MODELLING ISSUES IN RECORD LINKAGE: A BAYESIAN PERSPECTIVE

Marco Fortini, Alessandra Nuccitelli, ISTAT, Roma, Italy
Brunero Liseo, Università di Roma "La Sapienza", Italy
Mauro Scanu, ISTAT, via Cesare Balbo 16, 00184 Roma, Italy (scanu@istat.it)

Key Words: Exact matching, Monte Carlo Markov Chain methods, measurement error, mixture model.

Abstract:

Record Linkage (RL) refers to the use of an algorithmic technique to match records from different data sets that correspond to the same statistical unit. RL is ubiquitous in official statistics: estimation of population size via capture-recapture methods, testing of disclosure strategies and coverage measurement surveys are only few examples. A key difficulty for any statistical analysis with RL is the intensive computational burden. A Bayesian analysis of RL correctly clarifies the complexity of the problem and it helps in finding adequate solutions. In this paper we briefly discuss the validity of the more common statistical models for RL and we propose a fully Bayesian approach. We use standard MCMC algorithms to derive the marginal posterior distribution of a matrix-valued parameter which indicates the "configuration" of matches between the two lists.

1. Introduction

Record linkage is "the name given to any process which identifies the common reporting units in two different files" (Kelley, 1986). Such objective is very important in many different disciplines; among the others, medicine, business administration and official statistics (see, for instance, Newcombe, 1988, and Jabine and Scheuren, 1986). In these contexts it may happen that a unique data set with all the necessary information for a particular statistical analysis is not available. Furthermore time and cost constraints may make unfeasible to obtain such data set directly. Integration at the unit level of different data sets (sample surveys and/or administrative data sets) may solve this problem. A difficulty is represented by the lack of a unique identifier in the different data sets for each unit of interest. When a set of observed variables (key variables, henceforth) may be used as an identifier for connecting records that refer to the same unit, particular attention should be paid to errors and missing values. In such a case, many different methodologies have been introduced. Some methods are naïve, or heuristic, i.e. are based only on common sense (for instance the "iterative method" described in Armstrong and Mayda, 1993), whereas other methods refer to a typical statistical framework: i.e. a sample space and a sample probability distribution are defined. In this framework, usual statistical methods (estimators, tests) may be considered and an evaluation of their performance can be given. Generally speaking, most of these works are based on the formal mathematical framework provided by Fellegi and Sunter (1969) in a fundamental paper. Further advances are described in a number of papers in the 80's and 90's: among the others Jaro (1989), Winkler (1993), Belin and Rubin (1995) and Larsen and Rubin (2001). All of these papers assume that each single comparison between records in two different files provides new information, independently of the other comparisons. This assumption, as noted by Kelley (1986), is fundamentally incorrect, as illustrated in section 3. Also, Winkler (2000) states that "...because the underlying true probabilities have not been accurately estimated, estimated error rates (of the record linkage procedure) are not accurate". Consequently we propose a Bayesian model which makes comparisons among units independent of each other. We illustrate our ideas in the situation of a single continuous variable. Extensions to more general cases are sketched in the last section.

Brunero Liseo's research and Mauro Scanu's research are partially supported by MIUR grants.

2. The usual statistical model for record linkage

For the sake of simplicity, let us consider two data files $\mathcal{A}$ and $\mathcal{B}$, with respectively n_A and n_B units. Let us call A and B the two sets (lists) of observed units, $a = 1, ..., n_A$, $b = 1, ..., n_B$. We assume that some units are "common" to the two lists. The set of all ordered pairs

$$A \times B = \Big\{(a, b) \, : \, a \in A, \, b \in B\Big\}$$

can be splitted into

$$M = \Big\{(a,b) \in A \times B \,:\, a = b\Big\}$$

the set of matches, and

$$U = \Big\{(a,b) \in A \times B \,:\, a \neq b\Big\}$$

the set of non-matches. In order to decide whether a pair (a,b) is in M or U, we may compare variables observed in both the files (e.g. surname, name, sex, address, etc. for individuals). Let us assume we have k key variables, $k \geq 1$, whose observations in the two data lists are denoted by:

$$\mathbf{x}_a = (x_{a,1},\, x_{a,2},\, ...,\, x_{a,k}), \qquad a \in A,$$

and

$$\mathbf{x}_b = (x_{b,1},\, x_{b,2},\, ...,\, x_{b,k}), \qquad b \in B.$$

Generally speaking, the comparison $\mathbf{y}_{ab}$ of the observed values of the key variables between two units $a \in A$ and $b \in B$ is a function of $\mathbf{x}_a$ and $\mathbf{x}_b$:

$$\mathbf{y}_{ab} = f(\mathbf{x}_a,\, \mathbf{x}_b).$$

One commonly assumed comparison function is a vector of k elements, $\mathbf{y}_{ab} = (y^1_{ab}, ..., y^k_{ab})$ with:

$$y^h_{ab} = \begin{cases} 1 & \text{if } x_{a,h} = x_{b,h} \\ 0 & \text{otherwise} \end{cases} \qquad h = 1, ..., k. \quad (1)$$

In order to decide whether a pair (a,b) with comparison vector $\mathbf{y}_{ab}$ is a match or not, Fellegi and Sunter (1969) suggest to consider the distribution of the comparison vectors in M, say $m(\mathbf{y})$, and the corresponding distribution in U, $u(\mathbf{y})$. The decision rule is based on the likelihood ratio

$$t(\mathbf{y}) = \frac{m(\mathbf{y})}{u(\mathbf{y})} \quad (2)$$

(see Fellegi and Sunter, 1969, for a discussion on the optimality of such decision rule). Given that neither $m(\mathbf{y})$ nor $u(\mathbf{y})$ are known, most of the literature on record linkage studies how to estimate these distributions. The usual assumptions are that both the status of a pair (let's say C, where $C = 1$ when a pair (a,b) is a match and 0 otherwise) and the comparison vector ($\mathbf{Y}$) are random variables (r.v.) and:

1. the status c_{ab}, $(a,b) \in A \times B$, are i.i.d. observations of a Bernoulli r.v. C with $P(C = 1) = p$;
2. each comparison vector $\mathbf{y}_{ab}$, $(a,b) \in A \times B$, is an i.i.d. observation of the r.v. $\mathbf{Y}$ whose distribution is the mixture:

$$P(\mathbf{Y} = \mathbf{y}) = p\, m(\mathbf{y}) + (1-p)\, u(\mathbf{y});$$

3. the pairs $(c_{ab}, \mathbf{y}_{ab})$ are i.i.d. observations of the r.v. $(C, \mathbf{Y})$ whose distribution is:

$$P(C = c,\, \mathbf{Y} = \mathbf{y}) = \big(p\, m(\mathbf{y})\big)^c \big((1-p)\, u(\mathbf{y})\big)^{1-c},$$

with $c = 0, 1$.

The previous independence assumptions make particularly easy the computation of the likelihood function given the $n_A \times n_B$ observations $(c_{ab}, \mathbf{y}_{ab})$:

$$\prod_{(a,b) \in A \times B} \big(p\, m(\mathbf{y}_{ab})\big)^{c_{ab}} \big((1-p)\, u(\mathbf{y}_{ab})\big)^{1-c_{ab}}. \quad (3)$$

Maximum likelihood estimates of the distributions $m(\mathbf{y})$ and $u(\mathbf{y})$ may consequently be obtained, using for instance the EM algorithm (given that the status c_{ab} is unknown). Jaro (1989) assumes that the components of the comparison vector $\mathbf{Y}$ are independent, whereas Winkler (1993) and Larsen and Rubin (2001), among the others, consider the case of dependent key variables comparisons.

3. A modified model for record linkage

Kelley (1986) states: "...The decision procedure ... was developed under the hypothesis that the comparison vectors between separate record pairs are independent. However, since the record pairs that are considered for possible matches are elements of the cross product of the two files we are attempting to match, the comparison vectors are in fact dependent...". As a matter of fact, the r.v.'s $\mathbf{Y}_{ab}$ are deterministically dependent. For instance, consider the case of one key variable X and the comparison function in (1) (extensions to more than one key variable and more complex comparison functions are straightforward). Assume that $n_A = n_B = 2$; when

$$X_{a1} = X_{b1}, \qquad X_{a1} = X_{b2}, \qquad X_{a2} = X_{b1}$$

then, necessarily, $X_{a2} = X_{b2}$, i.e.:

$$P(Y_{22} = 1 | Y_{11} = 1, Y_{12} = 1, Y_{21} = 1) = 1.$$

Note that the problem of dependency among the Y_{ab}'s cannot be circumvented by eliminating redundant comparisons for the likelihood function (3), because the order with which pairs are considered would matter! Deterministic dependency affects also the r.v.'s C_{ab} when each record in A may be linked to at most one record in B (and vice versa). Such constraints:

$$\sum_b c_{ab} \leq 1, \qquad \forall\, a \in A \quad (4)$$

$$\sum_a c_{ab} \leq 1, \qquad \forall\, b \in B \quad (5)$$

have been studied by Jaro (1989) and Winkler and Thibaudeau (1991), but their procedures are only an additional and subsequent step of a record linkage procedure. On the other hand, a Bayesian model would naturally take into account such constraints, as developed in Fortini *et al.* (2001).

In the light of the previous considerations, we suggest the following:

- the statistical model is built upon *the statistical units*, $a \in A$ and $b \in B$, and not over *the pairs* $(a, b) \in A \times B$;
- observations over different statistical units can be considered independent.

The previous two considerations lead us to develop a model similar to the one in Copas and Hilton (1990). Moreover our approach refers explicitly to a measurement error framework (Fuller, 1995). For the sake of simplicity, let us consider only one continuous Gaussian key variable X. Each unit is an independent realization of $\mu_0 \sim N(\mu, \sigma^2)$. Whether observed in A or in B, a random measurement error occurs, so that

$$X|\mu_0 \sim N(\mu_0, \tau^2).$$

Denoting the r.v. X with the symbol X^A when observed on a unit $a \in A$ and X^B when observed on $b \in B$, we have that, marginally,

$$X^A \sim N(\mu, \sigma^2 + \tau^2), \qquad X^B \sim N(\mu, \sigma^2 + \tau^2).$$

Consequently, when we consider a pair (a, b), the bivariate variable (X^A, X^B) follows a bivariate Gaussian distribution according to the following rule:

1. when the pair (a, b) is not a match ($c_{ab} = 0$), X^A and X^B are independent and:

$$(X^A_a, X^B_b) \sim N_2\Big(\mu\mathbf{1}, (\sigma^2 + \tau^2)\mathbf{I}\Big),$$

   where $\mathbf{1}$ is a vector of 1's and $\mathbf{I}$ is the identity matrix of the appropriate dimension.

2. when the pair (a, b) is a match ($c_{ab} = 1$), only the random measurement errors in the two occasions are assumed independent; consequently X^A and X^B are correlated, with covariance equal to σ^2:

$$(X^A_a, X^B_b) \sim N_2(\mu\mathbf{1}, \sigma^2 + \tau^2\mathbf{I}).$$

Copas and Hilton (1990) adopt the previous set up using a Fellegi-Sunter approach, i.e. discriminating each couple distinctly via the likelihood ratio (2).

Following the approach in Fortini *et al.* (2001), we can instead write the likelihood function for all the observed units. Assuming that there are not duplications of units in a list, the number n of distinct units in A and B is such that

$$\begin{aligned} n &= \sum_{ab} c_{ab} + \Big[n_A - \sum_a c_{a\cdot}\Big] + \Big[n_B - \sum_b c_{\cdot b}\Big] = \\ &= n_A + n_B - \sum_{ab} c_{ab}. \end{aligned}$$

given that for each distinct unit there are three possibilities: the key variable is observed

1. only in A: $\sum_b c_{ab} = c_{a\cdot} = 0$;
2. only in B: $\sum_a c_{ab} = c_{\cdot b} = 0$;
3. both in A and B: $c_{ab} = 1$.

Consequently the likelihood function, induced by the observations $\{\mathbf{y}_{ab}\}$, is:

$$\begin{aligned} L(\theta, \mathbf{c}) &= \prod_{ab} \Big[N_2(\mu\mathbf{1}, \sigma^2 + \tau^2\mathbf{I})\Big]^{c_{ab}} \\ &\times \prod_a \Big[N(\mu, \sigma^2 + \tau^2)\Big]^{1-c_{a\cdot}} \\ &\times \prod_b \Big[N(\mu, \sigma^2 + \tau^2)\Big]^{1-c_{\cdot b}} \end{aligned}$$

where $\theta = (\mu, \sigma^2, \tau^2)$ and $\mathbf{c} = \{c_{ab}\}$.

4. Bayesian analysis of the new model

4.1 Prior distributions

The likelihood $L(\theta, \mathbf{c})$ has been characterized in terms of two sets of parameters: the parameters of interest (i.e. the matrix $\mathbf{c}$) and the parameter θ. For both the parameters, it is possible to have some kind of information that can be formalized in terms of suitable prior distributions. Following Fortini *et al.* (2001), the prior distribution for the random configuration matrix $\mathbf{C}$ can be given in two steps:

- the r.v. "number of matches" H follows a binomial distribution with parameters $(\xi, n_A \wedge n_B)$, $\xi \in (0, 1)$;
- conditional on $H = h$, $h = 0, 1, ..., n_A \wedge n_B$, the r.v. $\mathbf{C}|H = h$ follows a uniform discrete distribution over the set of configuration matrices with exactly h matches satisfying the constraints (4) and (5).

Table 1: Correct Match Rate (CMR=number of correctly guessed matches/number of true matches) and False Match Rate (FMR=number of wrongly guessed matches/number of true matches) for different values of τ^2

Rates	τ^2		
	0.01	0.006	0.001
FMR	0.6	0.4	0.2
CMR	0.4	0.6	0.8

(see Fortini *et al.*, 2001, for a justification for such priors). For the sake of simplicity, standard conjugate distributions may be considered for θ: then

$$\mu \sim N(\lambda, \omega^2),$$

$$\sigma^2 \sim IG(\alpha_\sigma, \beta_\sigma),$$

$$\tau^2 \sim IG(\alpha_\tau, \beta_\tau),$$

where $Z \sim IG(\alpha, \beta)$ stands for Inverse Gamma distribution with density function:

$$f(z) = \frac{\alpha^\beta}{\Gamma(\beta)} \frac{e^{\alpha/z}}{z^{\beta+1}}.$$

4.2 MCMC implementation

A closed form of the posterior distribution for the configuration matrix $\mathbf{C}$ is not available. For this reason we have adopted a MCMC approach to generate a sample from the posterior distribution. We use a standard Metropolis-Hastings algorithm to generate in turn candidate values of the parameters $(\mu, \sigma^2, \tau^2, \mathbf{c})$. The choice of the proposal distributions requires some attention.

The proposal distribution of μ was chosen to be Gaussian centered at the previous value of the chain and with the standard deviation tuned so to have a reasonable acceptance rate.

The same scheme was adopted for the logarithm of the two variances σ^2 and τ^2.

The update of the configuration matrix $\mathbf{c}$ is obviously more complicated: being in $\mathbf{c}_t$ at the t-th step, the algorithm proposes a new matrix $\mathbf{c}_{t+1}$ by choosing whether *i*) to add a link *ii*) to remove a link *iii*) to remain with the same number of links by deleting one link and adding a new one (see Fortini *et al.*, 2001, for details).

5. Simulation results

An example on fictitious data has been created for a first rough evaluation of the method. We have generated 20 units from a Gaussian distribution with mean $\mu = 100$ and variance $\sigma^2 = 9$. For each generated value, a random measurement error is added according to the rule:

$$X|\mu_0 \sim N(\mu_0, 0.001),$$

i.e. $\tau^2 = 0.001$. As far as the prior distributions are concerned, we have considered a high value for ξ ($\xi = 0.98$), given that we expect a high number of matches. For the sake of simplicity, we consider conjugate distributions for the parameters in θ:

- $\mu \sim N(100, 0.01)$;
- $\sigma^2 \sim IG(0.01, 0.01)$;
- $\tau^2 \sim IG(0.01, 0.01)$;

The two inverse gamma distributions are practically flat (i.e. non informative) in the log of the argument. The configuration with the maximum posterior distribution easily finds the majority of the 20 matched pairs (problems may occur for those pairs whose true values are close to each other). The previous case considers only the situation of a particularly low measurement error ($\tau^2 \sim \sigma^2/10000$). To test a more difficult situation, we have considered also the case of $\tau^2 = 0.006$ and $\tau^2 = 0.01$. Leaving unchanged the conjugate distributions, the performance of the algorithm worsen dramatically, but still a reasonable number of matches is found (see Table 1 for details). This situation can be reasonably expected and it may depend on some lack of discriminating information, which could be overcome by the introduction of other key variables. However we guess that room for improvement could be given by a better refinement of the MCMC algorithm.

6. The discrete case

The previous sections were devoted to the case of one continuous Gaussian key variable. Generally speaking, key variables are discrete (unless suitable transformations are adopted, as in Belin and Rubin, 1995). Let us consider the discrete r.v. α assuming the values $k = 1, ..., K$ with probabilities

$$P(\alpha = k) = p_k^\alpha, \qquad k = 1, ..., K.$$

Assuming the existence of a measurement error on the observations of α in the two occasions, X^A and X^B, let us define the following distributions (we use the notation adopted for latent class analysis, i.e. $p^{A|B}$ indicates the distribution of A conditional to B; see Goodman, 1974, for further details):

$$P(X^A = i, X^B = j) \quad = \quad p_{ij}^{AB} =$$

$$= \sum_k p_{ijk}^{AB\alpha} =$$
$$= \sum_k p_{ik}^{\overline{A}\alpha} p_{jk}^{\overline{B}\alpha} p_k^{\alpha},$$

for $i = 1, ..., K$, $j = 1, ..., K$ (note that, as in the continuous case, we assume that the random measurement errors are independent, conditional on the true value);

$$P(X^A = i) = p_i^A = \sum_k p_{ik}^{\overline{A}\alpha} p_k^{\alpha} \qquad i = 1, ..., K;$$

$$P(X^B = j) = p_j^B = \sum_k p_{jk}^{\overline{B}\alpha} p_k^{\alpha} \qquad j = 1, ..., K.$$

In this setting, the likelihood can be written as the following:

$$\begin{aligned} L(\theta, \mathbf{c}) &= \prod_{ab} \Big[\sum_k p_{ik}^{\overline{A}\alpha} p_{jk}^{\overline{B}\alpha} p_k^{\alpha} \Big]^{c_{ab}} \\ &\times \prod_a \Big[\sum_k p_{ik}^{\overline{A}\alpha} p_k^{\alpha} \Big]^{1-c_{a.}} \\ &\times \prod_b \Big[\sum_k p_{jk}^{\overline{B}\alpha} p_k^{\alpha} \Big]^{1-c_{.b}} \end{aligned}$$

Differently than in the Gaussian case, this likelihood contains an unidentifiable set of parameters. However, extensions of this model to more key values admit identifiable set of parameters via the use of loglinear models for latent variables by introducing appropriate constraints (Haagenaars, 1993). A second solution could be given by the use of the simplifying models adopted by Copas and Hilton (1990); this aspect should be studied in more detail.

7. Conclusions

In this work we deal with the problem of the misspecification of the statistical model for record linkage. Instead of a model based on the comparison between all the pairs of units, we consider a model based on each single distinct unit which can be observed twice when one unit is enlisted in both the data sets and only once otherwise. This approach refers explicitly to measurement error theory. We give a solution in the case of a single continuous key variable following a Gaussian distribution, affected by a Gaussian measurement error (which is assumed independent of each other in the two occurrences). Such solution is based on a Bayesian procedure where the posterior distribution is explored by means of a MCMC algorithm. An example on simulated data is given, reproducing situations for increasing values of measurement errors. This example shows the soundness of our method although its performance deteriorates when measurement errors become important. In this situation it is crucial to increase the number of key variables, as it usually happens in practice. Finally we sketch some possible developments to include the case of discrete key variables (which is more common in practice).

In this paper we have argued that the most popular statistical models for RL analysis fail to recognize the interdependency among record comparisons. Consequently, we have proposed an alternative approach which avoids the problem by building the model directly upon the observable key variables. At least in the simple setting considered in the paper, the new model can be implemented without any additional complexity in the MCMC algorithm.

However, in more general situations, the problem of the computational complexity is certainly to become crucial and some compromise between computational feasibility and theoretical adherence to the underlying mechanism should be made. In this direction a promising approach has been developed in de Freitas *et al.* (1999) and Blei *et al.* (2003) where "complicated" statistical models, which could produce slowly convergent MCMC algorithms, are approximated by simpler ones, with faster MCMC counterparts: the approximations are established using variational arguments. The importance of these new ideas in RL is twofold:

- it is possible to use this idea to quantify the loss of precision inherent in the use of the usual models in RL by considering them as approximations to the more realistic one, based on the observable key variables, as developed in this paper;
- starting from our model, new directions in the "simplification" strategy can be explored, and their practical impact on inference can be evaluated.

8. Bibliography

Armstrong J., 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 rates in record linkage". *Journal of the American Statistical Association*, **90**, 694-707.

Blei D.M., Jordan M.I., Ng A.Y. (2003). "Hierarchical Bayesian models for applications in information retrieval". *Bayesian Statistics VII*, (J.M. Bernardo et al. (Eds.)), Oxford University Press (to appear).

Copas J.R., Hilton F.J. (1990). "Record linkage: statistical models for matching computer records". *Journal of the Royal Statistical Society, A*, **153**, 287-320.

de Freitas N., Højen-Sørensen P., Jordan M.I., Russell S. (2001). "Variational MCMC". *Uncertainty in Artificial Intelligence Proceedings*, Morgan Kaufmann Publishers, 120-127.

Fellegi I.P., Sunter A.B. (1969). "A theory of record linkage". *Journal of the American Statistical Association*, **64**, 1183-1210.

Fortini M., Liseo B., Nuccitelli A., Scanu M. (2001). "On Bayesian record linkage". *Research in Official Statistics*, **4**, 185-198.

Fuller W.A. (1995). "Estimation in presence of measurement errors". *International Statistical Review*, **63**, 121-147.

Goodman L.A. (1974). "Exploratory latent structure analysis using both identifiable and unidentifiable models". *Biometrika*, **61**, 215-231.

Hagenaars J. (1993). *Loglinear models with latent variables*. Sage Publications, Newbury Park, USA.

Jabine T.B., Scheuren F.J. (1986). "Record linkages for statistical purposes: methodological issues". *Journal of Official Statistics*, **2**, 255-277.

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**, 414-420.

Kelley P. (1986). "Robustness of the Census Bureau's Record Linkage System". *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 620-624.

Larsen M., Rubin D.B. (2001). "Iterative automated record linkage using mixture models". *Journal of the American Statistical Association*, **96**, 32-41.

Newcombe H.B. (1988). *Handbook of record linkage methods for health and statistical studies, administration and business*. Oxford University Press, New York.

Saris W.E., Andrews F.M. (1991). "Evaluation of Measurement Instruments Using a Structural Modeling Approach". *Measurement Error in Surveys*, (P.P. Biemers et al. (Eds.)), Wiley, New York, 575-597.

Winkler W. (1993). "Improved decision rules in the Fellegi-Sunter model of record linkage". *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 274-279.

Winkler W. (2000). "Machine learning, information retrieval and record linkage". *Procedings of the Section on Survey Research Methods, American Statistical Association*, 20-29.

Winkler W.E., Thibaudeau Y. (1991). "An application of the Fellegi-Sunter model of record linkage to the 1990 U.S. decennial Census", *Bureau of the Census, Statistical Research Division, Statistical Research Report Series*, n. RR91/09.