Journal of Official Statistics, Vol. 27, No. 3, 2011, pp. 491–505

Bayesian Estimation of Population Size via Linkage of Multivariate Normal Data Sets

Brunero Liseo[1] *and Andrea Tancredi*[2]

We propose a Bayesian approach for matching noisy multivariate continuous vectors observed on different occasions but originating from the same closed population. The proposed methodology can be profitably adopted in record linkage and in capture-recapture problems where the size of a finite population is the main object of interest and the number of "recaptured" individuals is unknown. A Gibbs sampling scheme is used to simulate from the posterior distribution of the model parameters. The performance of the proposed approach is evaluated with simulated data sets.

Key words: Capture-recapture model; closed population; Gibbs sampling; measurement error; record linkage.

1. Introduction

Record linkage is "the name given to any process which identifies the common reporting units in two different files" (Kelley 1986). It is a powerful tool in the more general problem of multiple data set integration and it is ubiquitous in many different disciplines; among the others, medicine, business administration and official statistics (see, for instance, Newcombe (1988) or the more recent Herzog et al. (2007)). In many research projects or administrative tasks it is important to gather information about a single unit from more than one data sources. When a unique identifier – or a key – is available for all fields in every data source a deterministic linkage can be used. The deterministic record linkage procedures assume that data have been observed without errors so the linkage can only happen between records which exactly match on each field.

In practice, this situation is very uncommon and the same unit can be registered with, for example, different names and/or different values of some relevant variable, in different data sets. This implies the lack of a unique identifier and a probabilistic approach to record linkage is then necessary. The relevant literature in the field is vast and the same procedures are often given different names according to the field of application. In official statistics the role of record linkage is getting more and more important. Against the

[1] Department of Methods and Models for Economics, Territory and Finance. Sapienza Università di Roma. Viale del castro laurenziano 9, 00161, Roma, Italy. Email: brunero.liseo@uniroma1.it

[2] Department of Methods and Models for Economics, Territory and Finance. Sapienza Università di Roma. Viale del castro laurenziano 9, 00161, Roma, Italy. Email: andrea.tancredi@uniroma1.it

Acknowledgments: This work has been done with the contribution of Italian MIUR, grant 2008 CEFF37:002, Bayesian Methods for Finite Populations. The authors thank the Associate Editor and two anonymous referees for valuable comments on a previous version of this article.

© Statistics Sweden

background of a series of naive and heuristic record linkage methods, very popular in the '60s and in the '70s (see, for example, Armstrong and Mayda (1993)), Fellegi and Sunter (1969) were the first to set a record linkage problem into a formal statistical framework. Since then, there have been significant advances which are described in several important papers. Among the most influential, we mention Jaro (1989), Winkler (1993) and Belin and Rubin (1995), which made explicit the nature of the record linkage problem as a mixture model. Fortini et al. (2001; 2002) have proposed a Bayesian approach which is partially adopted in the present article.

All of these papers share the common approach of assuming a probabilistic model for the set of all possible comparisons among records from different data sets; moreover these comparisons are considered mutually independent. This assumption, as noted by Kelley (1986), is fundamentally incorrect. Using his words, ". . . *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 random variables related to different comparisons may be deterministically dependent. Consider for example, the case of a single key variable X and a 0/1 comparison function Y, that is

$$Y = \begin{cases} 1 & \text{if values on both units coincide} \\ 0 & \text{otherwise} \end{cases}$$

Let X_l^t be the observed value of the lth units of file t, $t = A, B$ and consider the first two records in data sets A and B. If

$$X_1^A = X_1^B, \quad X_1^A = X_2^B, \quad X_2^A = X_1^B$$

then, necessarily, $X_2^A = X_2^B$. Then comparison need not be independent and this problem cannot be circumvented by eliminating redundant comparisons for the likelihood function, because the order with which pairs are considered would matter!

Also, the problem of misspecification of the statistical model would bias the calculation of the error rates, as noted, for example, by Winkler (2000).

From this standpoint, we propose a model which is directly built up on the data X observed on the two occasions. In particular, we illustrate a Bayesian approach for matching noisy multivariate normal vectors observed on different occasions but originating from the same closed population. Extensions to other continuous variables are possible, losing some of the closed forms used in this article, with additional computational effort. The proposed methodology can be profitably adopted in record linkage and in capture-recapture problems where the size of a finite population is the main object of interest and the number of "recaptured" individuals is unknown.

The article is organized as follows. We continue this section by illustrating the relevant literature for our approach. In Section 2 we describe the model in terms of likelihood and prior distributions. In Section 3, a Gibbs sampling scheme is presented. Section 4 describes the performance of our approach with simulated data sets. Finally, in Section 5 the possibilities for different modelling and future research are discussed.

1.1. Relevant Literature

Most, if not all, of the previous approaches to population size estimation with matching uncertainty consider the matching step and the size estimation step as two logically and operationally well separated tasks. In this article we propose a unified framework where matching uncertainty is naturally accounted for in estimating population size. In particular we consider our proposal as a possible first answer to the issue raised by Fienberg and Manrique-Vallier (2009) who state: "*. . . it seems rather natural to ask whether there is a way to combine record linkage, covariates, and multiple system estimation methodologies using missing data framework and assumptions such as missing at random. While we have not attempted such a grand unification, we think that considering the problems in an integrated form will lead to new and improved statistical methodology. One of the main benefits we foresee in this unification is the acknowledgment and incorporation of the inherent uncertainty that probabilistic record linkage methods for merging multiple lists in a form directly suitable for multiple system estimates, which is ignored in virtually all applications*". In this article we will focus on the case where all the observed key variables are continuous. The case of categorical variables is discussed in Tancredi and Liseo (2011). In that paper, the authors consider the case where two independent random samples are drawn from a closed population generated from a superpopulation model (Ericson 1969) and a measurement error mechanism affects the sample data. While in the categorical data framework a multinomial distribution has been used as a superpopulation model and the hit-and-miss model has been used for the measurement error (Copas and Hilton 1990), in this article the population values are assumed to be random samples from a multivariate normal distribution and the sample data are affected by a normal measurement error.

The issue of linkage uncertainty in the presence of continuous variables is a serious one whose relevance goes beyond the population size estimation problems and falls into the broad category of inference problems in the presence of linkage uncertainty, deeply discussed by Judson (2007). Lahiri and Larsen (2005) provide a specific example for the case of multiple regression analysis with linked data.

Record linkage of survey or administrative data is not the only statistical problem where matching issues arise. An example which is relevant to our approach emerges in bioinformatics with the Bayesian model discussed by Green and Mardia (2006). DeGroot and Goel (1980) consider the situation where a random sample of size n, say (X_i, Z_i), $i = 1, \ldots, n$ is drawn from a bivariate normal distribution; however, before the sample values are recorded, each observation $(x_i; z_i)$ gets broken into two separate components. As a consequence, the available information assumes the form of the vector $x = (x_1, \ldots, x_n)$ and $y = (y_1, \ldots, y_n)$, where y is an unknown permutation of the original values $(z_1, \ldots, z_n)$.

A further example of matching is discussed in Lindley (1977), in a forensic scenario. Here the problem arises when some material is found at the scene of a crime and similar material is found on a suspect; in both cases material collection is generally subject to measurement error. Lindley describes a Bayesian method to establish whether the two materials are likely to come from the same source or not. Under the assumption of Gaussianity, Lindley shows that the Bayes factor comparing the alternative hypotheses that the suspect was (or was not) present at the scene of the crime, is actually the product of two components: the first one depends on the difference among the observed values, the second one depends on the distribution of the "true" values in the population. This is clearly so in that, for example, observing evidence of "blue hair" on the scene of the crime and on the suspect is much more informative than observing black or blond hair. The approach proposed in this article can be considered, in some respects, a multivariate generalization of Lindley's model.

2. The Model

Suppose we are given two data sets $\mathbf{X}^A$ and $\mathbf{X}^B$ of different sizes n^A and n^B randomly drawn from a finite population. Each record in each data set consists of observations related to h variables, denoted the *key* variables. Then,

$$\mathbf{X}^A = \left(\mathbf{x}_1^A, \ldots, \mathbf{x}_a^A, \ldots, \mathbf{x}_{n^A}^A\right) \quad \text{and} \quad \mathbf{X}^B = \left(\mathbf{x}_1^B, \ldots, \mathbf{x}_b^B, \ldots, \mathbf{x}_{n^B}^B\right),$$

and the single columns of $\mathbf{X}^A$ and $\mathbf{X}^B$ are respectively given by $\mathbf{x}_a^A = \left(x_a^{A_1}, \ldots, x_a^{A_h}\right)'$ and $\mathbf{x}_b^B = \left(x_b^{B_1}, \ldots, x_b^{B_h}\right)'$. For the sake of simplicity, the two data sets will be respectively called sample A and sample B. Conditionally on the unobservable h-dimensional vectors $\boldsymbol{\mu}_a^A = \left(\mu_a^{A_1}, \ldots, \mu_a^{A_h}\right)'$ and $\boldsymbol{\mu}_b^B = \left(\mu_b^{B_1}, \ldots, \mu_b^{B_h}\right)'$, for $a = 1, \ldots, n^A$ and $b = 1, \ldots, n^B$ and the diagonal matrix $\boldsymbol{\Gamma} = \operatorname{diag}(\gamma_1, \ldots, \gamma_h)$, we assume that vectors $\mathbf{x}_a^A$ and $\mathbf{x}_b^B$ are mutually independent for $a = 1, \ldots, n^A$ and $b = 1, \ldots, n^B$ with

$$\mathbf{x}_a^A \sim N_h\left(\boldsymbol{\mu}_a^A, \boldsymbol{\Gamma}\right) \quad \text{and} \quad \mathbf{x}_b^B \sim N_h\left(\boldsymbol{\mu}_b^B, \boldsymbol{\Gamma}\right). \tag{1}$$

The unknown quantity $\boldsymbol{\mu}_a^A\ \left(\boldsymbol{\mu}_b^B\right)$ is then the realization of a multivariate continuous variable $\boldsymbol{\mu} = (\mu^1, \ldots, \mu^h)$ over the population unit which occupies the ath (bth) position in sample A (B). The components of $\boldsymbol{\mu}$ represent the unobserved true values of the key variables. This way, assumption (1) implies that vectors $\mathbf{x}_a^A$ in sample A and $\mathbf{x}_b^B$ in sample B can be considered measurements subject to recording error of the variables $\boldsymbol{\mu}_a^A$ and $\boldsymbol{\mu}_b^B$. Also, conditionally on the unobserved true values of $\boldsymbol{\mu}_a^A$, $\boldsymbol{\mu}_b^B$ and $\boldsymbol{\Gamma}$, the recording errors are independent across the samples, across the observations of each sample and also between the components of each observation, since $\boldsymbol{\Gamma}$ is diagonal.

We now describe the distributional assumption about $\boldsymbol{\mu}^A = \left(\boldsymbol{\mu}_1^A, \ldots, \boldsymbol{\mu}_{n^A}^A\right)'$ and $\boldsymbol{\mu}^B = \left(\boldsymbol{\mu}_1^B, \ldots, \boldsymbol{\mu}_{n^B}^B\right)'$. We assume that both in sample A and sample B, the sampled units represent a simple random sample without replacement drawn from a finite population. Moreover, the unknown values of the variable $\boldsymbol{\mu}$ are assumed to be independently and identically distributed with an h-dimensional Normal $N_h(\boldsymbol{\theta}, \Sigma)$, in the spirit of the Bayesian superpopulation model proposed by Ericson (1969). This way one has $\boldsymbol{\mu}_a^A \sim N_h(\boldsymbol{\theta}, \boldsymbol{\Sigma})$ independently for $a = 1, \ldots n^A$ and $\boldsymbol{\mu}_b^B \sim N_h(\boldsymbol{\theta}, \boldsymbol{\Sigma})$ independently for $b = 1, \ldots n^B$.

However, one cannot state that $\boldsymbol{\mu}^A$ and $\boldsymbol{\mu}^B$ are independent. In fact, after "observing" values $\left(\boldsymbol{\mu}_1^A, \ldots, \boldsymbol{\mu}_{n^A}^A\right)$ in sample A, we know that these values are part of the population values and they have a positive probability of being observed again in sample B.

Then we need to explicitly model the dependence structure between $\boldsymbol{\mu}^A$ and $\boldsymbol{\mu}^B$ and we will make use of the latent matching matrix $\mathbf{C}$. This is an $n^A \times n^B$ matrix whose generic element C_{ab} indicates whether or not unit a in sample A and unit b in sample B are the same population unit, that is, for $a = 1, \ldots, n^A$ and $b = 1, \ldots, n^B$,

$$C_{ab} = \begin{cases} 1, & \text{if } a \text{ and } b \text{ refer to the same population unit} \\ 0, & \text{otherwise} \end{cases}$$

The matrix $\mathbf{C}$ is the actual quantity of interest in record linkage problems; a similar structure also appears in different statistical problems, such as Bayesian alignment (Green and Mardia 2006) or microarrays analysis (Do et al. 2005). We assume that multiple matches are not possible; then $\sum_a C_{ab} = C_{.b} \leq 1$, $\forall b = 1, \ldots, n^B$, $\sum_b C_{ab} = C_{a.} \leq 1$, $\forall a = 1, \ldots, n^A$; also, note that there are $\binom{n^A}{t}\binom{n^B}{t} t!$ different $\mathbf{C}$ matrices with exactly $t = \sum_{ab} C_{ab}$ matches, for all $t \leq \min(n^A, n^B)$. Given the values of the matching matrix $\mathbf{C}$, vectors $\boldsymbol{\mu}_a^A$ and $\boldsymbol{\mu}_b^B$ corresponding to the same unit $(C_{ab} = 1)$ assume identical values. Also, both $\boldsymbol{\mu}_a^A$ and $\boldsymbol{\mu}_b^B$ have marginal h-dimensional Normal distribution with mean $\boldsymbol{\theta}$ and covariance matrix $\boldsymbol{\Sigma}$. Setting

$$\boldsymbol{\mu}_{ab}^{AB} = \begin{bmatrix} \boldsymbol{\mu}_a^A \\ \boldsymbol{\mu}_b^B \end{bmatrix}; \ \boldsymbol{\theta}_2 = \begin{bmatrix} \boldsymbol{\theta} \\ \boldsymbol{\theta} \end{bmatrix} \quad \text{and} \quad \boldsymbol{\Sigma}_2 = \begin{bmatrix} \boldsymbol{\Sigma} & \boldsymbol{\Sigma} \\ \boldsymbol{\Sigma} & \boldsymbol{\Sigma} \end{bmatrix}$$

it follows that, when $C_{ab} = 1$,

$$\boldsymbol{\mu}_{ab}^{AB} \sim N_{2h}(\boldsymbol{\theta}_2, \boldsymbol{\Sigma}_2).$$

Finally, the dependencies among the components of $\boldsymbol{\mu}^A$ and $\boldsymbol{\mu}^B$ given $\mathbf{C}$ are restricted to the matched pairs. More precisely,

$$p(\boldsymbol{\mu}^A, \boldsymbol{\mu}^B | \mathbf{C}, \boldsymbol{\theta}, \boldsymbol{\Sigma}) = \prod_{a:C_{a.}=0} \phi_h(\boldsymbol{\mu}_a^A | \boldsymbol{\theta}, \boldsymbol{\Sigma}) \times \prod_{b:C_{.b}=0} \phi_h(\boldsymbol{\mu}_b^B | \boldsymbol{\theta}, \boldsymbol{\Sigma}) \times \prod_{ab:C_{ab}=1} \phi_{2h}(\boldsymbol{\mu}_{ab}^{AB} | \boldsymbol{\theta}_2, \boldsymbol{\Sigma}_2)$$

where $\phi_k(\cdot|\xi, \Omega)$ is the density of a k-dimensional multivariate Normal distribution with mean ξ and covariance matrix Ω.

The prior distribution for $\mathbf{C}$ should reflect the random selection mechanism of the two samples. Conditionally on $t = \sum_{ab} C_{ab}$, $\mathbf{C}$ has a uniform distribution on the set of all possible matching matrices with exactly t matches. Loosely speaking, in the absence of any information, all the possible couples are equally likely to be a match. Then, we have $p(\mathbf{C}|\boldsymbol{\theta}, \boldsymbol{\Sigma}) = p(\mathbf{C}|t)p(t|N)$ with

$$p(\mathbf{C}|t) = \begin{cases} 0 & \textit{if } \sum_{ab} C_{ab} \neq t \\ \left[\binom{n^A}{t}\binom{n^B}{t} t!\right]^{-1} & \text{otherwise} \end{cases}.$$

Finally, the total number t of common units across the two samples has a scalar hypergeometric distribution, that is,

$$p(t|N) = \binom{n^A}{t}\binom{N - n^A}{n^B - t} \Big/ \binom{N}{n^B} \tag{2}$$

It is easy to check that, by averaging out over $\mathbf{C}$ in the distribution $p(\boldsymbol{\mu}^A, \boldsymbol{\mu}^B, \mathbf{C}|\boldsymbol{\theta}, \boldsymbol{\Sigma})$, one reobtains that $\boldsymbol{\mu}^A$ and $\boldsymbol{\mu}^B$ are two random samples from a $N(\boldsymbol{\theta}, \boldsymbol{\Sigma})$.

An important feature of the model is that one can easily obtain the distribution of $\mathbf{X}^A$ and $\mathbf{X}^B$ given $\mathbf{C}$, $\boldsymbol{\theta}$, $\boldsymbol{\Sigma}$, $\boldsymbol{\Gamma}$. In fact, using standard results in multivariate Normal theory (see, for example Davison (2003), p. 456),

i) if $\mathbf{x}|\boldsymbol{\mu} \sim N_h(\boldsymbol{\mu}, \boldsymbol{\Gamma})$ and $\boldsymbol{\mu} \sim N_h(\boldsymbol{\theta}, \boldsymbol{\Sigma})$, then $\mathbf{x} \sim N_h(\boldsymbol{\theta}, \boldsymbol{\Gamma} + \boldsymbol{\Sigma})$

ii) if $\mathbf{x}|\boldsymbol{\mu}_2 \sim N_{2h}(\boldsymbol{\mu}_2, \boldsymbol{\Gamma}_2)$ and $\boldsymbol{\mu}_2 \sim N_{2h}(\boldsymbol{\theta}_2, \boldsymbol{\Sigma}_2)$ with

$$\boldsymbol{\mu}_2 = \begin{bmatrix} \boldsymbol{\mu} \\ \boldsymbol{\mu} \end{bmatrix} \quad \boldsymbol{\Gamma}_2 = \begin{bmatrix} \boldsymbol{\Gamma} & 0 \\ 0 & \boldsymbol{\Gamma} \end{bmatrix}$$

then $\mathbf{x} \sim N_{2h}\left(\boldsymbol{\theta}_2, \boldsymbol{\Psi}\right)$ with

$$\boldsymbol{\Psi} = \begin{bmatrix} \boldsymbol{\Sigma} + \boldsymbol{\Gamma} & \boldsymbol{\Sigma} \\ \boldsymbol{\Sigma} & \boldsymbol{\Sigma} + \boldsymbol{\Gamma} \end{bmatrix}.$$

This, in turns, implies that

$$p\left(\boldsymbol{X}^A, \boldsymbol{X}^B | \mathrm{C}, \boldsymbol{\theta}, \boldsymbol{\Sigma}, \boldsymbol{\Gamma}\right) = \prod_{a:C_{a\cdot}=0} \phi_h\left(\boldsymbol{x}_a^A | \boldsymbol{\theta}, \boldsymbol{\Sigma} + \boldsymbol{\Gamma}\right) \times \prod_{b:C_{\cdot b}=0} \phi_h\left(\boldsymbol{x}_b^B | \boldsymbol{\theta}, \boldsymbol{\Sigma} + \boldsymbol{\Gamma}\right)$$
$$\times \prod_{ab:C_{ab}=1} \phi_{2h}\left(\boldsymbol{x}_{ab}^{AB} | \boldsymbol{\theta}_2, \boldsymbol{\Psi}\right)$$

where $\boldsymbol{x}_{ab}^{AB} = \left(\left(\boldsymbol{x}_a^A\right)', \left(\boldsymbol{x}_b^B\right)'\right)'$.

We conclude this section by describing the prior assumptions about the other parameters of the model, namely N, $\boldsymbol{\theta}$, $\boldsymbol{\Sigma}$, $\boldsymbol{\Gamma}$. All these quantities are assumed a priori independent of each other; we also assume that, in practical situations, the information available on these parameters will be, at best, rather weak. In particular, we assume that $p(\boldsymbol{\theta}) \propto 1$; $p(N) \propto 1/N^2$ truncated on $\{1, N^*\}$, N^* being a reasonably large value of N, a diffuse inverse Wishart distribution for $\boldsymbol{\Sigma}$ and diffuse independent Gamma distributions for the elements of $\boldsymbol{\Gamma}$.

3. Bayesian Implementation

In this section we describe in detail the practical implementation of the proposed model. We will show how to produce a posterior sample from $p(\mathbf{C}, \boldsymbol{\theta}, \boldsymbol{\Sigma}, \boldsymbol{\Gamma}, N | \boldsymbol{X}^A, \boldsymbol{X}^B)$ using a Metropolis-within-Gibbs algorithm. The reader may refer to Robert and Casella (2004) for a general introduction to Markov Chain Monte Carlo methods.

For the updating step of the matrix $\mathbf{C}$ we adapt the algorithm proposed by Green and Mardia (2006) to our setup. In particular $\mathbf{C}$ is updated via single Metropolis-Hastings moves that, when accepted, can only increase (decrease) the total number of matches t by 1. Let $q(\mathbf{C}'|\mathbf{C})$ be the proposal distribution for the $\mathbf{C}$ move, that is the probability of proposing a new matrix $\mathbf{C}'$ given that the chain is in $\mathbf{C}$. The list of all possible moves from the matrix C to a new proposed matrix $\mathbf{C}'$ is listed below;

(a) adding a match, that is changing one entry C_{ab} from 0 to 1;
(b) deleting a match, that is changing one entry C_{ab} from 1 to 0;

(c) switching a match, that is changing, at the same time, one entry from 0 to 1 and another entry – in the same row or column – from 1 to 0.

In detail the algorithm proceeds as follows. First, a row or a column is randomly selected. Suppose, without loss of generality, we select row a: either the row already has an entry equal to 1 (i.e. $C_{ab}=1$, for some b) or, alternatively, $C_{ab}=0\ \forall b$. In the former case, then, with probability p* we propose the deletion of the match, i.e. $C'_{ab}=0$ and, with probability $1-p$* we propose a random switch to another entry of the row. In the latter case, when there are no matches in row a, the proposal distribution randomly chooses a b value among the nonmatched units of file B. This way, the acceptance probability,

$$\min\left\{1, \frac{p\big(\mathbf{C}', \boldsymbol{\theta}, \boldsymbol{\Sigma}, \boldsymbol{\Gamma}, N|\boldsymbol{X}^A, \boldsymbol{X}^B\big)q\big(\mathbf{C}|\mathbf{C}'\big)}{p\big(\mathbf{C}, \boldsymbol{\theta}, \boldsymbol{\Sigma}, \boldsymbol{\Gamma}, N|\boldsymbol{X}^A, \boldsymbol{X}^B\big)q\big(\mathbf{C}'|\mathbf{C}\big)}\right\}$$

in the case of a new match $C'_{ab}=1$, is equal to

$$\min\left\{1, \frac{\phi_{2h}\big(\boldsymbol{x}_{ab}^{AB}|\boldsymbol{\theta}_2, \boldsymbol{\Psi}\big)(n^B-t)p^*}{\phi_h\big(\boldsymbol{x}_a^A|\boldsymbol{\theta}, \boldsymbol{\Sigma}+\boldsymbol{\Gamma}\big)\phi_h\big(\boldsymbol{x}_b^B|\boldsymbol{\theta}, \boldsymbol{\Sigma}+\boldsymbol{\Gamma}\big)(N-n^A-n^B+(t+1))}\right\}. \tag{3}$$

Analogously, the acceptance probability for a switching proposal from (a, b) to (a, b') is

$$\min\left\{1, \frac{\phi_{2h}\big(\boldsymbol{x}_{ab'}^{AB}|\boldsymbol{\theta}_2, \boldsymbol{\Psi}\big)\phi_h\big(\boldsymbol{x}_b^B|\boldsymbol{\theta}, \boldsymbol{\Sigma}+\boldsymbol{\Gamma}\big)}{\phi_{2h}\big(\boldsymbol{x}_{ab}^{AB}|\boldsymbol{\theta}_2, \boldsymbol{\Psi}\big)\phi_h\big(\boldsymbol{x}_{b'}^B|\boldsymbol{\theta}, \boldsymbol{\Sigma}+\boldsymbol{\Gamma}\big)}\right\};$$

finally, for the deleting move $C'_{ab}=0$ the required probability is

$$\min\left\{1, \frac{\phi_h\big(\boldsymbol{x}_a^A|\boldsymbol{\theta}, \boldsymbol{\Sigma}+\boldsymbol{\Gamma}\big)\phi_h\big(\boldsymbol{x}_b^B|\boldsymbol{\theta}, \boldsymbol{\Sigma}+\boldsymbol{\Gamma}\big)\big(N-n^A-n^B+t\big)}{\phi_{2h}\big(\boldsymbol{x}_{ab}^{AB}|\boldsymbol{\theta}_2, \boldsymbol{\Psi}\big)\big(n^B-t+1\big)p^*}\right\}. \tag{4}$$

The mixing of the Markov chain is improved by proposing several updatings of the matrix $\mathbf{C}$, for each updating cycle of the other parameters ($\boldsymbol{\theta}$, $\boldsymbol{\Sigma}$, $\boldsymbol{\Gamma}$, N).

The conditional distribution of $\boldsymbol{\theta}$ is available in closed form. In fact, setting $\bar{\boldsymbol{x}}_{ab}^{AB}=\big(\boldsymbol{x}_a^A+\boldsymbol{x}_b^B\big)/2$ and $\boldsymbol{d}_{ab}^{AB}=\big(\boldsymbol{x}_a^A-\boldsymbol{x}_b^B\big)$, if $\boldsymbol{x}_{ab}^{AB}\sim N_{2h}(\boldsymbol{\theta}_2, \boldsymbol{\Psi})$ it follows that $\bar{\boldsymbol{x}}_{ab}^{AB}$ and $\boldsymbol{d}_{ab}^{AB}$ are independent and the Jacobian of the transformation is equal to 1. Then,

$$\phi_{2h}\big(\boldsymbol{x}_{ab}^{AB}|\boldsymbol{\theta}_2, \boldsymbol{\Psi}\big)=\phi_h\big(\bar{\boldsymbol{x}}_{ab}^{AB}|\boldsymbol{\theta}, \boldsymbol{\Sigma}+\boldsymbol{\Gamma}/2\big)\phi_h\big(\boldsymbol{d}_{ab}^{AB}|\mathbf{0}, 2\boldsymbol{\Gamma}\big) \tag{5}$$

and the conditional distribution of $\boldsymbol{\theta}$ is

$$\begin{aligned} p\big(\boldsymbol{\theta}|\boldsymbol{X}^A, \boldsymbol{X}^B\mathbf{C}, \boldsymbol{\Sigma}, \boldsymbol{\Gamma}, N\big) &\propto \prod_{a:C_{a\cdot}=0}\phi_h\big(\boldsymbol{x}_a^A|\boldsymbol{\theta}, \boldsymbol{\Sigma}+\boldsymbol{\Gamma}\big)\times\prod_{b:C_{\cdot b}=0}\phi_h\big(\boldsymbol{x}_b^B|\boldsymbol{\theta}, \boldsymbol{\Sigma}+\boldsymbol{\Gamma}\big)\\ &\quad\times\prod_{ab:C_{ab}=1}\phi_h\big(\bar{\boldsymbol{x}}_{ab}^{AB}|\boldsymbol{\theta}, \boldsymbol{\Sigma}+\boldsymbol{\Gamma}/2\big). \end{aligned}$$

Standard Bayesian calculations show that the conditional distribution of $\boldsymbol{\theta}$ is

N_h ($\boldsymbol{b}^*$, $\boldsymbol{B}^*$) with

$$\boldsymbol{B}^* = \left[t\left(\boldsymbol{\Sigma} + \boldsymbol{\Gamma}/2\right)^{-1} + \left(n^A + n^B - 2t\right)\left(\boldsymbol{\Sigma} + \boldsymbol{\Gamma}\right)^{-1}\right]^{-1}$$

and

$$\boldsymbol{b}^* = \boldsymbol{B}^*\left[t(\boldsymbol{\Sigma} + \boldsymbol{\Gamma}/2)^{-1}\bar{\bar{\boldsymbol{x}}}^{AB}_{\mathbf{C}=1} + (n^A + n^B - 2t)(\boldsymbol{\Sigma} + \boldsymbol{\Gamma})^{-1}\bar{\boldsymbol{x}}^{AB}_{\mathbf{C}=0}\right]$$

where $\bar{\bar{\boldsymbol{x}}}^{AB}_{\mathbf{C}=1} = \left[\sum_{ab:C_{ab}=1}\bar{\boldsymbol{x}}^{AB}_{ab}\right]/t$ and $\bar{\boldsymbol{x}}^{AB}_{\mathbf{C}=0} = \left[\sum_{a:C_{a.}=0}\boldsymbol{x}^A_a + \sum_{b:C_{.b}=0}\boldsymbol{x}^B_b\right]/(n^A + n^B - 2t)$

Incidentally, we notice that the acceptance probabilities for the matching matrix updating (3) and (4) reveal the close connection between our approach and Lindley's paper (Lindley 1977). For example, the acceptance probability (3) exactly corresponds to the Bayes factor for the hypothesis $C_{ab} = 1$ versus $C_{ab} = 0$, when all the other model parameters are known. Also, from (5), one can see that the evidence in favor of $C_{ab} = 1$ increases either when the distance $\boldsymbol{d}^{AB}_{ab}$ approaches zero or when the observed data $\boldsymbol{x}^A_a$ and $\boldsymbol{x}^B_b$ are far from their mean values $\boldsymbol{\theta}$: this last observation is a sort of generalization of Lindley's results (see also Davison (2003), p. 584).

The conditional distributions of $\boldsymbol{\Sigma}$ and $\boldsymbol{\Gamma}$ are not available in closed form. For both of them we will use a Metropolis step with a random walk proposal. The full conditional of $\boldsymbol{\Sigma}$ is

$$\begin{aligned}p(\boldsymbol{\Sigma}|\boldsymbol{X}^A,\boldsymbol{X}^B\mathbf{C},\boldsymbol{\theta},\boldsymbol{\Gamma},N) &\propto \prod_{a:C_{a.}=0}\phi_h(\boldsymbol{x}^A_a|\boldsymbol{\theta},\boldsymbol{\Sigma}+\boldsymbol{\Gamma}) \times \prod_{b:C_{.b}=0}\phi_h\left(\boldsymbol{x}^B_b|\boldsymbol{\theta},\boldsymbol{\Sigma}+\boldsymbol{\Gamma}\right)\\ &\times\prod_{ab:C_{ab}=1}\phi_h(\bar{\boldsymbol{x}}^{AB}_{ab}|\boldsymbol{\theta},\boldsymbol{\Sigma}+\boldsymbol{\Gamma}/2)p(\boldsymbol{\Sigma})\\ &\propto|\boldsymbol{\Sigma}+\boldsymbol{\Gamma}|^{-(n^A+n^B-2t)/2}\exp\left(-(n^A+n^B-2t)\,\mathrm{tr}\left((\boldsymbol{\Sigma}+\boldsymbol{\Gamma})^{-1}\boldsymbol{S}^{AB}_{\mathbf{C}=0}\right)/2\right)\\ &\times|\boldsymbol{\Sigma}+\boldsymbol{\Gamma}/2|^{-t/2}\exp\left(-t\ \mathrm{tr}\left((\boldsymbol{\Sigma}+\boldsymbol{\Gamma}/2)^{-1}\boldsymbol{S}^{AB}_{\mathbf{C}=1}\right)/2\right)p(\boldsymbol{\Sigma})\end{aligned}$$

with

$$\boldsymbol{S}^{AB}_{\mathbf{C}=0} = \left[\sum_{a:C_{a.=0}}\left(\boldsymbol{x}^A_a - \boldsymbol{\theta}\right)\left(\boldsymbol{x}^A_a - \boldsymbol{\theta}\right)' + \sum_{b:C_{.b=0}}\left(\boldsymbol{x}^B_b - \boldsymbol{\theta}\right)\left(\boldsymbol{x}^B_b - \boldsymbol{\theta}\right)'\right]\Big/(n^A + n^B - 2t)$$

and

$$\boldsymbol{S}^{AB}_{\mathbf{C}=1} = \sum_{ab:C_{ab}=1}\left(\bar{\boldsymbol{x}}^{AB}_{ab} - \boldsymbol{\theta}\right)\left(\bar{\boldsymbol{x}}^{AB}_{ab} - \boldsymbol{\theta}\right)'\Big/t$$

To update $\boldsymbol{\Sigma}$ we propose a draw $\boldsymbol{\Sigma}'$ from a Wishart distribution with mean equal to the current value of $\boldsymbol{\Sigma}$.

Similarly, the conditional distribution of $\mathbf{\Gamma}$ is given by

$$p(\mathbf{\Gamma}|\boldsymbol{X}^A,\boldsymbol{X}^B\mathbf{C},\boldsymbol{\theta},\mathbf{\Sigma},N)\propto|\mathbf{\Sigma}+\mathbf{\Gamma}|^{-(n^A+n^B-2t)/2}\exp\left(-(n^A+n^B-2t)\mathrm{tr}\left((\mathbf{\Sigma}+\mathbf{\Gamma})^{-1}\boldsymbol{S}^{AB}_{\mathbf{C}=0}\right)/2\right)$$

$$\times|\mathbf{\Sigma}+\mathbf{\Gamma}/2|^{-t/2}\exp\left(-t\,\mathrm{tr}\left((\mathbf{\Sigma}+\mathbf{\Gamma}/2)^{-1}\boldsymbol{S}^{AB}_{\mathbf{C}=1}\right)/2\right)$$

$$\times|\mathbf{\Gamma}|^{-t/2}\exp\left(-t\,\mathrm{tr}\left((2\mathbf{\Gamma})^{-1}\boldsymbol{D}^{AB}_{\mathbf{C}=1}\right)/2\right)p(\mathbf{\Gamma})$$

where $\boldsymbol{D}^{AB}_{\mathbf{C}=1}=\sum_{ab:C_{ab}=1}\boldsymbol{d}^{AB}_{ab}\boldsymbol{d}^{AB'}_{ab}/t$. We sequentially update each element γ_j via a random walk Metropolis-Hastings step: the new values are proposed from a Gamma distribution with mean equal to the current value of the chain γ_j. Both for $\mathbf{\Sigma}$ and $\mathbf{\Gamma}$ the shape parameters of the proposal distributions have been tuned up in terms of acceptance probability.

Finally, the conditional distribution of N is given by

$$p(N|\boldsymbol{X}^A,\boldsymbol{X}^B,\mathbf{C},\boldsymbol{\theta},\mathbf{\Sigma},\mathbf{\Gamma})\propto\binom{n^A}{t}\binom{N-n^A}{n^B-t}\Big/\left[N^2\binom{N}{n^B}\right]$$

restricted at the set {max $(n^A, n^B)+1$, N*}: in this case, it is easy to implement a Gibbs step.

4. Evaluating the Method

In this section we illustrate the performance of the proposed method via simulations. First, we discuss in detail the posterior analysis relative to a single simulation. Then we look at the frequentist properties of the Bayesian estimator for the population size N, under different scenarios.

We have drawn, without replacement, two samples of size $n^A = n^B = 30$, from a finite population with $N = 100$ units. The true population values are independent draws from a three-dimensional Gaussian random vector $N_3(\boldsymbol{\theta}, \mathbf{\Sigma})$ with $\boldsymbol{\theta}$ and $\mathbf{\Sigma}$ given by

$$\boldsymbol{\theta}=\begin{bmatrix}-10\\0\\10\end{bmatrix}\quad\mathbf{\Sigma}=\begin{bmatrix}10&8&1\\8&10&4\\1&4&10\end{bmatrix}. \tag{6}$$

Measurement errors in the samples are introduced via the covariance matrix – see Formula (1) $-$ Diag $(\mathbf{\Gamma})=(0.5,0.5,0.5)$.

In Figure 1 we show the observed values in the two samples. Also, solid and dashed lines indicate the $T = 14$ common units between the two samples. Those pairs can have been classified either as true matches or as false not-matches on the basis of the matching decision rule. To illustrate this, suppose we decide to declare a pair to be a match only when the posterior probability $P(C_{ab}=1|\boldsymbol{X}^A,\boldsymbol{X}^B)>0.15$ (see Tancredi and Liseo (2011) for a discussion about optimal decision rules for estimating the matching matrix $\mathbf{C}$). Using this decision rule we have found nine true matches which are represented, in Figure 1, by solid segments. It is important to stress that the performance of the method strongly depends on the number of key variables. With a

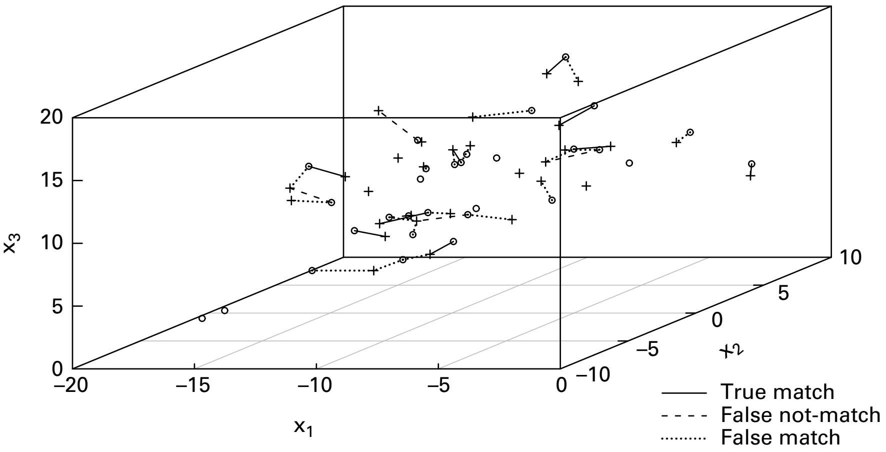

Fig. 1. Simulated data sets. Data in A are represented by circles, data in B by crosses. Pairs with a posterior probability $P(C_{ab} = 1|X^A, X^B) > 0.15$ are declared matches

larger dimension of key variables we have observed higher true match rates and lower false not-match rates.

Posterior analysis is based on an MCMC sample of size 10^5 with a burn-in period equal to 5×10^3. The prior distribution for *N* has been truncated on the set [0, 5000]: the results were not sensitive to the arbitrary upper bound, since it had never been touched during the

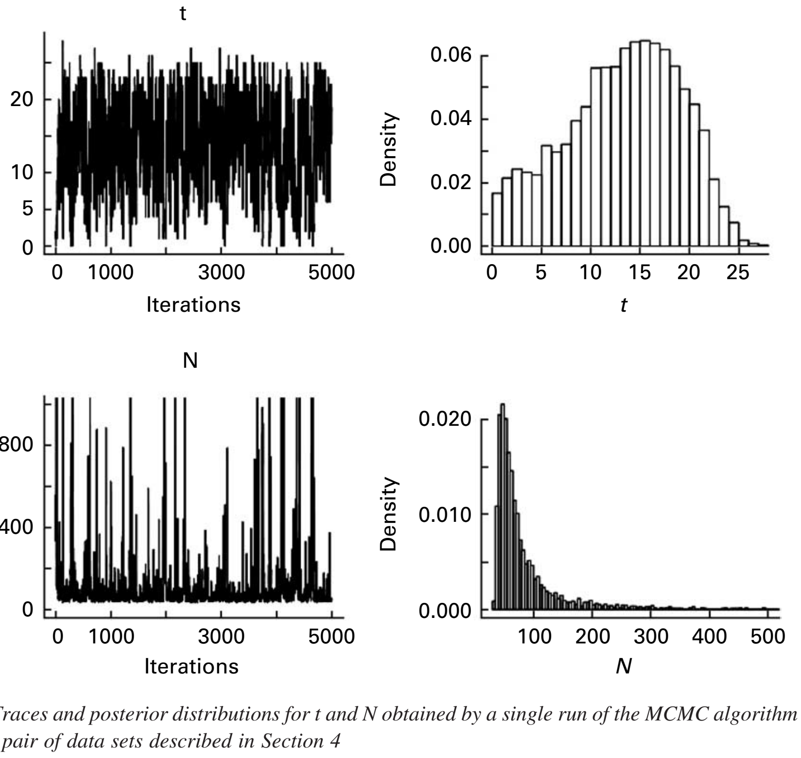

Fig. 2. Traces and posterior distributions for t and N obtained by a single run of the MCMC algorithm with the simulated pair of data sets described in Section 4

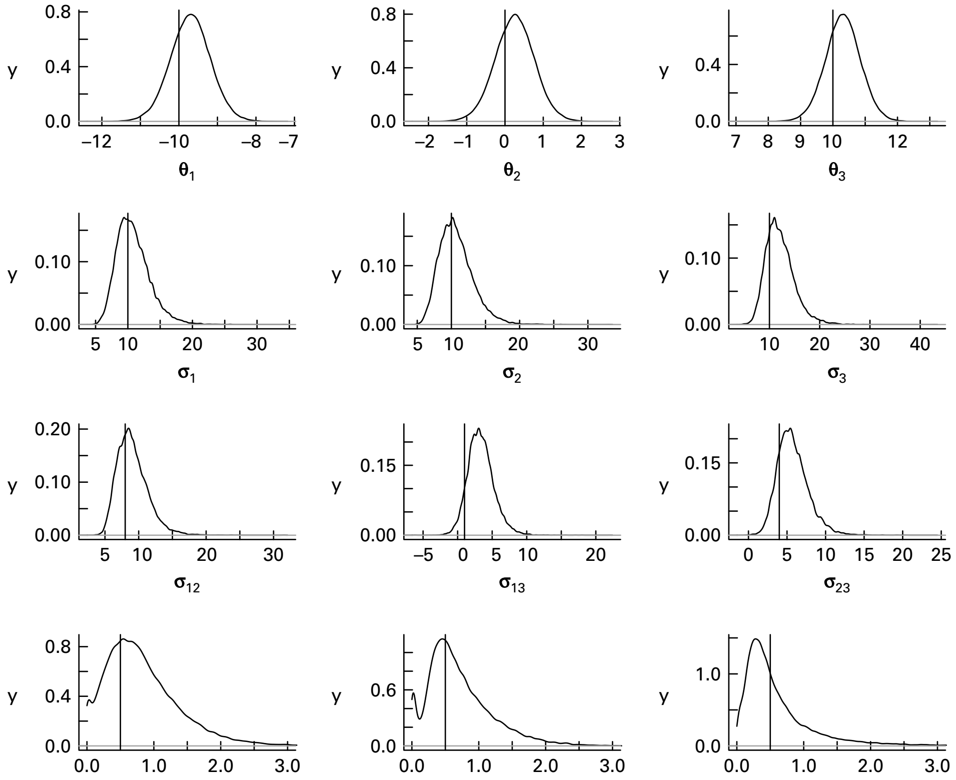

Fig. 3. Posterior densities for $\boldsymbol{\theta}$, $\boldsymbol{\Sigma}$ *and* $\boldsymbol{\Gamma}$ *obtained by a single run of the MCMC algorithm with the simulated pair of data sets described Section 4. The vertical lines indicate the true values of the parameters*

simulation. Figure 2 shows traces of the simulations for the main quantities of interest, namely t and N; to increase the mixing of the chain we consider one draw every 20 iterations. Convergence has been reached quite soon, even starting from a matching matrix $\boldsymbol{C}$ with all entries set to 0; also, repeating the simulation with different starting values we obtained identical results. Figure 2 also shows the posterior distributions of t and N; Figure 3 shows the posterior distributions of all the remaining parameters in the model. One can notice that the true values of the parameters always lie in regions of high posterior density.

4.1. Simulation Study

In this section we illustrate the performance of our method via a large-scale simulation study. We have considered three different frequentist scenarios. In the first scenario, whose results are reported in Table 1, we have replicated 100 times the simulation described above; more precisely, we have set $N = 100$, $\boldsymbol{\theta}$, $\boldsymbol{\Sigma}$, as in (6) and $\boldsymbol{\Gamma} = \text{diag}\,(0.5, 0.5, 0.5)$. We have considered different sample sizes n^A and n^B. Specifically, we have considered the values $n^A = n^B = 20, 50, 80$. For each simulated couple of data sets we have used the MCMC algorithm for approximating the posterior distribution of the

Table 1. Frequentist evaluation of the posterior estimates for N. Each row summarises data obtained with 100 simulations with n^A *and* n^B *given in the first column,* $N = 100$, $\boldsymbol{\theta}$ *and* $\boldsymbol{\Sigma}$ *as in (6) and* $\boldsymbol{\Gamma} = \text{diag}\ (0.5, 0.5, 0.5)$

$n^A = n^B$	$E(N\|\boldsymbol{X^A}, \boldsymbol{X^B})$	Med $(N\|\boldsymbol{X^A}, \boldsymbol{X^B})$	95% interval coverage	95% interval length
20	183	111	0.99	811
50	148	117	0.98	351
80	106	103	0.97	54

model parameters. The second simulation scenario, reported in Table 2, differs from the previous one mainly in terms of a larger amount of measurement error. In fact we have set $\boldsymbol{\Gamma} = \text{diag}\ (2.5, 2.5, 2.5)$. In the last scenario, reported in Table 3, we have simulated population data values from a 6-dimensional Normal distribution whose first three components are independent of the last three ones. Also the two groups of key variables shared the same mean vector $\boldsymbol{\theta}$ and the same covariance matrix $\boldsymbol{\Sigma}$, again given by (6).

Although our method can actually be used also for record linkage purposes, here we report the simulation results having the population size N as our primary parameter of interest and we have mainly focused on the marginal posterior distribution of N. In each table we reported the empirical mean of $E(N|\boldsymbol{X}^{\mathbf{A}}, \boldsymbol{X}^{\mathbf{B}})$, the empirical mean of the posterior median Med $(N|\boldsymbol{X}^{\mathbf{A}}, \boldsymbol{X}^{\mathbf{B}})$, the coverage of the 95% credible intervals and their mean length as well.

As one might expect, the larger the measurement error is, the more difficult estimating N will be. On the other hand, increasing the data information – that is, using a large number of key variables – will produce more accurate estimates of N; one may also notice that posterior estimates improve as n^A and n^B get larger.

5. Discussion

Record linkage techniques and population size estimation pose several interesting problems both from the methodological and the computational viewpoint. In particular, from a methodological perspective, the definition itself of a correct statistical model for representing comparisons among records is still debated.

In this article we have focused on the problem of the estimation of the size of a closed population when two surveys are available and continuous variables have been recorded. This situation is not uncommon and, to our knowledge, there are no well-established methods available in the literature to tackle this problem.

Table 2. Frequentist evaluation of the posterior estimates for N. Each entry is the empirical mean obtained with 100 simulations with n^A *and* n^B *given in the first column,* $N = 100$, $\boldsymbol{\theta}$ *and* $\boldsymbol{\Sigma}$ *as in (6) and* $\boldsymbol{\Gamma} = \text{diag}\ (2.5, 2.5, 2.5)$

$n^A = n^B$	$E(N\|\boldsymbol{X^A}, \boldsymbol{X^B})$	Med $(N\|\boldsymbol{X^A}, \boldsymbol{X^B})$	95% interval coverage	95% interval length
20	150	75	0.99	803
50	213	130	0.99	894
80	159	116	0.98	448

Table 3. Frequentist evaluation of the posterior estimates for N. Each entry is the empirical mean obtained with 100 simulations with n^A *and* n^B *given in the first column,* N = *100. True values were drawn from a multivariate Normal with six components. The first three components are independent of the other omponents. The two blocks of variables have mean vector and covariance matrices as in (6); measurement error is modeled via the matrix* $\boldsymbol{\Gamma}$ *= diag (0.5, 0.5, 0.5, 0.5, 0.5, 0.5)*

$n^A = n^B$	$E(N \vert \boldsymbol{X^A}, \boldsymbol{X^B})$	Med $(N \vert \boldsymbol{X^A}, \boldsymbol{X^B})$	95% interval coverage	95% interval length
20	165	119	0.97	554
50	213	130	0.96	60
80	101	100	0.95	16

Trying to figure out some possible alternative ways, we have implemented the Jaro approach for linking files, after an arbitrary, although unavoidable, discretization of the key variables. Then after selecting a given number of matches in the two databases, a conditional likelihood function can be written for the quantity of interest, *N*; this likelihood can be either used alone or combined with a prior to get an alternative Bayesian estimate. This approach, albeit reasonable, indeed fails to account for matching uncertainty and, consequently, the standard error of the likelihood estimator of *N* turns out to be underreported. Results of a small-scale simulation are reported in Table 4.

The model proposed in this article adopts an integrated approach, already discussed by Tancredi and Liseo (2011) in the case of categorical key variables. What we consider the major novelties of our proposal are these:

- The statistical model is built up on the data actually observed: no reduction of the available information to 0/1 comparisons is introduced.
- The model is able to account for matching uncertainty in the estimation of *N*. This point is rather important because it allows a more correct report of the "standard error" of the estimates.

This article represents one of the first attempts to deal with the problem of linking files in the presence of continuous variables: many improvements and extensions can be developed. Our method can be used with any kind of continuous distribution; admittedly,

Table 4. Frequentist evaluation of the posterior estimates for N *obtained via Jaro's approach, after discretization of each key variable into* h *classes. Each entry is the mean relative to 100 simulations from our model. Values of* h, n^A *and* n^B *are given in the first two columns,* N = *100,* $\boldsymbol{\theta}$ *and* $\boldsymbol{\Sigma}$ *are given by (6) and* $\boldsymbol{\Gamma}$ *= diag (0.5, 0.5, 0.5)*

h	$n^A = n^B$	$E(N \vert y)$	95% interval length	95% interval coverage
10	20	29	11	0.02
	50	59	6	0.02
	80	87	3	0.03
20	20	48	58	0.10
	50	48	58	0.10
	80	88	4	0.14
30	20	47	52	0.14
	50	74	30	0.14
	80	92	6	0.10

the Gaussian assumption makes computation much easier to perform and the results easier to understand. We are currently working on an extension to skewed and heavy-tailed distributions. Other possible generalizations are related to sampling. One could allow for different sampling schemes and/or different sampling probabilities among units. In some other situations, the two databases cannot be considered independent and/or the sample size should be considered random. All these situations can be easily framed into our model at a low computational cost. We are currently working on the more challenging problem of the linkage of more than two lists.

6. References

Armstrong, J. and Mayda, J. (1993). Model-Based Estimation of Record Linkage Error Rates. Journal of the American Statistical Association, 88, 137–147.

Belin, T. and Rubin, D. (1995). A Method for Calibrating False–Match Rates in Record Linkage. Journal of the American Statistical Association, 90, 694–707.

Copas, J. and Hilton, F. (1990). Record Linkage: Statistical Models for Matching Computer Records. Journal of the Royal Statistical Society, Series A, 153, 287–320.

Davison, A.C. (2003). Statistical Models. Cambridge, UK: Cambrige University Press.

DeGroot, M.H. and Goel, P. (1980). Estimation of the Correlation Coefficient from a Broken Random Sample. The Annals of Statistics, 8, 264–278.

Do, K.A., Mueller, P., and Tang, F. (2005). A Bayesian Mixture Model for Differential Gene Expression. Journal of the Royal Statistical Society, Series C, 54, 627–644.

Ericson, W. (1969). Subjective Bayesian Models in Sampling Finite Populations. Journal of the Royal Statistical Society, Series B, 31, 195–224.

Fellegi, I. and Sunter, A. (1969). A Theory of Record Linkage. Journal of the American Statistical Association, 64, 1183–1210.

Fienberg, S.E. and Manrique-Vallier, D. (2009). Integrated Methodology for Multiple Systems Estimation and Record Linkage Using a Missing Data Formulation. Advances in Statistical Analysis, 93, 49–60.

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

Fortini, M., Liseo, B., Nuccitelli, A., and Scanu, M. (2002). Modelling Issues in Record Linkage: A Bayesian Perspective. In Proceedings of the American Statistical Association, Section on Survey Research Methods, 1008–1013.

Green, P.J. and Mardia, K.V. (2006). Bayesian Alignment Using Hierarchical Models, with Application in Protein Bioinformatics. Biometrika, 93, 235–254.

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

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.

Judson, D.H. (2007). Information Integration for Constructing Social Statistics: History, Theory and Ideas Towards a Reserach Program. Journal of the Royal Statistical Society, Series A, 170, 483–501.

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

Lahiri, P. and Larsen, M.D. (2005). Regression Analysis with Linked Data. Journal of the American Statistical Association, 100, 222–230.

Lindley, D. (1977). A Problem in Forensic Science. Biometrika, 64, 207–213.

Newcombe, H. (1988). Handbook of Record Linkage Methods for Health and Statistical Studies, Administration and Business. New York: Oxford University Press.

Robert, C. and Casella, G. (2004). Monte Carlo Statistical Methods. New York, NY: Springer.

Tancredi, A. and Liseo, B. (2011). A Novel Bayesian Approach to Matching and Population Size Problems. Annals of Applied Statistics. Forthcoming.

Winkler, W. (1993). Improved Decision Rules in the Fellegi-Sunter Model of Record Linkage. Proceedings of the American Statistical Association, Section on Survey Research Methods, 274–279.

Winkler, W.E. (2000). Machine Learning, Information Retrieval and Record Linkage. Proceedings of the American Statistical Association, Section on Survey Research Methods, 20–29.

Received November 2009
Revised February 2011