Some advances on Bayesian record linkage and inference for linked data

Brunero Liseo and Andrea Tancredi
MEMOTEF, Sapienza Università di Roma
Viale del Castro Laurenziano 9, Roma 00161
{brunero.liseo, andrea.tancredi}@uniroma1.it

Abstract: In this paper we review some recent advances on Bayesian methodology for performing Record Linkage and for making inference using the resulting matched units. In particular we frame the record linkage issue into a formal inferential problem and we adapt standard model selection techniques to this context. Although the methodology is quite general, we will focus on the simple multiple regression set-up for expository convenience.

Keywords: Bayesian computational methods, capture-recapture, model selection.

1. Bayesian use of linked data

In general, from a statistical methodology perspective, the merge of two (or more) data files can be important for two different and complementary reasons:

- *per sé*, i.e. to obtain a larger and integrated reference data set;
- to perform a subsequent statistical analysis based on the additional information which cannot be extracted from either one of the two single data files.

The first situation need not any further comment: a new data set is created and appropriate statistical analyses will be performed based on it. However, the statistical theory behind the two situations must be different and we will comment on this problem later. On the other hand, the second situation is more interesting both from a practical and a theoretical perspectives. Let us consider a toy example to fix the ideas.

Suppose we have two computer files, say A and B, whose records respectively relate to units (e.g. individuals, firms, etc.) of partially overlapping populations P_A and P_B. The two files consist of several fields, or variables, either quantitative or qualitative. For example, in a file of individuals, fields can be "surname", "age", "sex", etc. The goal of a record linkage procedure is to detect all the pairs of units (a,b), with a in P_A and b in P_B, such that a and b refer actually to the same unit.

Suppose that the observed variables in A are denoted by $(Z, W_1, W_2, \ldots, W_h)$ and the observed variables in B are $(X, W_1, W_2, \ldots, W_h)$. Then one might be interested in performing a linear regression analysis (or any other more complex association model) between Z and X, restricted to those pairs of records which are declared matches after a record linkage analysis based on variables W_i's. The intrinsic difficulties which are present in such a simple problem are well documented and discussed in Scheuren and Winkler (1997) and Lahiri and Larsen (2005).

In statistical practice it is quite common that the linker (the researcher who matches the two files) and the analyst (the statistician doing the subsequent analysis) are two different persons working separately. However, we agree with Scheuren and Winkler (1997), which say " *... it is important to conceptualize the linkage and analysis steps as part of a single statistical system and to devise appropriate strategies accordingly.*"

In a more general framework, suppose one has $(Z_1,Z_2,...,Z_k, W_1,W_2,...,W_h)$ observed on n_A units in file A and $(X_1,X_2,...,X_p, W_1,W_2,...,W_h)$ observed on n_B units in file B. Our general goal can be stated as follows:

- use the key variables $W_1,W_2,...,W_h$ to infer about the true matches between A and B.
- perform a statistical analysis based on variables Z's and X's restricted to those records which have been declared matches.

To perform this double task, we argue that a fully Bayesian analysis allows for an integrate use of the information which improves the linkage step and allows to account for the matching uncertainty in the estimation of the regression coefficients. The main point to stress is that in our approach all the uncertainty about the matching process is automatically accounted for in the subsequent inferential steps. This approach is based on the Bayesian model for record linkage described in Fortini et al. (2001) and improved in Tancredi and Liseo (2011). We present the general theory and illustrate its performance via simple examples. In Section 2 we briefly recall the Bayesian approach to record linkage proposed by Fortini et al. (2001) to which to refer for details. Section 3 generalizes the method to include the inferential part. Section 4 concentrates on the special case of regression analysis, the only situation which has been already considered in literature: see Scheuren and Winkler (1993) and Lahiri and Larsen (2005) for an historical account and for a more detailed illustration.
Section 5 discuss the record linkage problem as one of model selection.

2. Bayesian Record Linkage

In Fortini et al. (2001) a general technique to perform a record linkage analysis is proposed. Starting from a set of key variables $W_1,W_2,...,W_h$, observed in two different sets of units, the method defines, as the main parameter of interest, the *matching matrix* C, of size n_A times n_B, whose generic element c_{ab} is either 0 or 1 according whether records a and b refer to the same unit. The parameter of interest C must satisfy some obvious constraints: we assume there are no duplicates either in P_A and P_B; this implies that the row and column sums of C will be either 0 or 1. In classical statistical inference matrix C would be defined to be a latent unobserved structure.
The statistical model is based on a multinomial likelihood function where all the of comparisons between key variables among units are measured on a 0/1 scale. As in the mixture model proposed by Jaro (1995) a central role is played by the parameter vectors m and u, both of length 2^h, with

$$m_i = P(Y_{ab}=y_i;\ c_{ab}=1); \qquad u_i = P(Y_{ab}=y_i;\ c_{ab}=0).$$

for $i=1, ...,2^k$, and Y_{ab} represents the 2^h-dimensional vector of comparisons between units $a \in A$ and $b \in B$. In the vast majority of the applications comparisons are based on a 0/1 scale, with Y_{ab} being a vector of 0's and 1's according whether the corresponding key variable matches or not between the two records. This approach implies an obvious loss of information; recently, Tancredi and Liseo (2011) have proposed a different approach which is based on the actual observed values of the key variables.
Then, independently of the way in which comparisons are performed, a Bayesian way to record linkage goes through the generation of a Markov Chain Monte Carlo sample

from the posterior distribution of the matrix valued parameter *C*. See Fortini et al. (2001) and Tancredi and Liseo (2011) for a discussion about the appropriate choices for the prior distribution on *C* and on the other parameters of the model, mainly *m* and *u*.

3. Inference with linked data

In this section we illustrate how to construct and calibrate a statistical model based on a data set which is the output of a record linkage procedure. As we already stressed, the final output provided by the procedure described in the previous section will be a simulated sample from the (approximated) joint posterior distribution of the parameters, say *(C, m, u;* ξ), where ξ includes all the other parameters in the model.
This can be used according to two different strategies. In fact we can either

1) compute a "point" estimate of the matrix C and then use this estimate to establish which pairs are passed to the second stage of the statistical analysis. In this case, the second step is performed with a fixed number of units (the declared matches). It must be noticed that, given the particular structure of the parameter matrix C, no obvious point estimates are available. The posterior mean of C is in fact useless since we need to estimate each single entry c_{ab} either with 0 or 1 values. The posterior median is difficult to define as well, and the most natural candidate, the maximum a posteriori (MAP) estimate typically suffers from sensitivity (to the prior and to Monte Carlo variability) problems: this last issue is particularly crucial in official statistics, where inferential results must be used for making decision. For a discussion on these issues see Tancredi et al. (2005) and, for related problems in a different scenario, Green and Mardia (2006).
2) Alternatively, one can transfer the "global" uncertainty relative to *C* (and to the other parameters), expressed by their joint posterior distribution, to the second step statistical analysis.

We believe that this latter approach is more sensible in the way it deals with uncertainty. Among other things, it avoids to over-estimate the precision measures attached to the output of the second step analysis.
The most obvious way to implement approach B simply consists in performing the second step analysis at the same time as the record linkage analysis, that is, including the second step analysis into the MCMC procedure. This will cause a feed-back propagation of the information between the record linkage parameters and the more specific quantities object of interest. Here we illustrate these ideas in a very general setting; in the next section we will consider the regression example in details.

Let *D=(Y, Z, X)* the entire set of available data where, as in the Introduction, Y_{ab} represents the vector of comparisons among variables which are present in both files (or the *2h* dimensional vector when the actual values of the key variables are observed), Z_a is the value of covariate *Z* observed on individual $a \in A$ and X_b is the value of covariate *X* observed on individual $b \in B$. The statistical model can then be written as

$$p(y,z,x|\ C,m,u,\theta,\ \xi)$$

where *(C; m, u, ξ)* are the record linkage parameters and θ is the parameter vector related to the joint distribution of *(X;Z)*. The above formula can always be re-expressed as

$$p(y|\ C,m,u,\theta,\ \xi)\ p(z,x|\ y,\ C,m,u,\theta,\ \xi)$$

It is then reasonable to assume that, given C, the comparison vector *Y* does not depend on θ; also, given *C*, the distribution of *(X;Z)* should not depend both on the comparison vector data *Y* and the parameters related to those comparisons. It follows that model can be simplified into the following general expression:

$$p(y \mid C,\ m,u)\ p(z,\ x \mid C,\ \theta) \qquad (1)$$

The first term in the last expression is related to the record linkage step; the last term refers to the second step analysis and must be specified according to the particular statistical analysis. The presence of *C* in both terms allows for the feed-back phenomenon we mentioned before. Approaches A and B can be re-phrased using the last formula.

In the case A) the first factor of the model is used to get an estimate $\hat{C}$ of *C*. Then $\hat{C}$ is plugged into the second factor and a standard statistical analysis is performed to get an estimate of θ. In approach B) the two factors are considered together within the MCMC algorithm thus providing a sample from the joint posterior distribution of all the parameters. In this case the Markov Chain which produces the posterior sample allows for an information feedback between *C* and θ.

There is actually a third possible approach to consider and we call it approach C). In fact, one can use a MCMC algorithm with the first factor only and, at each step *t=1, ..., T*, of the algorithm one can perform the statistical analysis expressed by the second factor of the model fixing the record linkage parameters at their values, say $C^{(t)}$, the value of the Markov chain for the parameter *C* at time *t*. This way, one can obtain an estimate $\hat{\theta}$ of θ at each step of the MCMC algorithm and then somehow summarize the set of estimates. In the next section we will illustrate the three approaches in the familiar setting of the simple linear regression.

We anticipate that approach A) seems to miss to account for the uncertainty in the first step of the process and, consequently, it tends to produce a false impression of accuracy in the second step inferences.

In general, we consider approach B) as the most appropriate in terms of the use of statistical information provided by the data. However, approach C) can be particularly useful especially if the set of linked data must be used more than one time, for different purposes. In fact, while in approach B) information flows back and forth from C to θ, in case C) the information goes one-way from C to θ and the record linkage step is not influenced by the information provided by *(X,Z)*.

4. Multiple linear regression

Consider again the toy example in the Introduction and assume that our object of interest is the linear relation between *X* and *Z,* say

$$Z = X\beta + \sigma\epsilon$$

with ϵ being a vector of i.i.d. standard normal random variables, and $\theta=(\beta,\sigma)$. One should notice that, the length of vectors Z and ϵ are not fixed in advance, since they depend on the number of matched units. Here we describe how to implement the three different approaches discussed in Section 3. In the following we assume that our statistical model can be simplified according to (1).

First we give a brief account of the method proposed by Larsen and Lahiri (2005), which generalize the pioneer approach developed in Scheuren and Winkler (1997) . They assume that the two datasets consist of the same number of units, say $n=n_A=n_B$. This assumption is quite restricted in practice. With respect to model (1), consider the matrix P where the generic element p_{ab} denotes the probability that the a-th unit of database A coincides with the b-th unit of database B. Assume that the main goal is the estimation of the regression parameters $(\beta_1,\beta_2,\ldots,\beta_h)$

Since the information about the true links is missing, it is useful to introduce the new variables ($V_1, V_2 \ldots, V_p$), where each V_i is any of the values of the response variable observed on the n units, each assumed with probability p_{ij}. Using our notation, their approach corresponds to introduce, for each unit in A, a latent vector

$$S_a = (S_{a1}, S_{a2}, \ldots, S_{an})$$

which consists of just *one 1* and *n-1* zeros; the *one* is of course the identifier of the unit a in file B. Also, $S_1, S_2, \ldots, S_n$ are assumed to be mutually independent with a multinomial distribution with parameters $(1, p_j)$, and

$$p_j = (p_{j1}, p_{j2}, \ldots, p_{jn})$$

Then it is easy to see that

$$E(Z_j \mid S_1, S_2, \ldots, S_n) = \Sigma_b S_{jb} X_j' \beta$$

and, by the law of the iterate mean, in matrix form

$$E(Z) = PX\beta$$

This produces an unbiased estimator of θ, that is

$$(X'P'PX)^{-1} X'PZ$$

In other words, in order to account for the uncertainty about matching, Larsen and Lahiri (2005) propose the use of a *weighted combination of covariates,* where the weights are *estimated* from the linkage model step. They also provide an estimate for the variance of the estimator via a parametric bootstrap approximation.

In general, linkage errors may weaken a linear regression analysis in several different ways;

a) If one fails to detect a match, standard error of the ML estimates increase.
b) If a false match is introduced in the analysis, on average, one introduces a bias which shrinks the ML estimates of the regression coefficients toward zero.

The same problem will be likely to happen for the posterior distribution of the regression coefficients in a Bayesian analysis.

Here we will try to go beyond the limitation of equal sample sizes for the two files and notice that, for a given matching matrix *C,* the *correctly linked* regression model can be written as

$$C'Z = C'CW\beta + C'\sigma\epsilon \qquad (2)$$

with the convention that one must eliminate the lines with zero components in the vector $C'Z$ in the above equation. From this perspective it is clear that the introduction of the matrix C allows for a direct generalization of the Larsen-Lahiri methods to the more general case of different sample sizes.
We now discuss the three different strategies illustrated in the previous section, with a particular emphasis to the multiple regression framework.

Method A.

I. Use any Record Linkage procedure to establish which pairs of records are true matches.
II. Use the subset of matched pairs to perform a linear regression analysis and provide an estimate of θ via ordinary least squares, maximum likelihood or Bayesian method.

This methodology corresponds to select a point estimate of C and to use it in the above regression expression. All the uncertainty about the matching procedure is clearly lost and not transferred to the regression analysis.

Method B.

I. Set a MCMC algorithm relative to model (1), that is, at each iteration $t=1, \ldots, T$,
II. draw $C^{(t)}$ from its full conditional distribution
III. draw $(m^{(t)}, u^{(t)}, \xi^{(t)})$ from the full conditional distribution
IV. draw $\theta^{(t)}$ from its full conditional distribution

From steps B-II and B-III one can notice that the marginal posterior distribution of C will be potentially influenced by the information on θ. In this case the posterior distribution of θ will account for the uncertainty related to linking procedure in a coherent way.
From a theoretical perspective, this is the coherent way to proceed. All the relations among variables and parameters are potentially considered and uncertainty is accounted for in the correct way.

Method C.

I. Set up a MCMC algorithm restricted to the first factor of (1) in order to produce a posterior sample from the joint posterior distribution of (C,m,u). This can be done using, for example, the algorithms illustrated in Fortini et al. (2001) and Tancredi and Liseo (2011).
II. At each iteration $t=1, \ldots, T$ of the MCMC algorithm, use $C^{(t)}$ to perform a linear regression analysis restricted to those pairs of records (a,b) such that $c^{(t)}_{ab}=1$, and produce a point estimate $\hat{\theta}$ of θ, for example the OLS estimate.

III. Use the list of estimates as an approximation of the "predictive distribution" of the used estimator.

In this third approach, setting S=C'C and using the fact that S is idempotent, from (2) one obtains, at each iteration, that $\hat{\theta}$ is equal to

$$(X'SX)^{-1} X'SC'Z$$

It follows that, in this approach, the estimation of C is not influenced by the regression part of the model. This method could be safer to use (and to be preferred) if the main goal of the record linkage step was to create an enriched and reference dataset to be repeatedly used in the future for different purposes.
Under the additional assumption that, given the matching matrix *C,* variables used in regression are unrelated to the key variable used in the record linkage analysis, methods B and C provide similar results.

From a computational perspective, method B is complicated by the fact that the full conditional of the extra-parameters given the record-linkage parameters, must be derived for any different statistical models; also the introduction of new parameters is likely to change the full conditionals of the record-linkage parameters and it might be not so simple to adjust the MCMC algorithm. This is another compelling, although practical, reason for preferring method C.

4. Selection of matches as a model selection problem

In this section we will rephrase the record linkage problem as one of variable selection in regression analysis. Suppose there are *p* potential explanatory variables available for the analysis and the researcher must select the *best subset* of variables among the 2^p possible choices.
Let K_j, $j=1, \ldots, 2^p$, the generic subset of covariates. In a Bayesian framework, one can usually compute, for each possible K_j, its posterior probability $P(K_j; data)$.
Then one can choose either

- The maximum a posteriori (MAP) model, that is the subset K_j with the highest posterior probability. This choice is optimal under a zero-one loss, although it typically suffers from a robustness problem.
- The median posterior model, (MeM) that is the subset of covariates which includes all the regressors which have a marginal posterior probability higher than 0.5. See Barbieri and Berger (2004) for details. One can show that this choice is optimal for predictive purposes under a large range of reasonable loss functions and it is also more robust than the MAP
- If prediction is the ultimate goal one need not necessarily choose a single model and an average prediction can be made using predictions from each single model weighted with their posterior probabilities. This approach is superior in terms of accounting for uncertainty since each single "inference" is weighed by its posterior probability. This methodology is generally known as Bayesian Model Averaging.

In record linkage problems models correspond to specific choices of set of matches to be selected. Given n_A and n_B there are

$$\sum_{k=0}^{min(n_A, n_B)} n! \binom{n_B}{n} \binom{n_A}{n}$$

possible models to choose from.

From a more theoretical perspective the correspondence between point estimates of C and models has the only drawback that in a record-linkage problem *there is* a correct model while this is almost never a correct perspective in applied statistics where models are, at best, more or less reliable approximation to reality, and it might be more reasonable to account for "model" uncertainty.
We are currently working on this particular perspective.

References

Barbieri, M.M. and Berger, J..O. (2004). Optimal predictive model selection. *The Annals of Statistics, 32,* pp. 870-897.

Fortini, M., Liseo, B., Nuccitelli, A. and Scanu, M. (2001) On Bayesian Record Linkage. *Research in Official Statistics,* 4, Vol.1, 185-198.

Green, P.J. and Mardia, K.V. (2006). Bayesian alignment using hierarchical models, with application in protein bioinformatics, *Biometrika, 93, pp. 235-254.*

Herzog, T. N., Scheuren, F., and Winkler, W.E., (2007), *Data Quality and Record Linkage Techniques,* New York, N. Y.: Springer.

Herzog, T. N., Scheuren, F., and Winkler, W.E., (2010), *Record Linkage*, in (D. W. Scott, Y. Said, and E. Wegman, eds.) Wiley Interdisciplinary Reviews: Computational Statistics, New York, N. Y.: Wiley, 2 (5), September/October, 535-543 .

Jaro, M. (1989). Advances in record-linkage methodology as applied to matching the 1985 census of Tampa, Florida *Journal of the American Statistical Association, 84, pp. 414-420.*

Lahiri, P., Larsen, M. (2005). Regression analysis with linked data. *Journal of the American Statistical Association,* 100, pp-222-230

Larsen, M. (2005) Advances in record linkage theory: hierarchical Bayesian record linkage. ASA proceedings

Lindley, D.V. (1977) A problem in forensic science. *Biometrika, 64, pp. 207-213.*

Liseo, B., Tancredi, A. (2011) Bayesian estimation of population size via linkage of multivariate normal data sets. *Journal of Official Statistics,* Vol. 27 No. 3, pp. 491—505.

Scheuren, F.,and Winkler, W. E. (1997), Regression analysis of data files that are computer matched, II, *Survey Methodology*, 23, 157-165.

Tancredi, A., Liseo, B., Guagnano, G. (2005) Inferenza statistica basata su dati prodotti mediante procedure di record linkage. In *L'integrazione di dati di fonti diverse: tecniche e applicazioni del Record Linkage e metodi di stima basati sull'uso congiunto di fonti statistiche e amministrative* (P.D. Falorsi, A. Pallara, A. Russo (eds.)) Franco Angeli, pp. 41-59.

Tancredi, A., Liseo, B. (2011) A hierarchical Bayesian approach to record linkage and population size estimation. *Annals of Applied Statistics,* Vol. 5, No. 2B, 1553—1585.