

Probabilistic Record Linkage: Relationship Between Parameter Estimations and Link Error Risk

Record linkage probabilistico: relazione tra la stima dei parametri e il rischio di errato abbinamento

N. Cibella, M. Fortini, T. Tuoto
ISTAT, tuoto@istat.it

Riassunto: Le tecniche di record linkage hanno suscitato l'interesse della comunità scientifica in quanto permettono di integrare informazioni provenienti da diverse fonti di dati al fine di migliorarne la qualità, sfruttando al meglio l'informazione già disponibile. Una procedura di record linkage deve essere corredata da un criterio che permetta di valutare la bontà della procedura stessa così che sia possibile minimizzare la probabilità di errato abbinamento. In questo studio, nel contesto delle procedure di record linkage probabilistico, si testa l'impatto che la scelta di parametri calcolati, attraverso l'algoritmo EM, su insiemi differenti di unità, ha sul rischio di errato abbinamento. I dati su cui viene effettuato lo studio, seguendo un approccio simulativo, sono quelli dell'indagine di Copertura del XIV Censimento Generale della Popolazione.

Keywords: Data Integration, Record Linkage, Matching Errors.

1. Introduction

In performing a record linkage technique is important to assess the quality of the procedures implemented and testing how this quality is related to the choices adopted. Moreover, also a criterion for evaluating the false match and the false non match rates need to be establish. In the context of probabilistic approach, the aim of this paper is to test alternatives methods to estimate the parameters of the procedure via the EM algorithm (Jaro, 1989). Actually, the parameters can be calculated on the whole set of data or they can be estimated on a sample of the original data in order to distinguish between the sets of matches and non-matches. The simulation study is performed on the Coverage Survey data of the XIV Italian Population Census

2. Study context and methods

In this paper we refer to the XIV Population Census Coverage Survey data. Actually, to estimate the unknown number of Italian Population at Census time, a dual system estimation model was implemented. In order to establish the number of record counted both by the census and the survey, complex record linkage techniques were performed, partly deterministic, partly probabilistic and partly manual (see Fortini et al, 2005). The linkage involved about 170.000 records, matched in a one-to-one process. Due to the accuracy of those procedures, it's possible to consider those matches as the "true

matches” in order to evaluate the quality of alternatives procedures in a simulation context.

Due to very strong constrain on the accuracy of the coverage results, the record linkage procedures actually implemented were very time-consuming. In this work we want to evaluate some probabilistic approaches that allow to save dramatically time, paying a cost in terms of accuracy.

The linkage approach here considered is the traditional Fellegi-Sunter approach (Fellegi and Sunter, 1969), in which within the set Ω of all pairs, where $\Omega = \{(a, b); a \in \text{Census}, b \in \text{Survey}\}$, the link/not link decision was based on weights w depending on the result of a comparison among some key variables (namely name, surname, gender, month and year of birth). The parameters, which the weights w are based on, are estimated on the whole set Ω , or alternatively on a subset of Ω to better achieve a distinguishing power of some variables. The traditional way to obtain better identifications consists in reducing the search space of parameters through the introduction of some blocking criteria. The basic idea of this work, starting from Yancey 2004, is to achieve better parameter estimates considering different subsets of Ω , obtained not only via blocking criteria but also via sampling of suitable subsets. The results are compared in terms of matching proportion, false match and false non-match rates; these errors are calculated with different methods, firstly starting from the known true matches status, but also through functions of the parameters themselves (Belin and Rubin, 1995; Torelli and Paggiaro, 1999).

References

- Armstrong, J. A., Mayda, J. E. (1993) Model-based Estimation of Record Linkage Error Rates, *Survey Methodology*, 19, 137-147.
- Belin, T.R., Rubin, D.B. (1995) A method for calibrating false-match rate 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.
- Fortini M., Carbonetti G., Mignoli N., Nuccitelli A. (2005) L'indagine sul grado di copertura del 14° censimento della popolazione: considerazioni di carattere metodologico e primi risultati. *Giornate di Studio sulla Popolazione*, Padova.
- 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.
- Torelli, N., Paggiaro, A. (1999) La Stima della Quota di Errori in Procedure di Abbinamento Esatto. *Atti del Convegno SIS 99: Verso i censimenti del 2000*.
- Yancey W.E. (2004) Improving EM Algorithm Estimates for Record Linkage Parameters, *Research Report Series (Statistics #2004-01)* U.S. Bureau of the Census.