

# Bayesian Inference with Linked Data

## *Inferenza Bayesiana con dati abbinati*

Andrea Tancredi and Brunero Liseo

**Abstract** In this paper we will describe some advances on a Bayesian methodology for performing record linkage and making inference using the resulting matched units. In particular, we frame the record linkage process into a formal statistical model which comprises both the matching variables and the other variables included at the inferential stage. This way, the researcher is able to account for the matching process uncertainty in inferential procedures based on probabilistically linked data, and at the same time, he/she is also able to generate a feed-back propagation of the information between the working statistical model and the record linkage stage.

**Abstract** *In questo articolo descriveremo, da un punto di vista bayesiano, alcuni avanzamenti metodologici riguardanti l'abbinamento di record (record linkage) appartenenti a file separati e la successiva inferenza con i dati abbinati. In particolare, il processo di record linkage verrà integrato con il modello inferenziale che si vuole stimare. In questo modo sarà possibile tenere conto dell'incertezza dell'abbinamento nella successiva inferenza e allo stesso tempo migliorare il processo di linkage grazie all'informazione contenuta nel modello statistico.*

**Key words:** Metroplis-Hastings algorithm, Record linkage, Regression

## 1 Introduction

From a methodological statistical perspective, the operation of merging two (or more) data sets can be important for two different and complementary reasons:(i) per sé, i.e. to obtain a larger reference data set or frame, suitable to perform more

---

Andrea Tancredi  
MEMOTEF, Sapienza Università di Roma, e-mail: andrea.tancredi@uniroma1.it

Brunero Liseo  
MEMOTEF, sapienza Università di Roma e-mail: brunero.liseo@uniroma1.it

accurate statistical analyses;(ii) to calibrate statistical models via the additional information which cannot be extracted from either one of the two single data sets. If the merging step can be accomplished without errors (maybe because a clear identification key is available and it can be used to match the units in the two dataset), there are no specific consequences on the statistical procedures undertaken in both the situations. In practice, however, identification key are often not available and linkage between statistical units is performed under uncertainty. This is when the record linkage process must be performed and the possibility to make wrong matching decisions is to be accounted for, especially when reporting the subsequent statistical analyses.

To briefly explain what record linkage is, let us suppose to have two data sets, say  $F_1$  and  $F_2$ , whose records respectively relate to statistical units (e.g. individuals, firms, etc.) of partially overlapping samples (or populations)  $S_1$  and  $S_2$ . The records in the two data sets consist of several fields, or variables, either quantitative or categorical, which may be observed together a potential dose of noise. For example, in a file of individuals, fields can be *surname*, *age*, *sex*, and so on. The goal of a record linkage procedure is to detect all the pairs of units  $(j, j')$ , with  $j$  in  $S_1$  and  $j'$  in  $S_2$ , such that  $j$  and  $j'$  refer actually to the same unit, via the information provided by the observed records in the two datasets. If the principal goal of the record linkage process is the former outlined above ( case (i)), a new data set is created by merging together three different subsets of units: those which are present in both data sets, those belonging to  $S_1$  only and those belonging to  $S_2$  only. Of course, information regarding the first group of individuals will be richer. Appropriate statistical data analyses may be then performed on this enlarged data set: however, since the linkage step is done with uncertainty, it may be affected from the presence of duplicate units and by a sort of lack of efficiency, mainly due to erroneous matching in the merging process.

On the other hand, the situation (ii) , which will be discussed throughout this paper, is even more challenging, both from a practical and from a methodological perspectives. Let us assume that the observed variables in  $F_1$  are denoted by  $(Y, V_1, V_2, \dots, V_h)$  while the observed variables in  $F_2$  are  $(X, V_1, V_2, \dots, V_h)$ . Then one might be interested in performing a linear regression analysis (or any other more complex association model) between  $Y$  and  $X$ , restricted to those pairs of records which are declared matches after a record linkage analysis based on variables  $(V_1, \dots, V_h)$ . The intrinsic difficulties which are present in such a simple problem are well documented in [ Neter (1965) ] and discussed in [Scheuren and Winkler (1993,1997), Larsen and Lahiri (2005)]. However, it is easily seen that, in the regression case, the presence of false matches (that is, matching record pairs which do not refer to the same statistical unit) reduces the observed level of association between  $Y$  and  $X$  and, as a consequence, they introduces a bias effect towards zero in estimating the slope of the regression line. Similar biased are caused in any statistical procedure and, in most cases, this bias takes a specific direction. As another example, when linkage procedures are used for estimating the size  $N$  of a population through a capture-recapture approach, the presence of false matches may severely reduce the final estimate of  $N$

We should also note, at this point, that in practical use of record linkage, it is quite usual that the linker (the researcher who matches the two files) and the analyst (the one which performs the statistical analysis) are two different persons, working separately. However, as [Scheuren and Winkler (1993)] states “... *it is important to conceptualize the linkage and analysis steps as part of a single statistical system and to devise appropriate strategies accordingly*”. Following such a suggestion, and putting it into a broader perspective, let us suppose to observe the variables  $(Y_1, Y_2, \dots, Y_k, V_1, V_2, \dots, V_h)$  on  $n_1$  units in file  $F_1$  and the variables  $(X_1, X_2, \dots, X_p, V_1, V_2, \dots, V_h)$  observed on  $n_2$  units in file  $F_2$ . In this set-up we consider the two-fold objective of using the key variables  $V_1, V_2, \dots, V_h$  to infer about the true matches existing between sources  $F_1$  and  $F_2$  and, at the same time, of using a statistical model  $\mathcal{M}$  to perform an analysis based on variables  $Y$ s and  $X$ s (or even including the  $V$ 's), restricted to those records which have been recognized as matches. In order to perform this double task, we argue that a fully Bayesian analysis is the simplest way to obtain an integrate use of the information which

- improves the performance of the linkage step (through the use of the extra information contained in the  $Y$ 's and  $X$ 's. This happens because pairs of records which do not adequately fit the model  $\mathcal{M}$  will be automatically down-weighted in the matching estimation;
- allows, in a natural way, to account for matching uncertainty in the estimation procedure related to model  $\mathcal{M}$  involving  $Y$ 's and  $X$ 's.
- improves the accuracy of the estimators of the parameters of model  $\mathcal{M}$  in terms of bias.

## 2 A Bayesian model for record linkage

In this section we briefly sketch the mathematical framework of the model. First we discuss in detail the Bayesian model for record linkage and then we show how to link it with the inference model  $\mathcal{M}$ . Suppose we have two matrices of record, say  $V_1$  and  $V_2$  of different sizes  $n_1$  and  $n_2$  respectively. Here  $V_1 = (v_{11}, \dots, v_{1n_1})$  and  $V_2 = (v_{21}, \dots, v_{2n_2})$  and each single  $v_{ij}$  can be represented as  $v_{ij} = (v_{ij1}, \dots, v_{ijh})$ , that is it contains the observed values of a categorical random vector  $v = (v_1, \dots, v_h)$  whose support is  $\mathcal{V} = \{v_{j_1 j_2, \dots, j_h} = (j_1, \dots, j_h) \mid j_1 = 1, \dots, k_1; \dots; j_h = 1, \dots, k_h\}$ . Also, consider the sets  $M$  and  $U$  of “true matches” and “true non matches” respectively. More precisely,  $M = \{(j, j') : \text{record } j \in V_1 \text{ and } j' \in V_2 \text{ refer to the same unit}\}$ , and, of course,  $U = M^c$ . The main goal of any record linkage technique is to identify which pair of records should be assigned to  $M$ .

We model the observed data matrices  $V_1$  and  $V_2$  of key variables, taking into account both the potential measurement error and the matching constraints. Let  $\tilde{v}_{ijl}$  be true unobserved value for the field  $l$  of the record  $j$  on data set  $V_i$ . We assume that  $p(v_{ijl} | \tilde{v}_{ijl}) = \gamma \delta_{\tilde{v}_{ijl}}(v_{ijl}) + (1 - \gamma) p(v_{ijl}) \quad \forall ijl$ . Notice that  $v_{ijl}$  is a mixture of two components, the former is degenerate on the true value while the latter is distributed over the other possible values of the variable;  $\gamma$  is the probability that the variable

$V_l$  is observed without noise. This model, known as “hit and miss”, was introduced by [Copas and Hilton (1990)]

To build a model for true values  $\tilde{v}_{ijl}$ s we introduce the matching matrix  $C$ . In particular, let  $C$  be a  $n_1 \times n_2$  unknown matrix whose entries are either 0 or 1, and  $C_{jj'} = 1$  represents a match,  $C_{jj'} = 0$  denotes a non-match: We assume that each data set does not contain replication of the same unit so that  $\sum_{j'} C_{jj'} \leq 1$ , and  $\sum_j C_{jj'} \leq 1$ . We assume that the true values distributions depends on the entries of  $C$ , that is,

$$p(\tilde{V}_1, \tilde{V}_2 | C) = \prod_{j: C_{jj'}=0 \forall j'} p(\tilde{v}_{1j}) \prod_{j': C_{jj'}=0 \forall j} p(\tilde{v}_{2j'}) \prod_{jj': C_{jj'}=1} p(\tilde{v}_{1j}, \tilde{v}_{2j'}). \quad (1)$$

Moreover, we take independent multinomial sampling for the non match true values, that is

$$p(\tilde{v}_{ij} = (j_1, \dots, j_h)) = \prod_{l=1}^h \theta_{lj_i}$$

and a degenerate joint distribution for the true match values

$$p(\tilde{v}_{1j}, \tilde{v}_{2j'}) = \begin{cases} 0 & \text{if } \tilde{v}_{1j} \neq \tilde{v}_{2j'} \\ \prod_{l=1}^h \theta_{lj_l} & \text{otherwise} \end{cases}$$

Notice that the above model represents a simplified version of the model proposed in [Tancredi and Liseo (2011)], where an additional layer - formulating a super-population model - was added at the top of the hierarchy. This simpler model can be obtained by integrating out with respect to the additional layer of hierarchy, under specific prior assumptions.

The model outlined cannot be analysed in closed form and some form of simulation from the posterior distribution is necessary. In particular, we have used a Metropolis-Hastings algorithm where the updating of the  $C$  matrix is performed by adding, or deleting or switching matches. For example, when proposing a move from  $C_{jj'} = 0$  to  $C_{jj'} = 1$ , we accept the move with probability given by

$$1 \wedge \frac{q(C|C')}{q(C'|C)} \frac{p(V_1, V_2 | C', \theta, \gamma)}{p(V_1, V_2 | C, \theta, \gamma)} \frac{p(C')}{p(C)}$$

where

$$\frac{p(V_1, V_2 | C', \theta, \gamma)}{p(V_1, V_2 | C, \theta, \gamma)} = \frac{p(v_j, v'_j | \theta, \gamma)}{p(v_j | \theta, \gamma) p(v'_j | \theta, \gamma)}. \quad (2)$$

If the move is accepted, we propose new true values  $(\tilde{v}_j, \tilde{v}'_j)$  by sampling from their full conditional distributions given the new status  $C_{jj'} = 1$ . Notice that the ratio (2) appearing in the above acceptance probability represents the Bayes factor comparing the hypothesis that the pair  $(j, j')$  is a match versus the alternative that it is not a match: see for example [Lindley (1977), Liseo and Tancredi (2011)] for similar expressions when Gaussian distributions are adopted. After that a reasonably large sample is drawn from the posterior distribution, we propose to estimate the matching configuration by declaring matches only the pairs where  $p(C_{ij} = 1 | V_1, V_2) \geq \frac{1}{2}$ .

### 3 Bayesian Regression with linked data

Suppose the first data set is a  $n_1 \times (h+1)$  matrix consisting of the variables  $(y, V_1)$ , while the other data set is a  $n_2 \times (h+p)$  matrix, including variables  $(V_2, \mathbf{X})$  where  $\mathbf{X} = (X_1, \dots, X_p)$ . Let  $\tilde{\mathbf{X}}$  be the matrix containing the true (unobserved) covariate values for  $Y$ .  $\tilde{\mathbf{X}}$  has dimension  $n_1 \times p$ . Conditionally on  $\tilde{\mathbf{X}}$  and on the true matching variables  $\tilde{V}_1$  and  $\tilde{V}_2$ , we assume a Gaussian linear regression model for  $Y$ , that is,  $Y|\tilde{\mathbf{X}}, \beta, \sigma^2 \sim N(\tilde{\mathbf{X}}\beta, \sigma^2 I)$ . In addition, given the matrices of true values  $\tilde{V}_1$  and  $\tilde{V}_2$ , we assume, for  $V_1$  and  $V_2$ , the “hit and miss” model as illustrated before. Conditionally on the matching matrix  $C$ , we also assume that the covariates for  $y_j$  are given by the vector  $x_{j'}$  (the  $x$ -part of the  $j'$ -th row of data set  $B$ ) only when  $C_{jj'} = 1$ ; otherwise we assume that they are unknown with a specific distribution  $p(\tilde{x})$ . The choice of  $p(\cdot)$  is not crucial and, in general, a multivariate Gaussian distribution will be used. More precisely, we have

$$p(\tilde{\mathbf{X}}|C) = \prod_{jj': C_{jj'}=1} \delta_{x_{j'}}(\tilde{x}_j) \prod_{j: C_{jj'}=0 \forall j'} p(\tilde{x}_j)$$

For the matrices of true values  $\tilde{V}_1$  and  $\tilde{V}_2$  we will adopt the same model expressed by (1). Notice that the covariates vector for non-matches pairs are handled as missing variables. The posterior simulation can easily be conducted via a standard Metropolis-Hastings algorithm. In this case, an “add-one-match” move will be accepted with a probability depending on

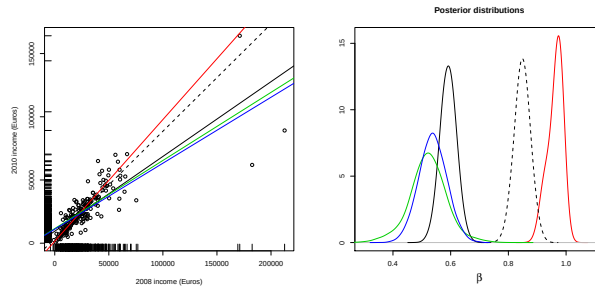
$$\frac{p(y, V_1, V_2 | C', \theta, \gamma)}{p(y, V_1, V_2 | C, \theta, \gamma)} = \frac{\phi(y_j; x_{j'}^T \beta, \sigma^2)}{\int \phi(y_j; \tilde{x}^T \beta, \sigma^2) p(\tilde{x}) d\tilde{x}} \frac{p(v_j, v_{j'} | \theta, \gamma)}{p(v_j | \theta, \gamma) p(v_{j'} | \theta, \gamma)}. \quad (3)$$

There are several important remarks and comments concerning formula (3). Firstly, there is a feedback effect between the parameters of the linkage block of the model and those of the regression part. This happens because all of them appear in the acceptance probability. This implies, for example, that the posterior distribution of the  $\beta$  vector will not be independent on  $C$ . This must be interpreted as a bias-correction effect. Secondly, a closed-form expression for  $p(y_j; \beta, \sigma^2) = \int \phi(y_j; \tilde{x}^T \beta, \sigma^2) p(\tilde{x}) d\tilde{x}$  can be obtained, for example, by assuming a multivariate normal for  $p(\tilde{x})$ . Thirdly, when the “add-one-match” move gets accepted, we update the true values  $(\tilde{x}_j, \tilde{v}_j, \tilde{v}'_j)$  by drawing from their full conditional distributions conditionally on the new status  $C_{jj'} = 1$ .

### 4 Application

We now show an application of the proposed methods using the data of the Italian Survey on Household Income and Wealth which is a sample survey conducted by the Bank of Italy every 2 years. The 2010 survey covers 7,951 households composed

of 19,836 individuals. Panel households and individuals represent 58% of the data. From the 2010 survey we consider the individual net disposable income ( $y$  variable) and the following matching variables: sex, age, marital status, employment status, working sector. From the 2008 survey we consider, in addition to the matching variables, the 2008 net disposable income which will be assumed as a covariate  $x$ . Aim of the application will be to estimate a regression model between  $y$  and  $x$  considering only the information provided by the matching variables. Figure 1 shows the results obtained for the Friuli region via our integrated record linkage and regression model. In particular, the left panel, shows the estimated regression line while the right panel reports the posterior distribution of the regression coefficient. Figure 1 shows also the results obtained by considering the true matches, a plug-in approach based on a regression conditional on the matching matrix estimated only via the linkage model and another approach where we propagate the matching uncertainty without exploiting the information of the regression model.



**Fig. 1** Results with a single region (Friuli),  $n_1 = 434, n_2 = 355$  Black line: (true) regression line given by the 203 true matches. Black dashed line: true regression line without 2 very influential obs. Red line: Bayesian estimation via the regression and linking model. Green line: Bayesian estimation via the linking model and posterior regression plug-in. Blue line: Bayesian estimation by matching uncertainty propagation from the linking model.

## References

1. Copas, J., Hilton, F.: Record linkage: statistical models for matching computer records. *Journal of the Royal Statistical Society, A*, **153** 287-320, (1990)
2. Lahiri, P., Larsen, M. D.: Regression analysis with linked data. *Journal of the American Statistical Association*. **100** 2222-2230, (2005)
3. Lindley, D.: A problem in forensic science. *Biometrika*, **64** 207-213, (1977)
4. Liseo, B., and Tancredi, A.: Bayesian estimation of population size via linkage of multivariate normal data sets. *Journal of Official Statistics*. **27** 491-505 (2011)
5. Scheuren, F., Winkler, W.E.: Regression analysis of data files that are computer matched. *Survey Methodology*. **19** 39-58, (1993)
6. Scheuren, F., Winkler, W.E.: Regression analysis of data files that are computer matched, Part II. *Survey Methodology*. **23** 157-165, (1997)
7. Tancredi, A., Liseo, B.: A hierarchical Bayesian approach to record linkage and population size problems. *Annals of Applied Statistics*. **5** 1553-1585, (2011)