Probabilistic record linkage with less than three matching variables

Record linkage probabilistico con meno di tre variabili di confronto

Tiziana Tuoto and Marco Fortini

Abstract Probabilistic record linkage based on Fellegi-Sunter theory is a methodology for integrating data collected in different sources when a unique common identifier is not available. It requires at least three matching variables are available to identify the probability model. In official statistics, it is emerging the need to join archives even with less than three common variables, this is the case for instance of addresses and business archives of poor quality. For this problem, we compare available common variables by means of string comparators and propose mixtures of continuous and categorical distributions rather than usual the latent class models to estimate linkage probabilities.

Abstract *Il record linkage probabilistico basato sulla teoria di Fellegi-Sunter è una metodologia per integrare dati raccolti in fonti diverse quando non è disponibile un codice identificativo comune univoco. In questo caso, sono necessarie almeno tre variabili di confronto per identificare il modello di probabilità. Nella statistica ufficiale, emerge la necessità di abbinare archivi anche quando sono disponibili meno di tre variabili in comune, come ad esempio nel caso di archivi di indirizzi o di imprese di scarsa qualità. Per questo problema, confrontiamo le variaibli disponibili mediante comparatori di stringhe e proponiamo misture di distribuzioni continue e categoriche piuttosto che i modelli a classi latenti solitamente utilizzati per la stima delle probabilità di abbinamento.*

Key words: Fellegi-Sunter record linkage, mixture models, string metrics

1 Introduction

[1] Marco Fortini, Istat; email: fortini@istat.it
Tiziana Tuoto, Istat; email: tuoto@istat.it

Data linkage is a common practice in National Statistical Offices and in many other Institutions to enlarge and enrich the availability of information without incurring the costs of new surveys and burden to respondents. Nowadays in many National Statistical Institutes a new statistical production system has been established, mainly based on integrated datasets, including both administrative archives and traditional sample surveys. This new statistical production system has been by the large availability of administrative data, often with unique identifiers for the units of interest, and almost unlimited computing and storage capacity. Integrated datasets provide complementary variables for the same units; they make it possible to discover relationships between different types of units (e.g. households and enterprises, households and schools) and to study the changes over time of the units and the variables.

When units unique identifiers are not available or corrupted, e.g. for privacy reasons or for lack of quality in the data sources, the data linkage is a not trivial task. The most widespread methodology to face linkage issues is the probabilistic record linkage. Probabilistic Record Linkage, according to the theory by Fellegi and Sunter (1969) and the implementation proposed by Jaro (1989) requires at least three matching variable to identify the probability model. This minimum number of common variables is easily available when the reference units are people (e.g. names, surnames, date and place of birth, gender). Unfortunately, this is not the case when integration is needed between other reference units, such as addresses or businesses. In this paper, we propose a model for probabilistic record linkage that uses less than three matching variables. The method is applied to link data from the Palestinian Business Census and an administrative business register. The behaviour of the proposed model is discussed by means of a simulation study.

2 Mixture models for probabilistic record linkage

In this section we shortly recall the well-known probability model for record linkage as proposed by Jaro (1989), that requires at least three matching variables are available, to move to the description of our proposal, that allows probabilistic record linkage with only two matching variables. Both models rely on mixture models.

2.1 Probabilistic record linkage

The goal of record linkage is to recognize records referred to the same unit even when this is differently represented in different sources. To fix the idea, let us consider two data sources, file A and file B, of size N_A and N_B, respectively. The whole comparisons between records *(a,b)* from A and B generate a comparison pairs space Ω of size $N_\Omega = N_A \times N_B$. The goal of linkage procedure is to identify in Ω two disjoint sets M and U such that $\Omega = M \cup U$ and $M \cap U = \emptyset$, where M is the set of Matches, i.e. $\omega = (a,b)$ represents the same unit, $a=b$; while U is the set of Non-matches, i.e.

$\omega = (a,b)$ refer to two different units. $a \neq b$. The pairs assignment to the sets M and U is determined on the basis of K common matching variables, and the comparison vector $\boldsymbol{\gamma} = \{\gamma_1, \gamma_2, \dots, \gamma_K\}$ reporting the agreement/disagreement between the matching variables. When the comparison on the matching variables admits dichotomous outcome, the vector $\boldsymbol{\gamma}$ assumes 2^K possible patterns. Jaro (1989) firstly approaches the record linkage problem by means of mixture models with latent variables. The not observed variable representing the real matching status is the latent one, to be predicted by observing the results of the comparisons vector $\boldsymbol{\gamma} = \{\gamma_1, \gamma_2, \dots, \gamma_K\}$ on the K observed matching variables.

Probabilistic record linkage models the 2^K observed frequencies of pairs of the comparison vector $\boldsymbol{\gamma}$ as a mixture coming from two different distributions, $p_{\boldsymbol{\gamma}} = p \cdot m_{\boldsymbol{\gamma}} + (1-p) \cdot u_{\boldsymbol{\gamma}}$, the distribution of the comparisons $\boldsymbol{\gamma}$ given that the pairs belong to the population of Matches, i.e. $m_{\boldsymbol{\gamma}} = P(\boldsymbol{\gamma} \mid M)$, and the distribution of comparisons $\boldsymbol{\gamma}$ given that the pairs belong to the population of Non-Matches, $u_{\boldsymbol{\gamma}} = P(\boldsymbol{\gamma} \mid U)$. The probability of the set Matches over the comparison space Ω, $p = P(M)$, is the weight of the mixture.

We aim to estimate of the probability of a pair to be a match among those showing the pattern $\boldsymbol{\gamma}$: $\pi_{\boldsymbol{\gamma}} = P(\omega \in \mathrm{M} \mid \boldsymbol{\gamma}), \forall \boldsymbol{\gamma}$. To solve this problem, Jaro (1989) suggests applying the EM algorithm: with at least three variables and under the conditional independence assumption, we obtain estimated distributions $\hat{m}(\gamma)$ and $\hat{u}(\gamma)$. Hence, applying the Bayes rule we can evaluate the probability of a pair to be a match given its pattern $\boldsymbol{\gamma}$:

$$\pi_{\boldsymbol{\gamma}} = (p \cdot m_{\boldsymbol{\gamma}}) / (p \cdot m_{\boldsymbol{\gamma}} + (1-p) \cdot u_{\boldsymbol{\gamma}}).$$

The most likely pairs to be Matches are those with values of $\pi_{\boldsymbol{\gamma}}$ close to 1.

To be identifiable, the model needs at least three matching variables, in this way we have $2^3 = 8$ observed frequencies and need to estimate 7 parameters $p, m_1, m_2, m_3, u_1, u_2, u_3$

2.2 Mixture of Beta and Bernoulli distribution for record linkage

In some real cases, some of them described in the following paragraph, the matching variables are less than three, preventing the application of the previous models. However, it is quite common the matching variables are strings of characters and are compared via string comparators that result in [0, 1] intervals rather than dichotomous outcomes. The most common and widespread string comparators for names are Levenstein, Jaro, Jaro-Winkler, qgram, Jaccard. In principle, to obtain comparison outcome in the range [0,1] we don't need to constrain ourselves to string variables, but the same reasoning can be applied to numeric variables, adopting the most convenient comparison function.

When only two matching variables are available, we propose the following model for record linkage: for each pair $\omega \in \Omega$ we consider a first variable with outcome $\delta_\omega \in [0,1]$ and the other variable with dichotomous outcome $\gamma_\omega \in \{0,1\}$.

The joint probability of observing δ_ω and φ_ω can be still modeled as a mixture of Matches and Non-matches distributions:

$$P(\delta_\omega, \varphi_\omega) = P(\omega \in M)P(\delta_\omega, \varphi_\omega|\omega \in M) + \big(1 - P(\omega \in M)\big)P(\delta_\omega, \varphi_\omega|\omega \in U).$$

Under the usual conditional independence assumption, we can factorize the joint observed probability:

$$P(\delta_\omega, \varphi_\omega) = P(\omega \in M)P(\delta_\omega|\omega \in M)P(\varphi_\omega|\omega \in M)$$

$$+ \big(1 - P(\omega \in M)\big)P(\delta_\omega|\omega \in U)P(\varphi_\omega|\omega \in M).$$

where $P(\omega \in M) = p; \quad P(\delta_\omega|M) = m_\delta(\omega); \quad P(\varphi_\omega|M) = m_\varphi(\omega);$

$$P(\delta_\omega|U) = u_\delta(\omega); \quad P(\varphi_\omega|U) = u_\varphi(\omega).$$

Let us assume the following probability distributions:

- $m_\delta(\omega) = Beta(\delta_\omega; \alpha_M, \beta_M), \quad \alpha_M, \beta_M > 0;$
- $u_\delta(\omega) = Beta(\delta_\omega; \alpha_U, \beta_U), \quad \alpha_U, \beta_U > 0;$
- $m_\varphi(\omega) = Bernoulli(\ \ _\omega; m_\varphi), \quad m_\varphi \in [0,1];$
- $u_\varphi(\omega) = Bernoulli(\varphi_\omega; u_\varphi), \quad u_\varphi \in [0,1].$

This allow us to write the compete likelihood, which includes also the latent variable, which can be factorized for the parameters $p, m_\varphi, u_\varphi, \alpha_M, \beta_M, \alpha_U, \beta_U$ and solved via an EM algorithm. Dealing with Beta distributions, the EM algorithm requires the solution of a system of partial derivatives involving non linear equations, its description and the related algebra can be provided in Appendix.

3 An application to real data and a simulation

The method proposed above might be applied when less than three matching variables are available, as it is the case, e.g. in the linkage of some residual addresses in the Italian Statistical Addresses Register. In this section, we propose a real-case application to the first version of the Palestinian Statistical Business Register, built through the linkage of several statistical and administrative registers. Moreover, to understand the behaviour of the proposed modelling, we show the results of a simulation with fictitious data where the linkage status is known.

Among others, the Palestinian Statistical Business Register links the 2017 Palestinian Business Census, managed by the Palestinian Central Bureau of Statistics PCBS and the Municipality Business Archive, managed by each Municipality for administrative reasons, as e.g. the delivery of services such as water, electricity, etc. For the municipality of Salfit, the census counts 662 establishments and the Salfit municipality archive reports 394 establishments. The data sets have some common

variables, i.e. owner name, kind of activity, address; unfortunately the kind of activity is not coded in the same way, preventing the comparison of the reported information, and the address are often missing or registered in an incomparable way, referring to rural areas. This implies the only available matching variable is the owner name. This info from the Salfit municipality archive can be compared to both the owner name and the commercial name in Census. In this exercise, we model the linkage process as follows: the Jaccard distance between owner names in both sources is modelled as a mixture of two Beta for M and U, the Jaro-Winker distance between owner name and commercial name is dichotomised and modelled as a Bernoulli. The proposed model identifies 198 matches, with $\pi_{\gamma} > 0.5$, whilst the deterministic linkage performed at PCBS mainly via manual inspection identifies 171 matches. Manual check to evaluate the goodness of the additional matches is not trivial, due to the Arabic language of the reported information.

To facilitate the performance evaluation of the proposed method a simulation is performed, using public synthetic data for which the true match status is known, i.e the dataset created for the ESSnet Data Integration project (Essnet DI, 2011). The database consists of over 26000 records, with matching variables such as names, dates of birth and addresses. The matching variables contain simulated missing values and typos, mimicking those encountered in reality. From this database, 100 samples of size 1000 are independently selected, each by simple random sampling without replacement. From each sample of 1000 units, two files A and B are independently created by Bernoulli sampling with probability of selection $p_A = 0.93$ and $p_B = 0.92$, respectively, i.e. the two files to be linked are of sizes $N_A = 930$ and $N_B = 920$ on average over the 100 replications, and the number of true matches between them is 858 on average.

Three linkage models are compared: the first (*mod1*) models the Jaccard distance on *Surname* with Beta distribution and Bernoulli distribution for the matching variable *Year of Birth*. In the second model (*mod2*) we create a new variable pasting "*Surname*" and "*Name*" and model the Jaccard distance on the *SurnameName* variable with Beta distribution and again Bernoulli distribution for the matching variable *Year of Birth*. This second model aims at introducing more information into the linkage, so to apply a standard procedure in the third model (*mod3*) where we consider the traditional Fellegi-Sunter model based on the three variables, *Name*, *Surname* and *Year of Birth*. It's worthwhile noting that we introduce a constraint on the Beta distribution for *mod2*, i.e. we fix the parameters of the Beta distribution for the M set to be $\alpha_M = \beta_M = 0.5$, in order to evaluate the modelling adequacy of the Beta distribution for linkage purpose, comparing results from *mod1* and *mod2*. Some results from the simulation are shown in table 1, where match rate and false match rate are reported, averaging over the 100 simulation. The simulation results allow for deeper analysis, presented in Appendix.

Table 1: *Results from the simulation*

Linkage Model	***Match rate***	***False Match rate***
Mod1	0.77	0.28
Mod2	0.82	0.05

Mod3	0.88	0.02

4 Concluding remarks

The proposed method seems a valid alternative to judgmental deterministic record linkage in situations with little information, when less than three matching variables are available.

The methodology can be extended in several ways, based on the specific features of the real data. A possible extension is presented in the simulation, where we compare the results of modelling Beta distributions where all the parameters vary in the parameter space, with Beta distributions whit the parameters constrained to fixed values. Other possible extensions include modelling a mixture of two Beta distributions, as well as extending the mixture of Beta and Bernoulli distributions to the comparison of three (or more) matching variables. To this extent, the simulation provides some insights yet. In the proposed setting, the standard Fellegi-Sunter approach seems exploiting the identification power of the matching variables in a way that over-performing the Beta-Bernoulli mixture, in terms of both match rate and false match rate. Obviously, to some extent, this is due to the use of three variables instead of two. Actually, whether the proposed modelling might substitute the standard approach based on multinomial distribution is still an open question and need further analysis. The intent of this contribution is not to provide an alternative of the Jaro implementation of the Fellegi Sunter model, but rather to provide a probabilistic solution in cases where the standard model is not applicable at all.

References

1. Fellegi, I.P., Sunter, A.B. (1969), A Theory for Record Linkage, Journal of the American Statistical Association, 64, pp. 1183-1210
2. 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.
3. Essnet DI - McLeod, Heasman and Forbes, (2011) Simulated data for the on the job training, http://www.cros-portal.eu/content/job-training