CONTRIBUTI ISTAT

n. 14/2008

Probabilistic Record Linkage for the Integrated Surveillance of Road Traffic Injuries when Personal Identifiers are Lacking

F. Chini, M. Fortini, T. Tuoto, S. Farchi,
P. Giorgi Rossi, R. Amato e P. Borgia

Istat

SISTEMA STATISTICO NAZIONALE
ISTITUTO NAZIONALE DI STATISTICA

Le collane esistenti presso l'ISTAT - *Rivista di Statistica Ufficiale*, *Contributi ISTAT* e *Documenti ISTAT* - costituiscono strumenti per promuovere e valorizzare l'attività di ricerca e per diffondere i risultati degli studi svolti, in materia di statistica ufficiale, all'interno dell' ISTAT, del SISTAN, o da studiosi esterni.

La *Rivista di Statistica Ufficiale* accoglie lavori che hanno come oggetto la misurazione dei fenomeni economici, sociali, demografici e ambientali, la costruzione di sistemi informativi e di indicatori, le questioni di natura metodologica, tecnologica o istituzionale connesse al funzionamento dei sistemi statistici e al perseguimento dei fini della statistica ufficiale.

I lavori pubblicati in *Contributi ISTAT* sono diffusi allo scopo di stimolare il dibattito intorno ai risultati preliminari di ricerca in corso.

I *Documenti ISTAT* forniscono indicazioni su linee, progressi e miglioramenti di prodotto e di processo che caratterizzano l'attività dell'Istituto.

Il Comitato di redazione esamina le proposte di lavori da pubblicare nelle tre collane sopra indicate. Quelli pubblicati nei Contributi ISTAT e nei Documenti ISTAT sono valutati preventivamente dai dirigenti dell'Istituto, mentre i lavori pubblicati nella Rivista di Statistica Ufficiale sono subordinati al giudizio di referee esterni.

ISTAT – Produzione libraria e centro stampa

DOCUMENTI ISTAT

n. 14/2008

Probabilistic Record Linkage for the Integrated Surveillance of Road Traffic Injuries when Personal Identifiers are Lacking

F. Chini( ), M. Fortini(**), T. Tuoto(***), S. Farchi(*), P. Giorgi Rossi(*), R. Amato(****) e P. Borgia(*)*

(*) Agenzia Sanità Pubblica, Regione Lazio
(**) ISTAT - Servizio Metodologie per i censimenti
(***) ISTAT - Servizio Progettazione e supporto metodologico nei processi di produzione
(****) ISTAT – Servizio Incidentalità stradale

Contributi e Documenti Istat 2008

Istituto Nazionale di Statistica
Servizio Produzione Editoriale

Produzione libraria e centro stampa:
Carla Pecorario
Via Tuscolana, 1788 - 00173 Roma

Sommario: Road traffic injuries are the leading cause of death among young adults in industrialised countries. All the data sources available for the surveillance of road traffic accidents have important limits, when taken separately. Therefore the integration of medical and non medical data is essential in order to build up a surveillance system so as to drive both preventive and repressive actions. This has been impossible up until now because of the absence of common variables that would allow an accurate joining of the lists. To overcome such difficulty, this study proposes the use of record linkage techniques, pointing out the feasibility of probabilistic linkage without the use of personal identifiers.
The linkage was carried out between the deaths collected in the road traffic injuries information system, managed by the Italian National Statistical Institute, and the records belonging to the mortality register, managed by the health authority. All data refers to the year 2000 and the Lazio Region.

Key words: probabilistic record linkage, deterministic linkage, exact matching, linkage error rates, injury surveillance

1. Introduction

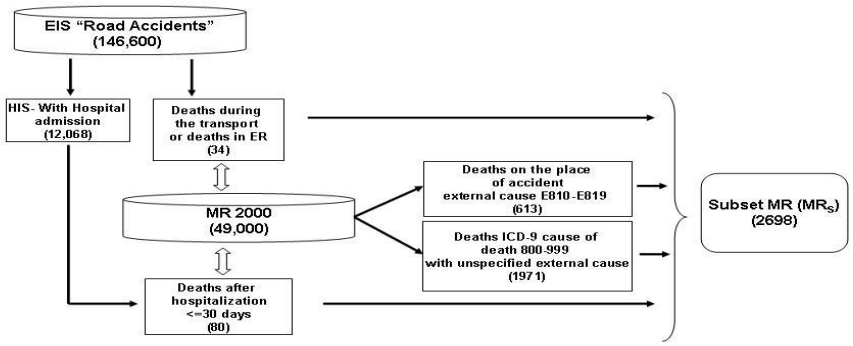
Road traffic injuries are the leading cause of death for young adults in industrialised countries (Peden et al. 2004). In 2004, in Italy there were 6,000 deaths and 320,000 injured caused by Road traffic injuries (Piffer 2004).
Injury prevention, particularly road accident prevention is one of the major challenges of the World Health Organization for both industrialized and developing countries (WHO 2002).
The European Union for road safety set the goal to halve the number of deaths by 2010 (European Council, 1997).
Few evidence-based campaigns for prevention have been set up in Italy and epidemiological surveillance of the health consequences of road accidents has been implemented only in a few local settings (Valent et al. 2002).
The available nationwide data comes from police reports which list the number of accidents, deaths, and "injured people" (ISTAT-ACI, 2001). Unfortunately, police reports provide little information on the health effects, indeed their role is legal, rather than medical. Several studies using health information systems have showed different figures of the diseases burden caused by road accidents, usually characterised by high incidence rates (Cercarelli et al. 1996; Langley 1995; Ferrante et al. 1993).
Otherwise, a source of personal information is available from death causes that are collected by means of the regional hospital information systems, and subsequently pooled by Istat. Moreover, a few Regions have recently augmented their systems with information coming from the emergency department admissions, both public and private.
All the different data sources available for the surveillance of road traffic accidents have important limits, when taken separately. Therefore the integration of medical and non medical data is essential to build up a surveillance system so as to drive both preventive and repressive actions (Langley 1995; Ferrante et al. 1993; Sniezek et al. 1989).
Nevertheless, this has been impossible up until now because of the absence of common variables that allow an accurate joining of the lists.
This study aims to assess the feasibility of a probabilistic record linkage, without personal identifiers (name, surnames), between the deaths collected in the road traffic information system and the records belonging to the mortality register, integrated with information in the sanitary database. Data sources are described in section 2, whereas in section 3 an overview of the probabilistic record linkage methodologies is decribed. Section 4 outlines a trial application carried out in order to tune the main procedure described in section 5. Finally in the last section, benefits, drawbacks and further developments of the procedure are discussed.

2. Data sources

The data considered here concerns the Lazio Region, populated with about 5,3 million inhabitants, which is the region of central Italy that includes Rome (about 3 million inhabitants).
The first source of data to be considered is the Road Traffic Injuries Information System for the years 2000 (RTIIS), managed by Istat, which collects information coming from police reports (663 units). 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, type of accident, types of vehicles, wheather and road conditions. Unfortunately, names and surnames of people involved in road traffic injuries are reported only for 149 of them.
The other source considered is the Mortality Register (MR) for the year 2000 (49,000 units), 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 previously augmented by information from two other sanitary databases managed by the regional health authority, namely:
Emergency Information System (EIS) for the year 2000, which collects all emergency ward admissions in Lazio. For each emergency department admission, EIS reports: the name, date and place of birth of

the patient, the date of admission, the triage code, up to four diagnoses and up to four therapeutic procedures (both, diagnoses and procedures coded according to ICD-9-CM), the outcome of the admission (hospitalisation, death, transfer or discharge) and, in the case of trauma, the place were the accident occurred ("road", "work", "home", "intentional violence" and "other").
Hospital Information System (HIS) for the year 2000, that collects all the hospital discharges which occurred in the Lazio region. It reports name and place of birth, the date of hospitalization, up to six diagnoses, up to six therapeutic procedures and the outcome of the discharge.
Integration of records in the EIS, HIS and MR systems was straightforward because these three systems were already integrated into a medical information system.
In our analysis we considered a subset of MR (2,698 units) obtained by selecting only the deaths related traffic accidents, according to the following rules:
Deaths coming from EIS with hospital admission or deaths during the transport and deaths in emergency room (114 units)
Deaths coming from MR where the external causes were road traffic injuries (coded in the range E810-E819) or, when not specified, the cause of death was coded in the range 800-999 ICD-9 CM (injury and poisoning) (2,584 units).
Figure 1 shows selection process.

Figure1. *Selection process subset of Mortality Register (MR)*



3. Probabilistic record linkage

Record linkage consists in matching the records belonging to different data sets when they correspond to the same statistical unit (Belin and Rubin, 1985). Let A and B be two partially overlapping computer files regarding the same type of units (e.g. individuals, firms). Suppose also that the two files consist of vectors of variables (X_A,Z_A) and (X_B,U_B), either quantitative or qualitative, assuming that X_A and X_B are sub-vectors of common identifiers, called key variables in what follows, so that any single unit is univocally identified by an observation x. The goal of record linkage is to find all the pairs of units $(a,b) \in \Omega=\{(a,b): a \in A, b \in B\}$, such that a and b refer actually to the same unit (a=b). Hence, a record linkage procedure is a decision rule based on the comparison of the key variables which, for each single pair of records, can take either one of the following decisions: link, possible link and non-link (Fellegi and Sunter, 1969). Since the key variables can be prone both to measurement errors and misreporting, the record linkage problem is far from being a trivial one and probabilistic techniques are used to minimize the incidence of false and missed links.
Following Fellegi and Sunter (1969), Ω is the set of all possible pairs of records from XA and XB, i.e.

$\Omega=\{(a,b): a\in A, b\in B\}$. In addition, M indicates the set of matches, i.e. the pairs related to the same unit [$M=\{(a,b): a = b\}$], whereas U denotes the set of nonmatches, namely the couples made up of different units [$U=\{(a,b): a \neq b\}$], so that $M\cap U=\varnothing$ and $\Omega= M\cup U$. Denoting with $|S|$ the cardinality of a set S, we can note that $|M|$ is typically much smaller than $|U|$.
If the x_A and x_B represent the vectors observed on the K key variables ($X_1, X_2, \ldots, X_k$) respectively for the file X_A and X_B, the comparison between two units $(a,b)\in\Omega$ can be expressed by a vector of K indicator functions $\gamma^{ab}=\left(\gamma_1^{ab}, \gamma_2^{ab}, \ldots, \gamma_k^{ab}\right)$, where

$$\gamma_j^{ab}=\begin{cases}1 & \text{if } x_{a,j}^{A}=x_{b,j}^{B}\\ 0 & \text{otherwise}\end{cases} \qquad (1)$$

The vector of comparison variables is used by Fellegi and Sunter (1969) for designating which pairs have to be considered as a match, defining a weight w(a,b) so to take into account the following likelihood ratio:

$$w_{(a,b)}=\frac{\Pr\left(\gamma^{ab} \mid M\right)}{\Pr\left(\gamma^{ab} \mid U\right)}=\frac{m_{(a,b)}}{u_{(a,b)}}$$

In doing so the decision can be taken as long as the log of the weight is either greater or smaller than two given thresholds K_1 end K_2 respectively,

$$\log(w) < K_1$$
$$K_1 \le \log(w) \le K_2$$
$$\log(w) > K_2$$

More precisely, a pair is classified as a link if log (w) is above the threshold K_2, and as a non-link if it lays below K_1; if log (w) falls in the range (K_1 , K_2) no-decision is made and the pair is held out for the clerical review so to be solved. A decision on the threshold levels has to be made in order to manage properly the trade off between the need of a small number of expected no-decisions (Fellegi and Sunter, 1969) and small misclassification error rates for the pairs.
To estimate m and u Jaro (1989) defines a latent vector

$$\mathbf{g}_{(a,b)}=\begin{cases}\langle 1,0\rangle & \text{if } (a,b)\in M\\ \langle 0,1\rangle & \text{if } (a,b)\in U\end{cases}$$

and the augmented log-likelihood for the observed vector x of the key variables and the vector g

$$\ln\left[f\left(\mathbf{x},\mathbf{g}|\mathbf{m},\mathbf{u},p\right)\right]=\sum_{(a,b)\in\Omega}\mathbf{g}_{(a,b)}\begin{pmatrix}\ln\left(\prod_k (m_k^{ab})^{\gamma_k^{ab}}\left(1-m_k^{ab}\right)^{\left(1-\gamma_k^{ab}\right)}\right)\\ \ln\left(\prod_k (u_k^{ab})^{\gamma_k^{ab}}\left(1-u_k^{ab}\right)^{\left(1-\gamma_k^{ab}\right)}\right)\end{pmatrix}$$
$$+\sum_{(a,b)\in\Omega}\mathbf{g}_{(a,b)}\begin{pmatrix}\ln(p)\\ \ln(1-p)\end{pmatrix} \qquad (2)$$

where p represents the probability that a randomly chosen pair (a,b) belong to the subset M.
Moreover, a conditional independence assumption is often made, so that

$$m^{a,b} = \prod_{k=1}^{K} m_k^{a,b} \; ; \qquad m^{a,b} = \prod_{k=1}^{K} m_k^{a,b} \; .$$

where

$$m_k^{ab} = \Pr\left(\gamma_k^{ab} = 1 \middle| (a,b) \in M\right); \qquad u_k^{ab} = \Pr\left(\gamma_k^{ab} = 1 \middle| (a,b) \in U\right);$$

Since the vector g and the subsets M and U cannot be directly observed, the probabilities m and u are estimated via the EM procedure (Dempster et al. 1977), providing initial values for m, u and p and estimating expected values for the vector g=<gm, gu> (STEP E)

$$\hat{g}_m\left(\gamma^{ab}\right) = \frac{\hat{p}\prod_{k=1}^{K}\left(m_k^{ab}\right)^{\gamma_k^{ab}}\left(1-m_k^{ab}\right)^{\left(1-\gamma_k^{ab}\right)}}{\hat{p}\prod_{k=1}^{K}\left(m_k^{ab}\right)^{\gamma_k^{ab}}\left(1-m_k^{ab}\right)^{\left(1-\gamma_k^{ab}\right)} + (1-\hat{p})\prod_{k=1}^{K}\left(u_k^{ab}\right)^{\gamma_k^{ab}}\left(1-u_k^{ab}\right)^{\left(1-\gamma_k^{ab}\right)}}$$

$$\hat{g}_u\left(\gamma^{ab}\right) = 1 - \hat{g}_m\left(\gamma^{ab}\right)$$

After this step, the g values can be placed into the log-likelihood [2] and a maximum likelihood estimate for m, u and p (STEP M) can be obtained from

$$\hat{m}_k = \frac{\sum_{(a,b)\in\Omega}\hat{g}_m\left(\gamma^{ab}\right)\gamma_k^{ab}}{\sum_{(a,b)\in\Omega}\hat{g}_m\left(\gamma^{ab}\right)}; \qquad \hat{u}_k = \frac{\sum_{(a,b)\in\Omega}\hat{g}_u\left(\gamma^{ab}\right)\gamma_k^{ab}}{\sum_{(a,b)\in\Omega}\hat{g}_u\left(\gamma^{ab}\right)}; \qquad \hat{p} = \frac{\sum_{(a,b)\in\Omega}\hat{g}_m\left(\gamma_k^{ab}\right)}{N}.$$

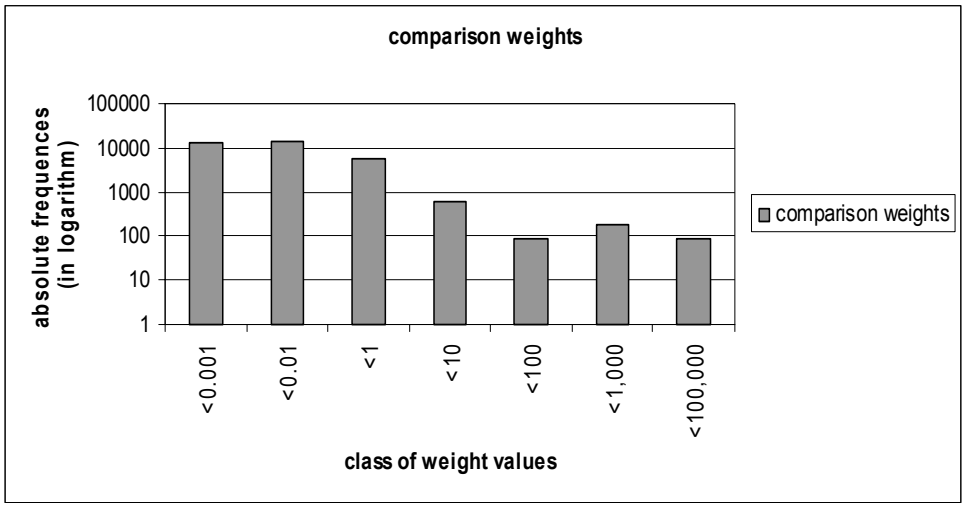
The Expectation and the Maximization steps are then iterated until the convergence of the parameters of interest is achieved. In order to avoid local maxima of the likelihood, in the procedure a lattice of different initial values should be considered.
As long as model parameters can be identified, the conditional independence assumption can be relaxed and the estimates of u and m can be still obtained in the framework of the latent class estimation model (Armstrong and Mayda, 1993; Hagenaars, 1993).

4. Training linkage

The two sources can be joined through information concerning both the victims (names, surnames, age and gender) and the accidents (date and place). Unfortunately only a few records (149 out of 663) from RTIIS report personal information, therefore names and surnames cannot be considered as key variables on the whole set of data. Moreover MR and its related sources (referred to as MR from now on) do not include date and place of the accident, but rather date and place of the death and of hospital admittance. The latter variables, though not identical to those in RTIIS, seem to be comparable with the former ones. In particular, the date of the accident can be compared with the date of death for those who died immediately, whereas it can be compared with the date of hospital admittance for those who died in 30 days from the accident. For the sake of brevity, in the following we will refer to this comparison variable only as "date of the accident". The same consideration can be applied to the "place of the accident".
As a starting point, we implemented a deterministic linkage on that subset of 149 people collected by the State Police, for which the names and the surnames of the victims were available (called RTIIS-e). For this subset, the two sources are linked with certainty, since names and surnames can be considered as perfect key variables. This result was then leveraged as a gold standard in order to evaluate the

amount of linkage errors made when using the less powerful key variables. The deterministic linkage between the RTIIS-e and the MR identified 138 matches.
Before proceeding to the probabilistic linkage on the whole set of data in RTIIS file, we tried a probabilistic record linkage between the RTIIS-e subset and the MR file, considering as key variables the gender and the age of victim, and the date and place of the accident. Therefore, the results of the probabilistic and deterministic record linkage were compared, obtaining useful hints on the main difficulties to be faced and the rate of the expected linkage errors.
The distribution of the comparison weights w, estimated via the EM algorithm, is shown in figure 2.

Figure 2. *Distribution of comparison weights on the training set.*

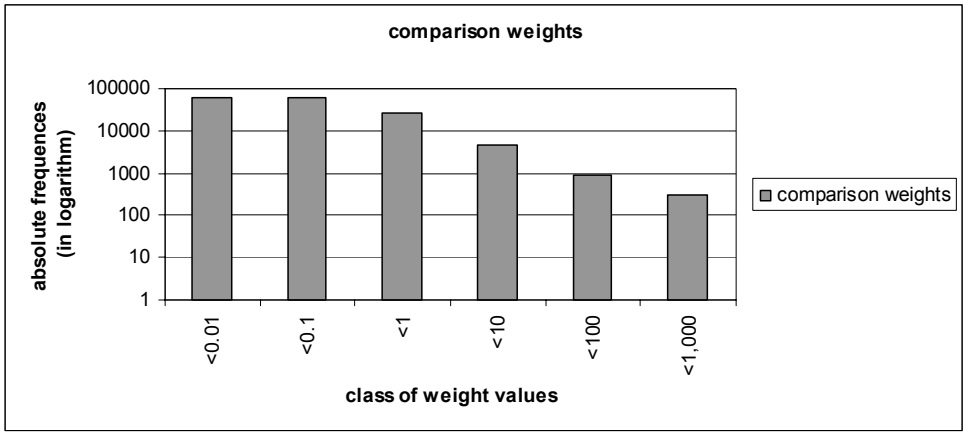


In accordance with the Fellegi-Sunter decision rule, two thresholds should be chosen in order to minimize the manual review, taking fixed the expected misclassification errors. Nevertheless, since the clerical review is not conducted due to practical constraints, we decided to determine only one threshold, whose value was fixed to 10 in order to decrease as much as possible the expected number of erroneously non-linked units. Finally, among the pairs lying beyond the threshold, only a subset were linked in order to satisfy the one-to-one constraint, explained in detail in the next section.
At the end of the procedure, 139 links are identified, instead of the 138 true links coming from the deterministic match based on the name and surname keys. This result suggests to apply the probabilistic record linkage to the whole set of data using the weak identifying keys. However, the analyses of the linkage errors required some cautions. The linkage performances were evaluated in terms of false match rate and false non-match rate (Winkler, 1995). The false match rate and the false non-match rate correspond to the well-known type II and type I errors in a one-tail hypothesis test context. False non-matches are the more common and occur when records which should have been assigned to the same unit are instead not matched together. False matches are less common but potentially more serious because of further analyses on erroneously linked data could lead to biased statistics. Other authors consider performance measures in terms of positive predicted value and sensitivity (Gu et al. 2003; Gomatam et al. 2002), that consist of the algebraic transformation of the false match rate and the false non-match rate.
Since the true matching status of the records on the RTIIS-e file is actually known, it is possible to calculate the false match rate and the false non-match rate, that take on the values 26% and 25%, respectively. Those figures are higher than those usually reported in literature (Ding and Fienberg, 1994), reflecting the difficulty of linking without strong identifying keys.
Different results can be reached by choosing different threshold values; for example, a more restrictive threshold, equal to 180, halves the false match rate, 12%, and consequently almost doubles the false non-match rate, 45%, due to the identification of less links, just 86 over the 138 true links.

5. Main linkage application

The figures shown allow us to be confident that, in the present context, the probabilistic record linkage techniques could successfully address the problem of matching two datasets in the presence of weak key variables. Therefore, the key variables considered in the experimental probabilistic linkage already described; age and gender of victim, date and place of accident, were used to conduct the probabilistic record linkage between the RTIIS and the MR files. The resulting 1,788,774 pairs, given by the cross-product of the two files, were reduced by deleting the pairs for which more than 30 days occurred between the accident and the death, according to the convention that a death which occurred 30 days after the road accident cannot be attributed to the accident itself. This method of operating a pair-reduction falls within the blocking procedures, in particular it resembles the sorted neighbourhood method (Baxter et al. 2003; Hernandez and Stolfo, 1998). Doing so, the number of pairs was finally reduced to 152,057. The linking weight w was then assigned to the pairs according to the Jaro (1989) estimation procedure.
Before comparing these weights with the thresholds, as suggested by Fellegi-Sunter, it is necessary to guarantee that the links will be chosen so that each record in the RTIIS file is assigned at most to one record in the MR file, and vice versa. According to Jaro (1989), we look at a linking assignment scheme that maximises the sum of the weights of the assigned links. This means solving a linear sum assignment problem and therefore such a linear programming model can be used.
Once the matching problem was constrained to comply a one-to-one assignment, the pairs can be classified as matches if the corresponding weights are greater than a threshold value. As was just explained referring to the training linkage, due to practical constraints, only one threshold was determined, classifying as matches (non-matches) all pairs with weights larger (smaller) than that threshold. To identify it, a visual analysis of the distribution of the record pairs comparison weights was carried out.

Figure 3. *Distribution of comparison weights.*



Looking at the distribution of weights, reported in figure 3, it must be pointed out that the mode of the non-matching population is easily recognisable whereas the matching mode is not so easily identifiable; this fact gives us an indication of the level of difficulty of the linkage task, and consequently predicts that the amount of errors of miss-classification will not be very small. Anyway the threshold value between the two populations could be determined to be around 10 in order to avoid both introducing too many non-matches and at the same time without loosing true matches.
Fixing the threshold, 431 pairs are recognised as matches, corresponding to 65% of the records of the RTIIS file. This represents a good outcome when compared with that one (lower than 21%) resulting from the exact record linkage based on name and surname. This means that future analyses involving variables coming from the two sources, if based on the probabilistic record linkage results, can make use of an amount of data three time larger than those resulting from exact record linkage.
However, the results have to be properly interpreted by taking into account the errors that the probabilistic linkage procedures generate. Generally, in order to measure the quality of record linkage,

the true matching status is investigated on a subset of pairs, suitably sampled from all the pairs, or coming from previous experiences on similar record linkage fields. The underling assumption is that the randomly chosen sample of pairs behaves similarly to the whole data with respect to the error rates. In this work, we adopt an estimate of the matching errors based on the knowledge of the true matching status with regard to the subsets of records with names and surnames. The match rate, defined as the number of linked record pairs divided by the total number of true match record pairs, calculated on the subset is considerably high, equal to 96%. At the same time, the false match rate is equal to 27% and the false non-match rate is 30%.
Also, in this case if we had taken a more restrictive threshold, e.g. equal to 100, the false match rate would have gone down to 7%, and the false no-match rate would have increased up to 49%, finally we would have obtained 216 link, 78 more than those coming from the deterministic linkage.

6. Concluding remarks

In the empirical study addressed up to this point, the probabilistic record linkage could lead to significant outcomes also when only weak matching variables are available. However, some comments are needed. As expected, the results (i.e. false match rate and false non-match rate) are worse for the main linkage procedure than for the training one, although the linkage procedure was the same, due to the higher number of possible pairs. In this work, we based the estimation of the linkage error rates on the knowledge of the true linkage status for a subset of pairs (Gomatam et al. 2002; Ding and Fienberg, 1994). This error evaluation will not be affected by bias if the discriminating power of the matching variables is similar between the two sets of pairs, both the whole set of pair candidates for the links and that where the linkage status is known. Comparing the distinguishing power of the four matching variables between the subset of the training linkage and the whole set of pairs, it should be noticed that whereas "gender" and "date of death" show similar distribution in the two sets, with respect to "age" and "place of death", some differences arise due to factors connected with the collection data procedures (i.e. 4% of RTIIS-e records falls in the age class 0-4 years, against 14% of the RTIIS records; moreover 25% of RTIIS-e records falls in the Municipality of Rome, against 64% of the RTIIS records). In fact, only the State Police, working in specific Municipalities, actually give more detailed information in reporting road accidents; furthermore when identification documents are available it is possible to collect names and surnames, as well as the exact ages of the persons involved in the accidents. These facts suggest that the estimates of the errors could be affected by a hard to evaluate bias. Some different approaches are suggested in literature in order to obtain a more accurate measure of the linkage errors, i.e. starting from the parameter estimations (Belin and Rubin, 1985; Armstrong and Mayda,1993). Also when exploring this way, one should bear in mind that in this kind of context the weakness of the matching variables could also affect the parameter estimations, implying unknown bias; so, despite a robustness of the linkage rule with respect to the parameter estimation, the measure of accuracy could be unreliable. Other methods could be exploited, related to application of re-sampling schemes on the data to link (Liseo and Tancredi, 2004), that is the direction of our future efforts. All these methods aim at taking into account the effect of these errors when analysing linked data.
Finally, in this application it is worth noting that the probabilistic record linkage approach, when based on weak matching variables, allows us to obtain a number of matches three times higher than the number of matches accomplished through a deterministic linkage procedure. In other circumstances, further additional matches could be found through a clerical review of pairs with weights belonging within suitably chosen thresholds. However, linked data have to be treated carefully, considering the errors associated with the linking probabilistic procedures.

References

Peden M., Scurfield R., Sleet D., et al. (2004), World report on traffic injury prevention, World Health Organization, Geneva.

Piffer S. (2004), Surveillance and prevention road accidents, Center Control of Disease, Ministry of Health, http://www.ccm.ministerosalute.it

WHO (2002), The Injury Chartbook: A graphical overview of the global burden of injuries, World Health Organization, Geneva.

European Council. Communication from the Commission to the Council, the European Parliament, the Economic and Social Committee and the Committee of the Regions: Promoting road safety in the European Union: - the programme for 1997-2001. COM(97) 131 final. Not published in the Official Journal.

Valent F., Schiava F., Savonitto C., Gallo T., Brusaferro S., Barbone F. (2002), "Risk factors for fatal road traffic accidents in Udine, Italy", Accident Analysis and Prevention, 34, 71-84.

ISTAT-ACI. (2001), Statistica degli incidenti stradali. Anno 2000. Collana ISTAT Informazioni n. 38, Roma. (In Italian).

Cercarelli L.R., Rosman D.L., Ryan G.A. (1996), "Comparison of accident and emergency with police road injury data", Journal of Trauma, 40, 805-809.

Langley J.D. (1995), "Experiences using New Zeland's hospital based surveillance system for injury prevention research", Methods of Information in Medicine, 34, 340-344.

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

Sniezek J.E., Finklea J.F., Graitcer P.L. (1989), "Injury coding and hospital discharge data", JAMA, 262, 2270-2.

Belin T.R, Rubin D.B. (1985), "A Method for calibrating false-match rates in record linkage", Journal of the American Statistical Association, 90, 694-707.

Fellegi I.P., Sunter A.B. (1969), "A Theory for record linkage", Journal of the American Statistical Association, 64, 1183-1210.

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.

Dempster A.P., Laird N.H., Rubin D.B. (1977), "Maximum likelihood from incomplete data via the EM algorithm", Journal of the Royal Statistical Society, B 39, 1-38.

Armstrong J.A., Mayda J.E. (1993), "Model-based Estimation of Record Linkage Error Rates", Survey Methodology, 19, 137-147.

Hagenaars J.A. (1993), Loglinear Models With Latent Variables, Sage Publications Inc.

Winkler W.E. (1995), "Matching and Record Linkage", in Cox, Binder, Chinnappa, Christianson, Colledge, Kott , Business Survey Methods, Wiley, New York, 355-384.

Gu L., Baxter R., Vickers D., Rainsford C. (2003), Record Linkage: Current Practice and Future Directions, Technical Report No. 03/83, CSIRO Mathematical and Information Sciences, http://datamining.csiro.au

Gomatam S., Carter R., Ariet M., Mitchell G. (2002), "An Empirical Comparison of Record Linkage Procedures", Statistics in Medicine. 21, 1485-1496.

Ding Y., Fienberg S.E. (1994), "Dual system estimation of Census undercount in the presence of matching error", Survey Methodology, 20, 149-158.

Baxter R., Christen P., Churches T. (2003), "A Comparison of Fast Blocking Methods for Record Linkage", Proceedings of the ACM Workshop on Data Cleaning, Record Linkage and Object Identification, Washington DC.

Hernandez M.A., Stolfo S.J. (1998), "Real-world Data is Dirty: Data Cleansing and The Merge/Purge Problem", Journal of Data Mining and Knowledge Discovery, 1, 9-37.

Liseo B., Tancredi A. (2004), "Statistical inference for data files that are computer linked", Proceedings of the International Workshop on Statistical Modelling, Firenze Univ. Press, pp. 224-228.

12/2007 – Daniele Frongia, Raffaello Martinelli, Fernanda Panizon, Bruno Querini e Andrea Stanco – *Il nuovo Sistema informatico Altri Servizi. Progetto di reingegnerizzazione dei processi produttivi delle indagini trimestrali di fatturato degli altri servizi*

1/2008 – Carlo De Gregorio, Stefania Fatello, Rosanna Lo Conte, Stefano Mosca, Francesca Rossetti – *Sampling design and treatment of products in Istat centralised CPI surveys*

2/2008 – Mario Abissini, Elisa Marzilli e Federica Pintaldi – *Test cognitivo e utilizzo del questionario tradotto: sperimentazioni dell'indagine sulle forze di lavoro*

3/2008 – Franco Mostacci – *Gli aggiustamenti di qualità negli indici dei prezzi al consumo in Italia: metodi, casi di studio e indicatori impliciti*

4/2008 – Carlo Vaccari e Daniele Frongia – *Introduzione al Web 2.0 per la Statistica*

5/2008 – Antonio Cortese – *La conta degli stranieri: una bella sfida per il censimento demografico del 2011*

6/2008 – Carlo De Gregorio, Carmina Munzi e Paola Zavagnini – *Problemi di stima, effetti stagionali e politiche di prezzo in alcuni servizi di alloggio complementari: alcune evidenze dalle rilevazioni centralizzate dei prezzi al consumo*

7/2008 – AA.VV. – *Seminario: metodi per il controllo e la correzione dei dati nelle indagini sulle imprese: alcune esperienze nel settore delle statistiche strutturali*

8/2008 – Monica Montella – *La nuova matrice dei margini di trasporto*

9/2008 – Antonia Boggia, Marco Fortini, Matteo Mazziotta, Alessandro Pallara, Antonio Pavone, Federico Polidoro, Rosabel Ricci, Anna Maria Sgamba e Angela Seeber – *L'indagine conoscitiva della rete di rilevazione dei prezzi al consumo*

10/2008 – Marco Ballin e Giulio Barcaroli – *Optimal stratification of sampling frames in a multivariate and multidomain sample design*

11/2008 – Grazia Di Bella e Stefania Macchia – *Experimenting Data Capturing Techniques for Water Statistics*

12/2008 – Piero Demetrio Falorsi e Paolo Righi – *A Balanced Sampling Approach for Multi-way Stratification Designs for Small Area Estimation*

13/2008 – AA.VV. – *Seminario: Strategie e metodi per il controllo e la correzione dei dati nelle indagini sulle imprese: alcune esperienze nel settore delle statistiche congiunturali*

14/2008 – Francesco Chini, Marco Fortini, Tiziana Tuoto, Sara Farchi, Paolo Giorgi Rossi, Raffaella Amato e Piero Borgia – *Probabilistic Record Linkage for the Integrated Surveillance of Road Traffic Injuries when Personal Identifiers are Lacking*