to use names and surnames as key variables, a further problem is that only age and sex of people can directly be used as key variables; the remaining information have rather to be properly modified for the purpose of being useful in the RL process. In order to better implement the RL, the subset of the 149 people collected in RA by the state police, for which the first name and the family name of the victims are available, has been used. Since these variables are available also for the people included in MR, for this subset it was possible to link the two sources with certainty. These results are then exploited as a reference to evaluate the performance of the RL procedure when the less powerful key variables are used.

3. Methods

Let RTA_1 be the subset obtained from the original RTA by considering only the 149 RTA people with known names and surnames. A deterministic linkage that used names and surnames as key variables was performed between RTA_1 and MR, identifying 141 match; this result was assumed as a benchmark for the RL. Moreover, a smaller MR, called MR_1 in the following, was built by adding the 141 matches to a subset of 200 people randomly sampled from the residual MR. This dramatic simplification made the

RL esier to treat at this initial stage. After these manipulations a RL with the Fellegi-Sunter approach (Fellegi and Sunter, 1969) was performed on RTA_1 and MR_1 in order to re-identify the matches. Within the set Ω of all ordered pairs $\Omega = \{(a,b); a \in RTA_1, b \in MR_1\}$, the decision to link/not link was based on the available key variables age, gender, date and place of accident. As already discussed, date and place of accident in RTA_1 were compared to the most appropriate between date and place of death or date and place of ED admission found in MR_1 , in the case the RA person dead immediately or within a 30 days period.

For the i -th key variable, $i=1, \dots, 4$, a γ agreement/disagreement indicator variable was established in order to create a vector of comparisons γ for every (a,b) pair so that:

$$\gamma_{(a,b)} = [\gamma_{(a,b)}^1, \gamma_{(a,b)}^2, \gamma_{(a,b)}^3, \gamma_{(a,b)}^4], \quad j=1 \dots 4.$$

A matching composite weight w was then computed for each possible γ via the EM algorithm (Jaro 1989, Winkler 2000) so that:

$$w_{(a,b)} = \ln \frac{\Pr(\gamma_{(a,b)} | M)}{\Pr(\gamma_{(a,b)} | U)} = \ln \frac{m_{(a,b)}}{u_{(a,b)}}$$

where M is the set of matching pairs and U is the set of not matching pairs, with $M \cup U = \Omega$ and $M \cap U = \emptyset$.

According to Fellegi and Sunter (1969) a pair (a,b) is classified as a match and assigned to the set M if $w_{(a,b)}$ was greater than a given threshold value t_I ; similarly the pair (a,b) is classified as a not match if $w_{(a,b)}$ was lower than a further threshold t_0 ($t_0 < t_I$). Finally, all the pairs that exhibit a value $t_0 < w_{(a,b)} < t_I$ are considered for a clerical review. As a consequence, it is important to carefully control the number of false match and false non-match errors that arise from the selection of the thresholds.

4. Results

By comparing the results of RL with the independent and more reliable matching obtained via names and surnames (henceforth true link, TL), it is possible to calibrate the RL procedure. Let y_{FM} be the number of pairs detected as matches by RL but resulted as not matches when checked against the TL. Moreover, let y_{FNM} be the number of pairs detected as non matches by RL but resulted as matches when checked against the TL. Finally, let y_{TM} be the number of pairs detected as matches both by the RL and the TL.

From the quantities defined above, the computation of false match rates (FMR) and false non match rates (FNR) is straightforward:

$$FMR = \frac{y_{FM}}{y_{FM} + y_{TM}}; \quad FNR = \frac{y_{FNM}}{y_{FNM} + y_{TM}}.$$

¹ Since in this context each RTA_1 unit cannot match to more than one MR_1 unit, some pairs (a,b) for which $w_{(a,b)} > t_I$ were discarded from M by means of the assignment procedure shown in Jaro (1989).

These quantities correspond to the well-known $1-\alpha$ and β error type, respectively, if a one-tail hypothesis test is considered. Table 1 shows the percentage of linked units, the FMR, the FNR and the number of pairs in the set M , depending on different values of the weight $w_{(a,b)}$. It can be seen that, by choosing the value 3.91 as the threshold t_l , a considerable linkage percentage of 90% can be achieved. Nevertheless, a value of 3% for the FMR, thought not so noticeable, is higher than the one obtained if t_l is kept equal to 6.55. These considerations lead to choose t_l equal to 6.55 and t_0 equal to 3.91. In this way, FMR and FNR can be chosen respectively as close as possible to 1% and to 9% by carrying out a clerical review on only 109 pairs, as it can be seen by subtracting 105 from 214 in the last column of table 1.

Table 1. Results of the comparison between RL and TL

$w_{(a,b)}$ (1)	Linked units (%) (2)	FMR (3)	FNR (4)	Nr. of pairs (a,b) so that $w_{(a,b)} > t_l$ (5)
6.55	55	0.01	0.44	105
6.21	87	0.03	0.12	170
4.61	89	0.03	0.09	190
3.91	90	0.03	0.09	214

5. Conclusions and further work

The figures shown above suggest that RL techniques could successfully approach the problem of matching two datasets even in absence of highly discriminating key variables. However, these results represent a starting point because of the larger dimension of the whole RA problem. Indeed, dealing with larger datasets it will require planning and introducing the most adequate blocking strategy in that context. Moreover, an accurate estimate of the parameters m and u will be necessary, in absence of known perfectly linked data, in order to correctly evaluate the FMR and FNR rates. In this direction, our efforts are towards the application of models belonging to the class of log-linear models with latent variables (Armstrong and Mayda, 1993).

References

- Armstrong, J. A., Mayda, J. E. (1993) Model-based Estimation of Record Linkage Error Rates, *Survey Methodology*, 19, 137-147.
- Fellegi I.P., Sunter A.B. (1969) A Theory for record linkage, *Journal of the American Statistical Association*, 64, 1183-1210.
- ISTAT-ACI. (2003) Statistica degli incidenti stradali - anno 2001. *Informazioni*, 18, Roma, ISTAT.
- 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*, 89, 414-420.
- Peden M., Scurfield R., Sleet D., et al.(2004) *World report on traffic injury prevention*. Geneva, World Health Organization.
- Taggi F, Giustini M, Fondi G, et al. (2001) L'epidemiologia degli incidenti stradali (I): i dati di base e i fattori di rischio. In Taggi F. Di Cristofaro Longo G. eds. *I dati socio-sanitari della sicurezza stradale*. Roma, Istituto Superiore di Sanità.