Towards an integrated surveillance system of road accidents

Tiziana Tuoto, Silvia Bruzzone, Luca Valentino, Giordana Baldassarre, Nicoletta Cibella and Marilena Pappagallo[1]

Abstract The aim of this paper is to provide an integrated overview on data included in Causes of death and Road accidents Istat registers. 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. Data integration, performed by record linkage techniques, focuses on the extension of information associated to each road accident. In particular, information on the circumstances of traffic accident and characteristic of roads and vehicles are linked to medical information on causes of death at individual level.

1 Introduction

Road traffic injuries are the leading cause of death for young adults in industrialised countries. Injury prevention is one of the major challenges of the World Health Organization for both industrialized and developing countries. 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.

The available official data for road accidents are often linked to different data sources; the various nature of data highlights the difficulties in data comparability and use of information, often based on dissimilar classifications. Even if some differences in definitions, the goal of the paper is to join individual information on deaths due to road accidents, using the two Istat register on Causes of death and Road accidents data. This approach is based on record linkage; the aim is to provide a set of integrated information for each killed person in road

[1] Tiziana Tuoto, Istat, tiziana.tuoto@istat.it
Silvia Bruzzone, Istat silvia.bruzzone@istat.it
Name of Author, Istat, name.surname@istat.it

accidents: data on cause and manner of death and data on role of the deceased, dynamic and circumstances of the road accident. A first study, using only a deterministic approach (Amato et al. 2006), was carried out in 2006; the results obtained were not fully satisfactory but represented a first step to highlight critical points of the method and to support the adoption of probabilistic techniques. After the required experiments of this preliminary stage, with some results showed in section 4, the medium-time ambitious intent is to build up a permanent integrated system for the surveillance of road accident, eventually adding further sources of medical data.

2 The road accidents and causes of death Istat registers

The survey on road accidents resulting in deaths or injuries, carried out by the Italian National Institute of Statistics (Istat), is an exhaustive and monthly based data collection. The survey collects all road accidents involving at least a vehicle, circulating on the national road net, resulting in deaths or injuries and documented by a Police authority. The collection of information is done by members of the Italian military corps or Police for which the internal organisation is usually variable at local level. A flexible data flow model has been adopted by Istat, through the subscription of special agreements with regions and provinces, to facilitate the local authorities information needs and to improve the timeliness and quality of data collected.

Data on mortality by cause is annually collected, processed and published by Istat. Causes of death data is collected by the Istat certificate, according to WHO and Minister of Health recommendations. Therefore, for every death, socio-demographic variables and epidemiological information are available. The section of the death certificate treated by physician contains the complete sequence of diseases (whether fatal or non fatal) and, when it applies, the traumatic circumstances that have occurred to the individual before death. Nevertheless, data refer to the underlying cause of death, i.e. the one that has mostly contributed to death.

3 The Integration Procedure with formalization

In a context of increasing demand of statistical information with stricter budgetary constraints and the desire of limit the response burden, bringing together for statistical purpose huge amount of data coming from different sources is largely widespread. Record linkage techniques are a multidisciplinary set of methods and practices with the main purpose of accurately recognize the same real world entity at individual level, even when differently stored in sources of various type. The complexity of record linkage resides on several aspects, mainly related to lack and errors in personal identifiers. In the following, a very short view to the classical theory, due to Fellegi and Sunter (1969) is summarised.

Given two data sets A and B of size N_A and N_B respectively, let us consider $\Omega=\{(a,b), a\in A$ and $b\in B\}$ of size $N=N_A\times N_B$. The linkage between A and B can be defined as the problem of classifying the pairs that belong to Ω in two subsets M and U independent and mutually exclusive, such that: M is the set of matches ($a=b$) and U is the set of non-matches ($a\neq b$). In order to classify the pairs, K common identifiers (matching variables) have to be chosen so

that, for each pairs, a comparison function is applied and a comparison vector $\gamma = \{\gamma_1, \gamma_2, ..., \gamma_K\}$ is obtained.

Following Fellegi and Sunter (1969), the ratio

$$r = \frac{P(\gamma | (a,b) \in M)}{P(\gamma | (a,b) \in U)} = \frac{m(\gamma)}{u(\gamma)}$$

between the probabilities of γ given the pair (a,b) membership either to the subset *M* or *U* is used so as classifying the pair. In practice, once the probabilities *m* and *u* are estimated, for instance by means of the EM algorithm, all the pairs can be ranked according to their ratio *r* and a classification criterion based on two thresholds T_m and T_u ($T_m > T_u$) is applied. More precisely, those pairs for which *r* is greater than T_m can be considered as linked; those pairs for which *r* is smaller than T_u can be considered as not-linked, if *r* falls in the range (T_m , T_u) no-decision is made and the pair is held out for the clerical review so to be solved. The thresholds are chosen so to minimize false match rate and false non-match rate. The Fellegi and Sunter approach is heavily dependent on the accuracy of $m(\gamma)$ and $u(\gamma)$ estimates.

4 First Results

For sake of simplicity, in the first attempt of building an integrated system on road accident, we focus on the Toscana region and the 2008 year. As far as the data from road traffic accident are concerned, only records with at least one dead person are considered. This selection corresponds to 291 records. The variables useful for the linkage purpose are: name, surname, gender and age of the victim, day, month, municipality, province of the accident. Related to data from causes of death register, at this first step only deaths due to road accidents are considered, according to ICD-10 codes for motor vehicle traffic accidents on public roads. This corresponds to 321 records. The variables selected for the linkage purpose are: name, surname, gender and age of the dead person, day, month, municipality, province of the death.

The first main difficulty to face in linking data from road accidents and from causes of death register regards the difference in the reference units, that is the accident for the former set and the single person for the latter one. This fact is mainly influential when an accident involves more than one person (that occurs 13 times, corresponding to 27 records), because from the road accidents data is not possible to put together, at individual level, variables "name and surname" with variables "age and gender". It is easy to guess that, while the whole set of variables has a high identification power, the use of only a subset causes a serious loss. At this stage, the variables "name and surname" have been preferred when an accident involves more than one dead person. Moreover, the variables name and surname are missing for 34 records of the road accidents register, and in 2 records name, surname, age and gender are simultaneously missing.

In order to successfully apply probabilistic linkage methods, new linking variables are generated starting from the former ones. In particular, the "date" new variable is the concatenation of day and month of the event (accident or death), the "place" variable is the concatenation of municipality and province of the event, the "name and surname" and "age and gender" concatenations are considered as well. Note that variables "year" and "region" are equal for all records due to the starting selection; in fact, they can be considered as blocking variables in a standard search space reduction procedure. The size of the selected

data sources do not requires further reduction procedures, and the cross product of all records can be considered.

Probabilistic model has been applied. Several attempts have been tried, corresponding to different models. In the following, for seek of brevity, only the best is described, in terms of identified matches and estimated errors. All the new four variables have been used as linking keys. For the comparison between the values of "name and surname" a distance function, based on the Jaro string comparator, has been considered. Prior error rates are set in order to accept as matches those pairs with posterior linking probability greater or equal to 0,95 and to refuse as non-matches those pairs with posterior linking probability smaller or equal to 0,50. The described linkage process identifies 189 pairs as Matches and 14 pairs as Possible-Matches. A clerical review of the Possible-Matches suggests to accept 13 of them as Matches and to reject one of them. So, the whole linkage result proposes 202 matches. The associate errors, estimated from the model, are 0,054 and 0,002 for the false non-match rate and false match rate, respectively.

5 Concluding remarks and future works

The experiment described in the previous session has been performed via the RELAIS software. RELAIS is configured as an open source project with the aims of facing the record linkage complexity by decomposing the whole problem in its constituting phases and dynamically adopting the most appropriate technique for each step. In this way, it is possible to define for each project the most suitable strategy depending on application and data specific requirements (Cibella et al. 2010). The methodological core of RELAIS is based on the Fellegi-Sunter theory, allowing its usage by both researchers and non-experts.

The first results described in the previous session encourage to continue in this way in order to build an integrate system to explain the complex world of road traffic accidents and their consequences. As a matter of fact, further analyses are needed before to extend the results to the whole set of data. The first natural benchmark in this kind of experiments is the result of the deterministic linkage by the strong keys and, in this sense, probabilistic procedures overcomes that performances. Anyway, good results in terms of matching errors are still ensured by the relative small number of handled records, so, some cautions have to be applied when the whole set of data are treated. Moreover, further step of the linkage procedure could take into account the processing of those records that don't match within a single region or a single year, removing the blocks, in order to recover residual matches.

Finally, it is important to underline the role of the set of Possible matches: it allows to identify a limited number of pairs for the clerical review, ensuring a significant increase of the whole quality of the result, even in time of strict budgetary constraints.

References

1. Amato R. , Bruzzone S., Del monte V., Fagiolo L. (2006). "Le statistiche sociali dell'ISTAT e il fenomeno degli incidenti stradali: un'esperienza di record linkage", Istat Contributi n. 4
2. Fellegi, I.P., A.B. Sunter, (1969). "A theory for record linkage", *JASA*, Volume 64.
3. Cibella N., Fortini M., Ichim D., Tuoto T. (2010) "Record linkage methods and techniques as proposed in RELAIS", in Proceedings of International Methodology Symposium 2010, Statistics Canada, 26-29 October, Ottawa, Canada.
4. Jaro, M.A. (1989), "Advances in record-linkage methodology as applied to matching the 1985 Census of Tampa, Florida", *JASA*, Volume 84.

5. Relais 2.2. User Guide, Istat, at http://www.istat.it/strumenti/metodi/software/analisi_dati/relais/