

Bayesian Point Estimators for Record Linkage

Stimatori Puntuali Bayesiani per il Problema dell'Abbinamento Esatto

Alessandra Nuccitelli
Istituto Nazionale di Statistica, Roma
e-mail: nuccitel@istat.it

Riassunto: con il termine ‘abbinamento esatto’ si intende l’uso di tecniche algoritmiche per identificare coppie di registrazioni contenute in archivi diversi e relative alla stessa unità statistica. In questo lavoro si discute l’utilizzo di alcuni possibili stimatori puntuali del parametro oggetto di interesse in un’impostazione bayesiana del problema.

Keywords: Exact matching, Posterior probability of matching, Decision rules

1. Introduction

Record linkage refers to the use of an algorithmic technique to identify pairs of records (a, b) , a from file **A** and b from file **B**, that correspond to the same statistical unit (e.g. individual, business, dwelling). This technique can be very useful and cost-efficient in the design, production, analysis and evaluation of statistical data. Clerks can review pairs of records, but that is time consuming and costly. Given a decision rule, computers can quickly designate each pair as a link or a non-link. Decision rules are based on the comparison between the common fields (or matching variables) of the two data files. Unfortunately, matching variables are likely to be observed with errors and that makes the linkage process not trivial. An overview of methods and systems for record linkage is in Winkler (1999). The fully Bayesian approach proposed in Fortini *et al.* (2001) is particularly suitable for the solution of the following problems: 1) exact computation of the probability that each pair is a match, conditionally on the observed data (the comparisons between the matching variables); 2) computation of conditional probabilities that several pairs are simultaneously matches. This second point represents an improvement on the classical methods, where decision rules establish separately for each pair if records correspond to the same unit or not, without considering the compatibility constraints. In this approach a matrix **C**, which indicates the pattern of matches in the two files, is considered as the parameter of interest. For the application presented in this work, based on simulated data, the knowledge on truth and falsehood is used to evaluate the performance of the posterior mode in comparison with alternative inferential summaries defined in a decision theoretic context by introducing two different loss functions (a comparative study between the posterior mode and the classical decision rule implemented in *Record Linkage Software*, developed by U. S. Bureau of the Census, is in Nuccitelli, 2001).

2. Bayesian decision rules

The configuration matrix $\mathbf{C} = \{C_{ab}\}$ can take values on the set **C** of all the matrices whose elements satisfy the following obvious compatibility constraints:

$$\sum_{b=1} C_{ab} \leq 1 \quad a = 1, 2, \dots, n_A, \quad \sum_{a=1} C_{ab} \leq 1 \quad b = 1, 2, \dots, n_B,$$

where C_{ab} is the indicator function of the membership of the pair (a,b) to the set of matches and n_A and n_B indicate the number of records respectively in files **A** and **B**. Once the data $\mathbf{y}$ (the comparisons between the matching variables) have been observed, the posterior distribution $P(\mathbf{C} = \mathbf{c}|\mathbf{y})$ contains all the current information about $\mathbf{C}$.

Unfortunately, the number of matrices over which the posterior is defined is factorially large and there is not a natural ordering among their possible values. The analytic use of $P(\mathbf{C} = \mathbf{c}|\mathbf{y})$ is practically impossible but a sample from the posterior can be generated by using a Metropolis-Hastings algorithm.

The posterior mode seems to be an appealing estimator of $\mathbf{C}$, although it could suffer from a certain sensitivity to prior inputs¹.

As measure of performance of a decision rule, the False Match Rate (*FMR*), defined as the number of false declared matches divided by the total number of declared matches, can be adopted.

A more formal Bayesian approach implies the use of a loss function to produce a final decision.

Let $\mathbf{c}^v$ be the true configuration matrix and let $I_D(\cdot)$ be the indicator function of the set $\mathbf{D}$. If we adopt the following loss function²

$$L(\mathbf{c}, \mathbf{c}^v) = FMR + FUR = \frac{\sum_{a,b} c_{ab} I_{\{c_{ab}^v=0\}}(c_{ab}^v)}{\sum_{a,b} c_{ab}} + \frac{\sum_{a,b} (1-c_{ab}) I_{\{c_{ab}^v=1\}}(c_{ab}^v)}{n_A \times n_B - \sum_{a,b} c_{ab}},$$

it can be shown that the posterior expected loss associated to the decision $\mathbf{c}$ is

$$\begin{aligned} W_L(\mathbf{c}) &= \frac{\sum_{a,b} c_{ab} P(C_{ab} = 0|\mathbf{y})}{\sum_{a,b} c_{ab}} + \frac{\sum_{a,b} (1-c_{ab}) P(C_{ab} = 1|\mathbf{y})}{n_A \times n_B - \sum_{a,b} c_{ab}} = \\ &= \frac{\sum_{a,b} c_{ab} - \sum_{a,b} c_{ab} P(C_{ab} = 1|\mathbf{y})}{\sum_{a,b} c_{ab}} + \frac{\sum_{a,b} P(C_{ab} = 1|\mathbf{y}) - \sum_{a,b} c_{ab} P(C_{ab} = 1|\mathbf{y})}{n_A \times n_B - \sum_{a,b} c_{ab}}. \end{aligned} \quad (1)$$

Another interesting loss function is the squared error loss function $L_{SE}(\mathbf{c}, \mathbf{c}^v) = \sum_{a,b} (c_{ab}^v - c_{ab})^2$.

Its posterior expected loss $W_{L_{SE}}(\mathbf{c})$ corresponds to the sum of the numerators in (1).

The minimisation of $W_L(\mathbf{c})$ (or $W_{L_{SE}}(\mathbf{c})$) can be faced by an algorithm able to discover for each h , $h = 0, 1, \dots, n_A \wedge n_B$, the partial optimal solution in the subset $\mathbf{C}^h = \{\mathbf{c} \in \mathbf{C} : \sum_{a,b} c_{ab} = h\}$. For $h = 0$ the solution is trivial; for fixed h , $h \geq 1$, this

⁽¹⁾ In the application presented, in order to deal with the mode-finding problem, a simulated annealing algorithm was applied (Kirkpatrick *et al.*, 1983).

⁽²⁾ *FUR* is the False Un-match Rate, defined as the number of false declared non-matches divided by the total number of declared non-matches.

problem can be formulated in terms of the following programming problem: find the c_{ab} such that $\sum_{a,b} c_{ab} P(C_{a,b} = 1 | \mathbf{y})$ is maximum, subject to the constraints

$$\begin{aligned} \sum_{b=1} c_{ab} &\leq 1 \quad a = 1, 2, \dots, n_A; \quad \sum_{a=1} c_{ab} \leq 1 \quad b = 1, 2, \dots, n_B; \\ \sum_{a,b} c_{ab} &= h; \\ c_{ab} &\geq 0 \quad a = 1, 2, \dots, n_A, \quad b = 1, 2, \dots, n_B. \end{aligned} \quad (2)$$

An interesting feature of this problem (for fixed h) is that its optimal solution, obtained by using a continuous linear programming method as the simplex method (Dantzig, 1963), is integer. Precisely, it can be shown that the coefficient matrix of the linear constraints (2) is totally unimodular³ and that, consequently, every basic feasible solution of this problem is an integer solution.

Then, denoting with $\mathbf{c}^h$ the configuration matrix with lowest posterior expected loss in $\mathbf{C}^h$, the global optimal solution is $\mathbf{c}^* = \arg \min \{W_L(\mathbf{c}^h), h = 0, 1, \dots, n_A \wedge n_B\}$ (or $\mathbf{c}^* = \arg \min \{W_{LSE}(\mathbf{c}^h), h = 0, 1, \dots, n_A \wedge n_B\}$).

3. Application to simulated data

The following application is based on individual data collected in October 1998 in the context of the Test for the 2001 Italian Census of Population and Housing. Matching variables are *sex*, *day of birth*, *month of birth*, *year of birth*, *marital status*, *occupational status*, *relationship to head of household*, and *highest educational qualification*. Two data files with a known number of matches were perturbed by introducing errors at various extents according to a ‘‘Completely at Random’’ mechanism. Two different error rates were used (respectively in columns 2 and 3 of Table 1); they represent different quality standards you can expect for the matching variables. The input characteristics of the 8 distinct scenarios considered are presented in Table 2: f indicates the fraction of

Table 1: Total error rates introduced in each file

Matching variables	Low degree of distortion	High degree of distortion
<i>sex</i>	0.02	0.04
<i>day of birth</i>	0.04	0.08
<i>month of birth</i>	0.03	0.06
<i>year of birth</i>	0.03	0.06
<i>marital status</i>	0.06	0.12
<i>occupational status</i>	0.24	0.48
<i>relationship to head of household</i>	0.04	0.08
<i>highest educational qualification</i>	0.22	0.44

individuals in the two data files which are actually matches; x is the prior expected fraction of individuals which are matches. In simulations two different values of x were

(³) A matrix $\mathbf{m}$ is called totally unimodular if each square submatrix of $\mathbf{m}$ has determinant 0 or ± 1 .

considered, namely $x_1 = 0.6$ and $x_2 = 0.9$. Results on the goodness of the estimates are shown in Table 3: *CMR* indicates the Correct Match Rate, defined as the number of true declared matches divided by the number of matches; E and E_{SE} respectively denote the point of minimum of $W_L(\mathbf{c})$ and $W_{LSE}(\mathbf{c})$.

E appears an useless estimator: in all eight cases $FMR = 0$ but the *CMR* is far too low. Probably the *FUR* is not ‘well’ defined: the second denominator in $L(\mathbf{c}, \mathbf{c}^v)$ is very high compared to the first, so that the *FMR* completely overrides the *FUR*.

Results for the posterior mode and E_{SE} are encouraging in spite of a certain sensitivity to prior input when the error rates are larger and f is lower.

Finally, E_{SE} presents a better performance in terms of *FMR* in comparison with the posterior mode.

Table 2: *Input parameters*

Input parameters	case 1	case 2	case 3	case 4	case 5	case 6	case 7	case 8
degree of distortion	low	high	low	high	low	high	low	high
f	0.73	0.73	0.93	0.93	0.73	0.73	0.93	0.93
x	0.60	0.60	0.60	0.60	0.90	0.90	0.90	0.90

Table 3: *Output parameters for each point estimator*

Estimator	Output parameters	case 1	case 2	case 3	case 4	case 5	case 6	case 7	case 8
posterior mode	<i>CMR</i>	1.00	0.45	0.98	0.94	0.97	0.95	1.00	0.97
	<i>FMR</i>	0.02	0.00	0.00	0.02	0.19	0.17	0.02	0.09
E	<i>CMR</i>	0.02	0.02	0.02	0.02	0.02	0.05	0.84	0.52
	<i>FMR</i>	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
E_{SE}	<i>CMR</i>	0.97	0.37	0.97	0.63	1.00	0.90	0.98	0.98
	<i>FMR</i>	0.00	0.00	0.00	0.02	0.07	0.05	0.02	0.02

References

- Dantzig, G. B. (1963) *Linear Programming and Extensions*, Princeton University Press, Princeton.
- Fortini M., Liseo B., Nuccitelli A., Scanu, M. (2001) On Bayesian Record Linkage, *Research in Official Statistics*, 4, 1, 185-198 .
- Kirkpatrick S., Gelatt C. D., Vecchi, M. P. (1983) Optimization by Simulated Annealing, *Science*, 220, 671-680.
- Nuccitelli A. (2001) *Integrazione di Dati mediante Tecniche di Abbinamento Esatto: una Rassegna Critica ed una Proposta in Ambito Bayesiano*, Tesi di dottorato in *Metodi Statistici per l'Economia e per l'Impresa*, Università degli Studi di Roma Tre, Roma.
- Winkler W. E. (1999) *The State of Record Linkage and Current Research Problems*, Reserch Report RR99/04, U. S. Bureau of the Census, Washington.