Bayesian Analysis (2019) **TBA**, Number TBA, pp. 1–26

A Unified Framework for De-Duplication and Population Size Estimation

Andrea Tancredi[*], Rebecca Steorts[†], and Brunero Liseo[‡]

Abstract. Data de-duplication is the process of detecting records in one or more datasets which refer to the same entity. In this paper we tackle the de-duplication process via a latent entity model, where the observed data are perturbed versions of a set of key variables drawn from a finite population of N different entities. The main novelty of our approach is to consider the population size N as an unknown model parameter. As a result, a salient feature of the proposed method is the capability of the model to account for the de-duplication uncertainty in the population size estimation. As by-products of our approach we illustrate the relationships between de-duplication problems and capture-recapture models and we obtain a more adequate prior distribution on the linkage structure. Moreover we propose a novel simulation algorithm for the posterior distribution of the matching configuration based on the marginalization of the key variables at population level. We apply our method to two synthetic data sets comprising German names. In addition we illustrate a real data application, where we match records from two lists which report information about people killed in the recent Syrian conflict.

Keywords: cluster analysis, entity resolution, partition models, record linkage.

1 Introduction

De-duplication (record linkage or entity resolution) is the process of merging together potentially noisy lists, data sets, or databases, often in the absence of a unique identifier, both to remove duplicated information and to increase the informative content of each single file. In fact, from a statistical perspective, performing de-duplication is paramount for obtaining a more reliable or a larger reference data set. Indeed, on one hand, the identification of duplications of the same entity would allow to increase the quality of the information associated to it. On the other hand, merging different files, once the common entities have been correctly detected, leads to a new, larger and richer data set. This new data may be suitable to perform accurate model-based statistical analyses via the additional information which could not be extracted from a single data set, because the original data may not comprise some of the model variables.

When unique identifiers are known exactly, the linkage process can be accomplished without errors. In this case, there are no specific consequences on the statistical procedures undertaken in the aforementioned situations. However, in practice, unique iden-

[*] Department of Methods and Models for Economics, Territory and Finance. Sapienza University of Rome, andrea.tancredi@uniroma1.it

[†] Department of Statistical Sciences, Duke University, beka@stat.duke.edu

[‡] Department of Methods and Models for Economics, Territory and Finance. Sapienza University of Rome, brunero.liseo@uniroma1.it

https://doi.org/10.1214/19-BA1146

tifiers are rarely available and the researcher must deal with the uncertainty related to the linking step. The problem of how to account for the matching uncertainty has then caused an active line of recent research among the statistical, the machine learning, and the computer science communities. In fact, in practical applications of record linkage procedures, the concrete possibility to make wrong matching decisions should be accounted for, especially when the result of the linking step, namely the fused data set, will be used for further statistical analyses, such as regression, capture-recapture methods or small area estimation: see for example Tancredi and Liseo (2011, 2015), Briscolini et al. (2018), and Sadinle (2018).

The classical record linkage approach with two data sets was formalized by Jaro (1989), following the seminal paper by Fellegi and Sunter (1969). This standard method is based on the comparison vectors — data vectors obtained by comparing the common fields, also known as key variables, for each pair of records. Since the distribution of the comparison vectors depends on the unknown match or non-match status of the record pairs, a mixture model fitted to the entire collection of comparison vectors can be used to classify all the pairs in two or more sets concerning their matching status (Belin and Rubin, 1995; Larsen and Rubin, 2001). Recently, Sadinle and Fienberg (2013) extended the Fellegi-Sunter approach to allow situations with three or more files, while also preserving transitive closures.

To our knowledge, Fortini et al. (2001) proposed the first Bayesian approach to record linkage, where the likelihood function provided by the set of multiple comparison vectors was used to estimate the matching configuration through the use of Markov Chain Monte Carlo (MCMC) methods. This approach, together with Larsen (2005) and Sadinle (2017), can be interpreted as a Bayesian version of the classical Fellegi-Sunter record linkage approach. Note that these papers do not assume the presence of "within file" duplications. That is, it is only possible to match a record in a file to a single record of another file and vice versa. A clear advantage of the Bayesian approach is that one can naturally account for this constraint by simply selecting appropriate prior distributions on the matching status to incorporate this assumption.

Tancredi and Liseo (2011) recently proposed a Bayesian record linkage method that is well suited for categorical data. The authors deviate from the Fellegi-Sunter approach in two major ways — they do not work with comparison data and allow for record linkage uncertainty to be accounted for in population size estimation. To handle the former, they explicitly model the fully observed records through a particular measurement error model, inspired by the so called "hit-and-miss" strategy proposed by Copas and Hilton (1990). The latter is naturally handled through the joint estimation of the record linkage model and the capture–recapture model used for population size estimation. In the same spirit, Liseo and Tancredi (2011) have introduced a record linkage model for continuous data based on a multivariate normal model with measurement error. The de-duplication problem for a single list framework has been tackled from a Bayesian perspective in Sadinle (2014) by using the information provided by the comparison data. Steorts et al. (2014, 2016) were the first to perform simultaneous record linkage and de-duplication on multiple files through the use of the fully observed records, creating a scalable record linkage algorithm. Steorts (2015) extended this work further to the case of string and categorical data, where arbitrary distance metrics between strings have been considered.

In this paper we extend both the work of Tancredi and Liseo (2011) and Steorts et al. (2016). We develop a unified framework for population size estimation by using multiple files that require both linkage and de-duplication. In fact the former paper considered only the case of two files without duplication inside each of the single lists while the latter assumed a generating population with a fixed and known size.

The rest of the paper proceeds as follows. Section 2 introduces the basic framework of our generalized Bayesian record linkage and de-duplication model and specifies the measurement error model for the key variables, namely the hit and miss model. Section 3 illustrates how the task of estimating a population size can be rephrased in terms of the partition associated with the observed records. Moreover, we provide new insights about the prior modeling of the matching configuration in a de-duplication problem and show some connections between our prior partition modeling and capture-recapture models with non homogeneous capture probabilities and duplication rates. Section 4 shows how to simplify the model by integrating out the unknown population values. Section 5 discusses the computational aspects of our proposed model. In particular, in comparison with respect to Steorts et al. (2016), we propose a novel simulation algorithm for the posterior distribution based exactly on the marginalization of the records values at the population level. Section 6 illustrates the results of our unified model for de-duplication and population size estimation applied to the synthetic data sets `RLdata500` and `RLdata10000` from the `RecordLinkage` package in `R`, presenting an intensive sensitivity analysis with respect to all model hyperparameters. In Section 7 we fit the model to a real data set reporting the names of victims of the recent Syrian conflict. Finally, Section 8 provides a brief discussion of our work.

2 The key variables model

We first introduce the methodological framework of the record linkage process. Assume L lists $F_1, F_2 \ldots, F_L$, whose records respectively relate to statistical units (e.g. individuals, firms, etc.) of partially overlapping samples. The records in the lists consist of several categorical variables which may contain corruptions, noise, and errors. Moreover we do not handle missing fields across lists, and assume that all lists have p fields in common, representing the key variables. For example, in lists regarding individuals, the common fields, might be surname, name, age, sex. Denoting the j-th record of file F_i as (i, j), the main goal of a standard record linkage procedure is to identify all pairs of records, say (i_1, j_1) and (i_2, j_2), with $i_1 \neq i_2$, that actually refer to the same unit, by using the key variables of the observed records of L lists. An additional difficulty in record linkage arises when some records in the same file, say $(i, j_1), \ldots, (i, j_n)$, refer to a single entity—known as duplicate detection.

Assume that the L sets of records have been collected from a given population with N entities, that is, $\tilde{U}_N = \{\tilde{u}_1, \tilde{u}_2, \ldots, \tilde{u}_N\}$ where $N < \infty$ and that the lists are independent, that is population entities occur independently across the lists in the same framework as Steorts et al. (2016). Assign to each member of the population the label

j' resulting from its position in the ordered list $\tilde{U}_N$. Hence $j' = 1, \ldots, N$. We assume that N is unknown; thus knowing the labels of the entities observed in the data sets would produce strong information about N if only because N should be greater of the maximum label. However these labels cannot be observed and neither estimated via the information provided by the L list of records. In fact, we anticipate that the data can be informative only on how many distinct population entities have been observed at the sample level and which sample records gather around each one of them. The former information will be used to estimate N, the latter to perform the matching process.

Let $\tilde{v}_{j'} = (\tilde{v}_{j'1}, \ldots, \tilde{v}_{j'p})$ be the vector of the p categorical key variables for the population individual j'. Denote by $\tilde{v} = (\tilde{v}_1, \ldots, \tilde{v}_N)$ the entire set of population records. Assume the set of population records $\tilde{v}$ is generated independently, for $j' = 1, \ldots, N$, from a vector of independent categorical variables $\tilde{V} = (\tilde{V}_1, \ldots, \tilde{V}_\ell, \ldots, \tilde{V}_p)$ such that $\tilde{V}_\ell \in \mathcal{V}_\ell = \{v_{\ell\,1}, \ldots, v_{\ell\,M_\ell}\}$ and that given the probability vector $\theta_\ell = (\theta_{\ell v_{\ell 1}}, \ldots, \theta_{\ell v_{\ell M_\ell}})$, $p(v_{\ell\,s}|\theta_\ell) = \theta_{\ell\,v_{\ell\,s}}, \quad s = 1, \ldots, M_\ell$, where M_ℓ is the number of categorical values for the ℓth field. Note that here and later, to simplify notations we let the arguments define the density and mass functions. Hence, the model for the population records can be written as

$$p(\tilde{v}|\theta, N) = \prod_{j'=1}^{N} \prod_{\ell=1}^{p} p(\tilde{v}_{j'\ell}|\theta_\ell) = \prod_{j'=1}^{N} \prod_{\ell=1}^{p} \theta_{\ell \tilde{v}_{j'\ell}} \tag{2.1}$$

where $\theta = (\theta_1, \ldots, \theta_\ell, \ldots, \theta_p)$.

At the sample level we assume that one does not observe the population "true" values, due to measurement and reporting errors. In fact, each set of observed records, which is a list of size n_i, $i = 1, \ldots, L$, comprises contaminated versions of subsets of the vectors $\tilde{v}_{j'}$. Let $v_{ij} = (v_{ij1}, \ldots, v_{ijp})$ denote the observed values for the j-th record of the i-th file, with $i = 1, \ldots, L$ and $j = 1, \ldots, n_i$. Moreover, denote with $v = (v_{11}, \ldots, v_{1n_1}, \ldots, v_{L1}, \ldots v_{Ln_L})$ the entire set of observed records across the L lists.

Let $\lambda_{ij} \in \{1, 2, \ldots\}$ $j = 1, \ldots, n_i, i = 1, \ldots, L$ be the unknown population labels of the sample units. This way $\lambda = (\lambda_{11}, \ldots, \lambda_{1n_1}, \ldots, \lambda_{L1}, \ldots, \lambda_{Ln_L})$ denotes the unknown matching pattern between the observed records v and the population records $\tilde{v}$, where $\lambda_{ij} = j'$ indicates that the observed record v_{ij} is a version of the population record $\tilde{v}_{j'}$. The relation $\lambda_{ij_1} = \lambda_{ij_2}$, with $j_1 \neq j_2$, implies that records j_1 and j_2 of the i-th list are co-referent to the same population record. This is an instance of duplicate-detection within the same list. Instead, when $\lambda_{i_1 j_1} = \lambda_{i_2 j_2}$, with $i_1 \neq i_2$, one has the usual record linkage framework with the same individual appearing in two different lists.

Let us now formalize the generative distortion mechanism when the population records are observed on the L lists. In particular, we assume the *hit-and-miss model* proposed by Copas and Hilton (1990) and also adopted in Steorts et al. (2014, 2016) and Steorts (2015). Let $V_{ij\ell}$ be the random variable generating $v_{ij\ell}$. Assume that $V_{ij\ell} \in \mathcal{V}_\ell$, that is $V_{ij\ell}$ has the same support of $\tilde{V}_\ell$. Moreover, set $\delta(a, b) = 1$ if $a = b$ and $\delta(a, b) = 0$ if $a \neq b$, let $\alpha_{j'} = (\alpha_{j'1}, \ldots, \alpha_{j'\ell}, \ldots, \alpha_{j'p})$ be the vector with the measurement error probabilities of the p key variables for the population individual j' and denote by $\alpha = (\alpha_1, \ldots, \alpha_{j'}, \ldots, \alpha_N)$ the entire set of distortion probabilities. We firstly assume that

$$p(v_{ij\ell} \mid \tilde{v}, \lambda, \alpha, \theta) = (1 - \alpha_{\lambda_{ij}\ell})\delta(v_{ij\ell}, \tilde{v}_{\lambda_{ij}\ell}) + \alpha_{\lambda_{ij}\ell}\theta_{\ell\, v_{ij\ell}} \quad \forall i\, j\, \ell. \tag{2.2}$$

This way, for the $\ell - th$ key variable, the true population value of the individual j' generating the record ij is observed with probability $1 - \alpha_{j'\ell}$, while, with probability $\alpha_{j'\ell}$, we observe a different value drawn from the random variable $\tilde{V}_\ell$ generating the population values.

Finally, assuming the conditional independence among all the sample records and all the key variables given their respective unobserved population counterparts, one obtains

$$p(v \mid \tilde{v}, \lambda, \alpha, \theta) = \prod_{i=1}^{L} \prod_{j=1}^{n_i} \prod_{\ell=1}^{p} p(v_{ijl} \mid \tilde{v}, \lambda, \alpha, \theta). \tag{2.3}$$

The model summarized by equations (2.1), (2.2) and (2.3) can be viewed as a part of a hierarchical model where N unobserved population records $\tilde{v}_{j'}$, drawn from a super-population model parametrized by the probability vectors θ_ℓ, generate the observed records v_{ij} with the vectors $\alpha_{j'}$ acting as record distortion parameters. The key variables probabilities θ_ℓ and the distortion probabilities $\alpha_{j'}$ are unknown quantities. For the probability vectors θ_ℓ we assume independent Dirichlet priors for $\ell = 1, \ldots, p$. An exchangeable prior will be assumed for the distortion probabilities $\alpha_{j'\ell}$ for $j' = 1, \ldots, N$. In particular the logit transformation of $\alpha_{j'\ell}$, that is $\beta_{j'l} = \log(\alpha_{j'l}/(1 - \alpha_{j'l}))$ will be Normal with mean β_{0l} and variance s^2, for $j' = 1, \ldots, N$ and β_{0l} will be Normal with mean m_0 and variance s_0^2. Note also that distortion probabilities for different key variables will be assumed independent.

3 The prior for the records partition and the population size

The interpretation and the prior specification of the labeling variables λ is more challenging with respect to all other model variables and parameters. One interpretation of λ is that its values are drawn from a known and specific sampling design, which generates the labels allowing for duplications within each list. Consider the simplest situation, where L independent simple random samples are drawn with replacement from a population of size $N < \infty$. It follows that

$$p(\lambda|N) = \prod_{i=1}^{L} \prod_{j=1}^{n_i} p(\lambda_{ij}|N) = \left(\frac{1}{N}\right)^n \tag{3.1}$$

where $n = \sum_{i=1}^{L} n_i$. Therefore, with fixed N, one has a uniform prior over the set of all possible configurations of the λ values. This is exactly the prior used in Steorts et al. (2016) and we will call this distribution the uniform prior on the label space. Note that a similar scheme was considered also by Tancredi and Liseo (2011) in the context of two-file record linkage without duplication. There, their matching matrix prior distribution was based on the assumption that the lists were two simple random samples without replacement.

We now investigate an alternative aspect of the uniform prior distribution of λ given N. Let $Z = Z(\lambda)$ denote the partition of the n records determined by λ. For example assuming $N = 3$, $L = 1$, $n_1 = n = 3$ and $\lambda = (1, 2, 2)$ we have the partition $Z = 1|23$ indicating that the second and third sample units share the same population label which is different from the one of the first sample unit. Note that, in this case, λ may assume 27 different vectors, all with equal probability, producing the five different partitions of the $n = 3$ records, namely $\{123,\ 1|23,\ 13|2,\ 12|3,\ 1|2|3\}$. Moreover the partition 1|23 can be obtained when λ is one of the following vectors $(1, 2, 2), (1, 3, 3), (2, 1, 1), (2, 3, 3), (3, 1, 1), (3, 2, 2)$. Thus the probability of the partition 1|23 given $N = 3$ is 6/27. When $N = 4$, λ may assume 64 different vectors and it is simple to verify the probability of the partition 1|23 is now 12/64. Thus the distribution on the sample labels λ given N induces a distribution on the partition space which depends on N. This means that the simple knowledge of the partition of the sample records is able to produce information on the population size N. Furthermore, matches and duplicates are completely specified given the knowledge of Z. Thus estimating the partition will permit at the same time to produce inference on N and to estimate the linkage structure of the data at hand.

In the following we will indicate with $\mathcal{P}$ the set containing all the possible partitions of the n observed records and with $z \in Z$ a single block of the partition Z. Moreover, let $u_z(\lambda)$ be the label identifying the block z on the vector λ and $U = U(\lambda) = \{u_z(\lambda), z \in Z\}$ be the set of the block labels ordered accordingly to the sequence $z \in Z$. Hence $\lambda = (3, 5, 1, 5)$ and $\lambda = (5, 3, 1, 3)$ produce the same partition $Z = 1|24|3$ but different label vectors $U = (3, 5, 1)$ and $U = (5, 3, 1)$. Note that (Z, U) and λ are in one to one correspondence, thus $p(Z, U|N) = p(\lambda|N)$.

We now obtain the prior distribution on the partition space $\mathcal{P}$, for a given N, resulting from the uniform prior on the label space. Let $k = k(Z)$ be the observed number of blocks of the partition Z. The number of elements λ producing the partition Z is $N_k = N!/(N - k)!$. In fact we have $\binom{N}{k}$ ways to select the unordered labels for the blocks of Z and for each of them $k!$ ordered labellings U. Thus

$$p(Z|N) = \sum_{\lambda : Z(\lambda) = Z} p(\lambda|N) = \sum_{U|Z} \left(\frac{1}{N}\right)^n = \left(\frac{1}{N}\right)^n N_k, \quad \forall Z \in \mathcal{P}. \tag{3.2}$$

Moreover

$$p(U|Z, N) = \frac{1}{N_k}.$$

Note also that $N^n = \sum_{k=0}^{n} N_k S(n, k)$, where $S(n, k)$ is the Stirling number of the second kind, that is the number of possible partitions of the n records into k non empty sets, so we have

$$p(Z|N) = \frac{N_k}{\sum_{r=0}^{n} N_r S(n, r)}, \quad \forall Z \in \mathcal{P}. \tag{3.3}$$

Following Pitman (2006), equation (3.3) defines a special case of Gibbs partitions. Moreover, the distribution of the random number of blocks K is given by

$$p(k|N) = \frac{N_k \, S(n,k)}{N^n} \qquad k = 1, \dots, n.$$

The mean and the variance of K are easily obtained as

$$E(K|N) = N\,(1 - (1 - 1/N)^n) \tag{3.4}$$

and

$$\mathrm{Var}(K|N) = N\big[(N-1)(1-2/N)^n - N(1-1/N)^{2n} + (1-1/N)^n\big]$$

(see Appendix A in Supplementary Material, Tancredi, Steorts, and Liseo, 2019). For fixed N, as the number of records $n \to \infty$, the distribution of $K|N$ concentrates on N, since $E(K|N)$ tends to N and the variance vanishes. Also observe that, for a fixed number of records n and large values of N, the distribution of the distinct entities $K|N$ concentrates on n. That is, the prior probability of observing links or duplicates approaches 0 in the limit, as intuition suggests.

To complete the prior modeling of the linkage structure we need to specify the prior for the population size N. Throughout this paper we assume

$$p(N) = \frac{1}{\zeta(g)\,N^g} \qquad N = 1, 2, \dots \tag{3.5}$$

where $\zeta(g) = \sum_{N=1}^{\infty} 1/N^g$ is the Riemann zeta function. Such a prior is proper $\forall\, g > 1$. Note that the use of heavy-tailed priors $p(N) \propto 1/N^g$ as non informative distributions is quite diffuse in population size Bayesian estimation, see for example George and Robert (1992) or Wang et al. (2007). Straightforward calculations (see Appendix A in Supplementary Material, Tancredi, Steorts, and Liseo, 2019) show that, under this class of priors, the marginal prior mean for K is

$$E(K) = \sum_{s=1}^{n} \binom{n}{s} (-1)^{s+1} \frac{\zeta(s+g-1)}{\zeta(g)}. \tag{3.6}$$

Notice that, as $g \to 1$ $E(K)$ converges to n which is the upper end point of the support of K; hence when g approaches to 1 the whole distribution of K concentrates on n.

The left part of Table 1 reports, for different values of g, the mean and the standard deviation for K when the total number of records is $n = 500$ as in the first application that will be illustrated in this paper. Such summaries are obtained by simulating 10^7 draws from $p(N, \lambda)$ via the accept reject algorithm for $p(N)$ proposed in Devroye (1986) §10.6 and by direct simulation of $p(\lambda|N)$. Note that even for values of g close to 1, the standard deviation of K is quite high. Thus, such values of g have the important role to induce a priori a high number of clusters with few observation per cluster, i.e. the microclustering effect, see for example Zanella et al. (2016) and Johndrow et al. (2018), without being too much informative. The right part of Table 1 reports the mean and the standard deviation for K when we use the uniform prior for λ by fixing the values of N. Note that the assumption of a uniform distribution on the label space conditioned

g	$E(K)$	$SD(K)$	N	$E(K)$	$SD(K)$
2	4	20	250	216	4.5
1.5	30	87	500	316	7.0
1.1	271	224	1000	394	7.4
1.05	375	198	2500	453	6.0
1.02	452	137	5000	476	4.6
1.01	477	98	10000	488	3.4
1.001	498	31	100000	499	1.1

Table 1: Mean and standard deviation of the random variable K with different values of g and of $K|N$ with different values of N.

on the value of N might not be adequate in real applications of record linkage and de-duplication even when we are only interested on the linkage structure and we do not need to make inference on N. In fact the resulting distribution on the number of distinct entities K will be generally too concentrated as illustrated by the extremely low standard deviation.

3.1 Estimation for the population size N when the partition Z is known

When the partition Z of the n records is known and the model generating the partition is given by (3.2), inference on N can be conducted via the posterior distribution

$$p(N|Z) \propto p(Z|N)p(N) \propto \frac{N_k}{N^n} p(N) I_{\{N \geq k\}} \tag{3.7}$$

where $I_{\{N \geq k\}}$ denotes the indicator function of the set $N \geq k$. Notice that the distribution (3.7) is exactly the posterior for N obtained from a T-stage homogeneous capture recapture model when $T = n$, we observe k different individuals across the samples and we condition on one capture in each occasion, see for example Marin and Robert (2014) §5. Note also that assuming the prior (3.5), the posterior (3.7) is proper $\forall g \geq 0$ when $k < n - 1$.

It is also interesting to observe that the mode of the posterior for N when $g = 0$ is approximated by the moment estimator of N obtained by the expression (3.4). In fact, by approximating the logarithm of $p(Z|N)$ using the Stirling formula, we have that

$$\log p(Z|N) = N \log N - (N-k) \log (N-k) - n \log N + O(\log N) - O(\log (N-k))$$

and the mode of the posterior distribution $p(N|Z)$ when $g = 0$, i.e. $p(N) \propto c$, is approximately given by the solution of the equation $k = N\,(1 - e^{-n/N})$ which can be further approximated by solving

$$k = N(1 - (1 - 1/N)^n) \tag{3.8}$$

that is the equation providing the expected value of K as a function of N.

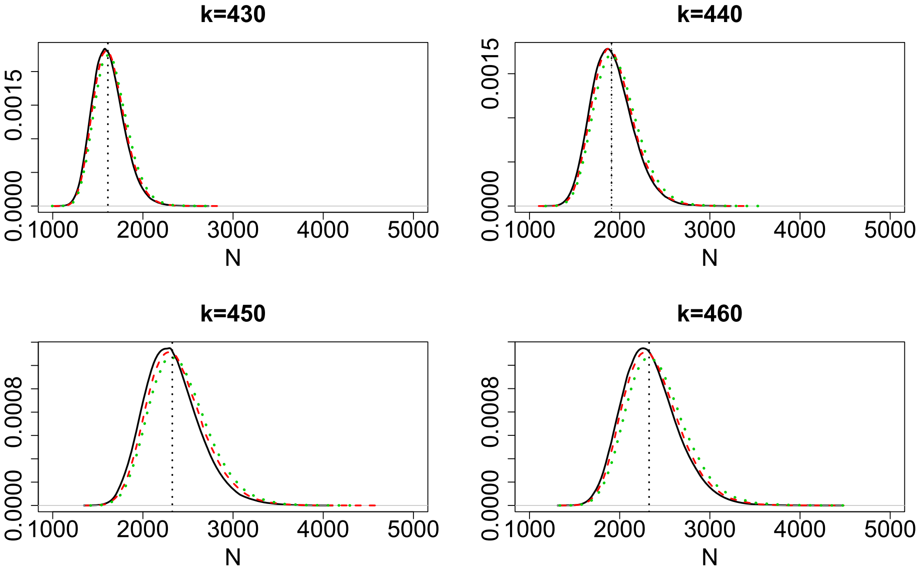

Figure 1: Posterior distributions for N when $n = 500$, k is known and equal to 430,440,450,460 and $p(N) \propto 1/N^g$ with $g = 0$ (dotted line), $g = 1$ dashed line and $g = 2$ solid line.

Figure 1 shows the posterior distributions for N when $k = 430, 440, 450, 460$ and $p(N) \propto 1/N^g$ and $g = 0, 1, 2$. The posterior distributions have been obtained with a Metropolis-Hastings algorithm with Poisson proposals. Note that different values of k produce quite different posterior distributions, while sensitivity with respect to the proposed values is g is limited. The vertical line is the solution of the equation (3.8), which properly approximates the maximum a posteriori estimate when $g = 0$.

3.2 Connections with capture-recapture models with non homogeneous capture probabilities and duplication rates

Now suppose that, in order to form the jth list, each one of the N population units is subject to being captured a random number of times. That is for each label j' and for each list j there are $T_{jj'}$ attempts to capture the population unit $U_{j'}$ and for each attempt, the capture probability is p_j. Moreover assume that the random variables $T_{jj'}$ are independent Poisson with mean δ_j. Hence $\delta_1, \ldots, \delta_L$ are list dependent parameters providing the "within- list" duplication rates while $p_1, \ldots, p_L$ are the different list capture probabilities.

Now let $X_{jj't}$ for $t = 1, \ldots, T_{jj'}$, $j' = 1, \ldots, N$ and $j = 1, \ldots, L$ be a random number of independent Bernoulli variables with probability p_j indicating if, in list j, unit $U_{j'}$ has been captured at the attempt t and let $X_{jj'} = \sum_{t=1}^{T_{jj'}} X_{jj't}$ be the number of times that $U_{j'}$ has been captured in list j. Note that the mean of $X_{jj'}$ is $\delta_j p_j$ and it is Poisson distributed being the sum of a Poisson number of Bernoulli variables. Now let n_j be the list size, for $j = 1, \dots, L$, and observe that $n_j = \sum_{j'}^{N} X_{jj'}$ is Poisson distributed with mean $N\delta_j p_j$, the conditional distribution of $X_{jj'}|n_j$ is Binomial$(n_j, 1/N)$ and

$$p(x_{j1}, \dots x_{jN}|n_j) = \frac{n_j!}{\prod_{j'=1}^{N} x_{jj'}!} \prod_{j'=1}^{N} (1/N)^{x_{jj'}} = \frac{n_j!}{\prod_{j'=1}^{N} x_{jj'}!} \frac{1}{N^{n_j}} \quad j = 1, \dots, L.$$

Moreover, each label sequence of the j list, that is the vector $\lambda_j = (\lambda_{1j}, \dots, \lambda_{n_j j})$ has probability

$$p(\lambda_j|n_j) = p(\lambda_j|x_{j1}, \dots x_{jN}, n_j)p(x_{j1}, \dots x_{jN}|n_j) = \frac{1}{N^{n_j}}.$$

Assuming duplication and capture independence across the lists, we also have that $p(\lambda|n_1, \dots n_L) = \frac{1}{N^n}$. Then, the conditioning on list sizes has eliminated the duplication rates and the capture probabilities, thus providing a conditional likelihood for N which depends on the non identifiable population labels which, in turn, provides the likelihood function for N (3.2) given the observable partition Z. In summary, the proposed prior (3.1) for λ exactly embeds the sampling information, conditional on list sizes, provided by a capture-recapture model with non homogeneous capture probabilities and duplication rates.

Notice that the elimination of the capture probabilities and the duplication rate parameters from the prior model for λ automatically implies that two records of the same list and two records of two different sets would have the same prior probability to be duplicates. Such assumption, which follows directly from the prior (3.1), is admittedly unlikely to be true in practice. We simply consider this assumption a convenient and operative starting point for performing matching estimation.

4 The hit-miss marginal model for record clustering

A convenient property of the hit-miss model illustrated in Section 2 is that one can integrate out the unknown population values $\tilde{v}$ to directly obtain the distribution $p(v|Z, U, N, \alpha, \theta)$, as it is illustrated below. The resulting marginal distribution is the product of within-blocks distributions. In fact, records belonging to different blocks are independent because they refer to different and independent population records, while records within the same block are dependent, since they are observations on the same population individual. Clustering approaches based on similar dependence structures are discussed in Booth et al. (2008) and McCullagh and Yang (2008).

Let $z \in Z$ be a partition block, let $v_z = (v_{ij} : ij \in z)$ denote the corresponding cluster of records and let $v_{z\,\ell} = (v_{ij\ell} : ij \in z)$ denote the cluster of observed records for the ℓ-th key variable. Also let u_z denote the label in U corresponding to the block z and let $\tilde{v}_U = (\tilde{v}_{u_z}, z \in Z)$ and $\alpha_U = (\alpha_{u_z}, z \in Z)$ be the relative sets of population records and distortion probabilities.

Firstly, note that equation (2.3) can be re-expressed, taking into account the partition imposed by λ, in the following way

$$\begin{aligned} p(v|\tilde{v}, \lambda, N, \alpha, \theta) &= p(v|\tilde{v}, Z, U, N, \alpha, \theta) = \prod_{j'=1}^{N} \prod_{ij:\lambda_{ij}=j'} p(v_{ij}|\tilde{v}, Z, U, N, \alpha, \theta) \\ &= \prod_{z \in Z} p(v_z|\tilde{v}_{u_z}, Z, U, N, \alpha, \theta). \end{aligned}$$

Hence, observing that $p(\tilde{v}_U|Z, U, N, \alpha, \theta) = \prod_{z \in Z} p(\tilde{v}_{u_z}|Z, U, N, \alpha, \theta)$, and marginalizing out the true values $\tilde{v}_U$, one obtains

$$p(v|Z, U, N, \alpha, \theta) = \prod_{z \in Z} p(v_z|Z, U, N, \alpha, \theta).$$

Now, let us consider a block with only a single record, i.e., $z = \{(i\,j)\}$. Then the marginal distribution of the observed value for the l-th field of this record is

$$\begin{aligned} p(v_{z\ell}|Z, U, N, \alpha, \theta) &= \sum_{\tilde{v}_{u_z\ell} \in \mathcal{V}_\ell} p(v_{z\ell}, \tilde{v}_{u_z\ell}|Z, U, N, \alpha, \theta) \\ &= \sum_{\tilde{v}_{u_z\ell} \in \mathcal{V}_\ell} p(v_{z\ell}|\tilde{v}_{u_z\ell}, Z, U, N, \alpha, \theta) p(\tilde{v}_{u_z\ell}|Z, U, N, \alpha, \theta) \\ &= \sum_{\tilde{v}_{u_z l} \in \mathcal{V}_\ell} [(1 - \alpha_{u_z\ell})\delta(v_{z\ell}, \tilde{v}_{u_z\ell}) + \alpha_{u_z\ell}\theta_{\ell\, v_{z\ell}}]\theta_{\ell\tilde{v}_{u_z\ell}} = \theta_{\ell v_{z\ell}}. \end{aligned}$$

Since we have assumed conditional independence among the key variables, one has

$$p(v_z|Z, U, N, \alpha, \theta) = \prod_{\ell=1}^{p} p(v_{z\ell}|Z, U, N, \alpha, \theta) = \prod_{\ell=1}^{p} \theta_{\ell\, v_{z\ell}}.$$

After simple algebra, an analytical expression can also be found for a cluster $z = \{(i_1\, j_1), (i_2\, j_2)\}$ with two records, that is,

$$\begin{aligned} p(v_z|Z, U, N, \alpha, \theta) \;=\; & \prod_{\ell=1}^{p} \big[\delta(v_{i_1 j_1 \ell}, v_{i_2 j_2 \ell})\theta_{\ell\, v_{i_1 j_1 \ell}}(1 - \alpha_{u_z\ell})^2 + \\ & (2\alpha_{u_z\ell} - \alpha_{u_z\ell}^2)\theta_{\ell\, v_{i_1 j_1 l}}\theta_{\ell\, v_{i_2 j_2 \ell}}\big]. \end{aligned}$$

Furthermore, it is straightforward (see Appendix B in Supplementary Material, Tancredi, Steorts, and Liseo, 2019) to obtain a general and recursive formula for the marginal distribution of a cluster with n records, $z = \{(i_1\, j_1), \ldots, (i_n\, j_n)\}$:

$$\begin{aligned} p(v_{z\,\ell}|Z, U, N, \alpha, \theta) \;=\; & \alpha_{u_z\ell}\theta_{\ell v_{i_n j_n \ell}} p(v_{z\setminus(i_n\, j_n)\,\ell}|Z, U, N, \alpha, \theta) + \\ & (1 - \alpha_{\ell u_z})\theta_{\ell v_{i_n j_n \ell}} \prod_{h=1}^{n-1} \Big[(1 - \alpha_{u_z\ell})\delta(v_{i_h j_h \ell}, v_{i_n j_n \ell}) + \alpha_{u_z\ell}\theta_{\ell v_{i_h j_h \ell}}\Big], \end{aligned}$$

where $v_{z\setminus(i_n\, j_n)\,\ell}$ indicates the cluster values for the ℓ-th key variable excluding those observed on the record (i_n, j_n).

As a final note, observe that, for all z, $p(v_z|Z,U,N,\alpha,\theta)$ depends on α and on the partition block z along with the corresponding label u_z. Then $p(v|\lambda,N,\alpha,\theta) = p(v|Z,U,\alpha,\theta)$, that is the distribution of the observed data depends on Z,U,α,θ and not on the population size N.

5 Posterior simulation

De-duplication and population size inference can be carried out by simulating from the posterior $p(Z,N|v)$, that is the marginal distribution of $p(\lambda,N,\beta,\beta_0,\theta|v)$ where β is the vector with the logit transformations of the distortion probabilities of the N population entities, β_0 is the vector with their means for each key variable and

$$\begin{aligned} p(\lambda,N,\beta,\beta_0,\theta|v) &\propto p(Z,U,N,\beta,\beta_0,\theta|v) \qquad (5.1)\\ &\propto p(v|Z,U,\beta,\theta)p(U|Z,N)p(Z|N)p(\beta|\beta_0)p(N)p(\beta_0)p(\theta)\\ &\propto \prod_{z\in Z} p(v_z|Z,U,\beta_{u_z},\theta)p(U|Z,N)p(Z|N)p(\beta|\beta_0)p(N)p(\beta_0)p(\theta). \end{aligned}$$

Note that the marginal posterior $p(Z,N,\beta_0,\theta|v)$ is

$$\begin{aligned} &p(Z,N,\beta_0,\theta|v)\\ &\propto \sum_U \left[\int_{R^{N\times p}} \prod_{z\in Z} p(v_z|Z,U,\beta_{u_z},\theta)p(\beta|\beta_0)d\beta\right] p(U|Z,N)p(Z|N)p(N)p(\beta_0)p(\theta)\\ &\propto \sum_U \left[\int_{R^{k\times p}} \prod_{z\in Z} p(v_z|Z,U,\beta_{u_z},\theta)p(\beta_U|\beta_0)d\beta_U\right] p(U|Z,N)p(Z|N)p(N)p(\beta_0)p(\theta)\\ &\propto \sum_U \left[\prod_{z\in Z}\int_{R^p} p(v_z|Z,U,\beta_{u_z},\theta)p(\beta_{u_z}|\beta_0)d\beta_{u_z}\right] p(U|Z,N)p(Z|N)p(N)p(\beta_0)p(\theta). \end{aligned}$$

Note that by integrating out the measurement error parameters β_{u_z}, the integrals inside the square brackets in the last expression do not depend on the population labels $\{u_z, z\in Z\}$. Hence we have that

$$p(Z,N,\beta_0,\theta|v) \propto \prod_{z\in Z} q(v_z|\beta_0,\theta)p(Z|N)p(N)p(\beta_0)p(\theta) \qquad (5.2)$$

where $q(v_z|\beta_0,\theta)$ is the marginal distribution of the block z given β_0 and θ.

Now let η be an alternative set of labels for the sample records where $\eta_{ij}\in\{1,\ldots,n\}$ $\forall ij$. Let Z be the partition generated by η and U' the set of labels assigned by η to the blocks $z\in Z$. Note that $\eta \leftrightarrow (Z,U')$. Assume that $p(Z|N) = N_k/N^n$, as for the random partition generated by λ, while $p(U'|Z,N) = 1/\left(\binom{n}{k}k!\right)$ so that

$$p(\eta|N) = p(Z|N)p(U'|Z,N) = \frac{N_k}{N^n}\frac{1}{n_k}.$$

Moreover let $\beta'_{j'}$ for $j' = 1, \dots, n$ be a vector with L measurement error parameters with the same prior model of the original variable dimension vector β. Then the posterior (5.2) can also be seen as the marginal, with respect to U' and β' of the distribution

$$\begin{aligned} p(\eta, N, \beta', \beta_0, \theta | v) \quad &\propto \quad p(Z, U', N, \beta', \beta_0, \theta | v) \qquad (5.3) \\ &\propto \quad \prod_{z \in Z} p(v_z | Z, U', \beta'_{u'_z}, \theta) p(U'|Z, N) p(Z|N) p(\beta'|\beta_0) p(N) p(\beta_0) p(\theta). \end{aligned}$$

Note that simulating the distribution (5.3) instead of (5.1) may imply a considerable saving of computing time since the label indicators η_{ij} vary in $\{1, \dots, n\}$ and no longer in $\{1, \dots, N\}$ without any loss of information for the de-duplication and population size inference. Drawings from the distribution (5.3) can be obtained updating the elements η, β', β_0, N and θ via the following Gibbs sampler algorithm.

In particular, the updating of the vector η which leads to the consequent updating of both Z and U' is the most critical step of the algorithm. Denote $\eta_{(-ij)}$ the vector η without the element η_{ij}. Moreover let $z \setminus (ij)$ be a partition block without the record ij, and let z_q be the partition block such that $u'_{z_q} = q$. Then, the full conditional distribution of η_{ij} can be written as

$$\begin{aligned} p(\eta_{ij} = q | \eta_{(-ij)}, N, \beta', \beta_0, \theta, v) &\propto \prod_{z \in Z} p(v_z | Z, U', \beta'_{u'_z}, \theta)\, p(\eta_{ij} = q | \eta_{(-ij)}) \qquad (5.4) \\ &\propto \prod_{z \in Z} \frac{p(v_z | Z, U', \beta'_{u'_z}, \theta)}{p(v_{z \setminus (ij)} | Z, U', \beta'_{u'_z}, \theta)} p(\eta_{ij} = q | \eta_{(-ij)}) \\ &\propto \frac{p(v_{z_q} | \eta, \beta'_q, \theta)}{p(v_{z_q \setminus (ij)} | \eta, \beta'_q, \theta)} p(\eta_{ij} = q | \eta_{(-ij)}) \quad q = 1, \dots n. \end{aligned}$$

This occurs because, in equation (5.4), setting $\eta_{ij} = q$, one has $z = z \setminus (ij)$, $\forall z \neq z_q$ so that

$$\frac{p(v_z | \eta, \beta'_{u'_z}, \theta)}{p(v_{z \setminus (ij)} | \eta, \beta'_{u'_z}, \theta)} = 1 \quad \forall z \neq z_q.$$

Equation (5.4) suggests that the conditional posterior probability $p(\eta_{ij} | \eta_{(-ij)}, N, \beta', \beta_0, \theta, v)$ depends on the ratio between the probability of the cluster of records referring to the label q considering $\eta_{-(ij)}$ and $\eta_{ij} = q$ and the probability of the same cluster with the exclusion of the record ij.

The above ratio, when the label q identifies an already existing block given $\eta_{-(ij)}$, exploiting the recursive formula (4.1), can also be written as

$$\begin{aligned} &\frac{p(v_{z_q} | \eta, \beta'_q, \theta)}{p(v_{z_q \setminus (ij)} | \eta, \beta'_q, \theta)} \\ &\quad = \prod_{\ell=1}^{p} \left[\beta'_{q\ell} \theta_{\ell\, v_{ij\ell}} + (1 - \beta'_{q\ell}) \frac{\prod_{(i_h, j_h) \in z_q \setminus (ij)} \left((1 - \beta'_{q\ell}) \delta(v_{i_h j_h \ell}, v_{ij\ell}) + \beta'_{q\ell} \theta_{\ell v_{i_h j_h \ell}} \right)}{p(v_{z_q \setminus (ij)\, \ell} | \eta, \beta'_{q\ell}, \theta)} \right]; \end{aligned}$$

however, it gets simplified into

$$\frac{p(v_{z_q}|\eta, \beta'_q, \theta)}{p(v_{z_q \setminus (ij)}|\eta, \beta'_q, \theta)} = \prod_{\ell=1}^{p} \theta_{\ell, v_{ij\ell}}$$

when the label q identifies a new block.

Thus we can update η_{ij} with the following distribution

$$\begin{aligned} & p(\eta_{ij} = q|\eta_{(-ij)}, N, \beta', \beta_0, \theta, v) \\ & = \begin{cases} \frac{p(v_{z_q}|\eta,\beta'_q,\theta)}{p(v_{z_q \setminus (ij)}|\eta,\beta'_q,\theta)} p(\eta_{ij} = q|\eta_{(-ij)}) & \text{if } q \text{ labels an observed cluster} \\ \prod_{\ell=1}^{p} \theta_{\ell, v_{ij\ell}} p(\eta_{ij} = new|\eta_{(-ij)})/(n - k_{(-ij)}) & \text{if } q \text{ labels a new cluster} \end{cases} \end{aligned} \tag{5.5}$$

for $i = 1, \ldots, L, j = 1, \ldots, n_i$, where $k_{(-ij)}$ is the number of clusters without the label η_{ij} and

$$p(\eta_{ij} = q|\eta_{-(ij)}) \propto 1 \quad \text{and} \quad p(\eta_{ij} = \text{new}|\eta_{-(ij)}) \propto (N - k_{-(ij)}).$$

Such a way to update the cluster composition is a standard approach for mixture models when the marginal likelihood of the cluster observations is known or it can be easily calculated, as in our case via the recursive formula (4.1); see for example MacEachern (1994) and Neal (2000).

The full conditional distribution

$$p(\beta'_{j'\ell}|\beta'_{-(j'\ell)}, \beta_0, \eta, N, \theta, v) \propto p(v_{z_{j'}\ell}|\beta'_{j'\ell}, \eta, \theta) p(\beta'_{j'\ell}|\beta_0)$$

can be updated using a Metropolis step when j' labels a record cluster or directly by the prior distribution $p(\beta'_{j'l}|\beta_0)$ when j' does not identify any cluster. A Metropolis step can also be used to update the parameters β_{0l} whose conditional distribution is

$$p(\beta_{0\ell}|\beta_{-(0\ell)}, \beta', \eta, N, \theta, v) \propto \prod_{z \in Z} p(\beta'_{u_z \ell}|\beta_{0\ell}, \eta) p(\beta'_{0\ell}).$$

Anyway, to improve the mixing of the chain we have adopted a non centered parameterization (Papaspiliopoulos et al., 2003), for $\beta'_{j'\ell}$, updating the differential effects $\beta'_{j'l} - \beta_{0\ell}$ slightly modifying the Metropolis steps for $\beta_{j'l}$ and β_{0l}.

The full conditional distribution of N is given by

$$p(N|\eta, \beta', \beta_0, \theta, v) \propto p(Z|N) p(N) \propto \frac{N_k}{N^{n+g}} I_{\{N \geq k\}}$$

and an exact Gibbs step truncating N to a very large integer or a Metropolis step with integer proposals can be easily implemented. Lastly, note that the full conditional distribution of the probability vector θ_ℓ is

$$p(\theta_\ell|\theta_{-(\ell)}, \eta, \beta', \beta_0, N, v) \propto \prod_{z \in Z} p(v_{z\ell}|\beta'_{u'_z, \ell}, \theta_\ell, \eta) p(\theta_\ell)$$

which can be updated using a Metropolis-Hastings steps with a Dirichlet proposal distribution. Finally note that having all n records from a single set or from $L > 1$ sets would not make a difference for the whole proposed algorithm. This is a direct consequence of the use of the uniform prior distribution $p(\lambda|N)$ which, although based on overly restrictive assumptions, has the advantages of simplifying the computation of the posterior distribution. In fact, more elaborated prior distributions for λ would require more complex posterior sampling schemes.

6 Experiments with synthetic data

To investigate the performance of our proposed methodology we first consider the `RLdata500` data set from the `RecordLinkage` package in `R`. This synthetic data set consists of 500 records, each comprising first and last name and full date of birth. This data set contains 50 records that are intentionally constructed as "duplicates" of other records. Hence the true value of k is 450 and the true partition is represented by 400 clusters of size one and 50 clusters of size two. In order to apply a model with categorical variables only, we partially modify the data set by transforming names and surnames via the English soundex algorithm. This way we obtain records with 14 fields; 4 of them are produced by the name, 4 comes from the surname and the last 6 are obtained from the date of birth (4 given by the year, 1 by the month, and 1 by the day). Table 2 shows the first 6 records of the transformed data set.

	name fields				*surname fields*				*day of birth fields*					
									year				*month*	*day*
1	C	6	2	3	M	6	0	0	1	9	4	9	7	22
2	G	6	3	0	B	6	0	0	1	9	6	8	7	27
3	R	1	6	3	H	6	3	5	1	9	3	0	4	30
4	S	3	1	5	W	4	1	0	1	9	5	7	9	2
5	R	4	1	0	K	6	2	6	1	9	6	6	1	13
6	J	6	2	5	F	6	5	2	1	9	2	9	7	4

Table 2: First 6 records of the `RLdata500` data set with names and surnames transformed via the soundex algorithm.

We fit our de-duplication and size estimation model to the modified `RLdata500` data set by taking $p(N) \propto 1/N^g$ with g=1.02. Note that with this choice, as reported in Table 1, the prior mean for K is approximately 450, that is the true number of clusters for this file, and the dispersion is quite large as we can see also from the upper left panel of Figure 2 where the prior for K has been plotted. The probability vector θ_ℓ are uniform on the simplex. The prior variance of the logit transformations $\beta_{j'\ell}$ of the distortion probabilities is equal to $s^2 = 0.5$ while the mean and the variance of their common mean $\beta_{0\ell}$ are $m_0 = logit(0.01)$ and $s^2 = 0.1$. Such a prior specification leads to a prior mean and a 0.99 prior quantile for $\alpha_{j'l}$ respectively equal to 0.013 and 0.058 indicating strong belief towards low block distortion probabilities. We observe that this is a condition to facilitate the micro-clustering effect since larger distortion

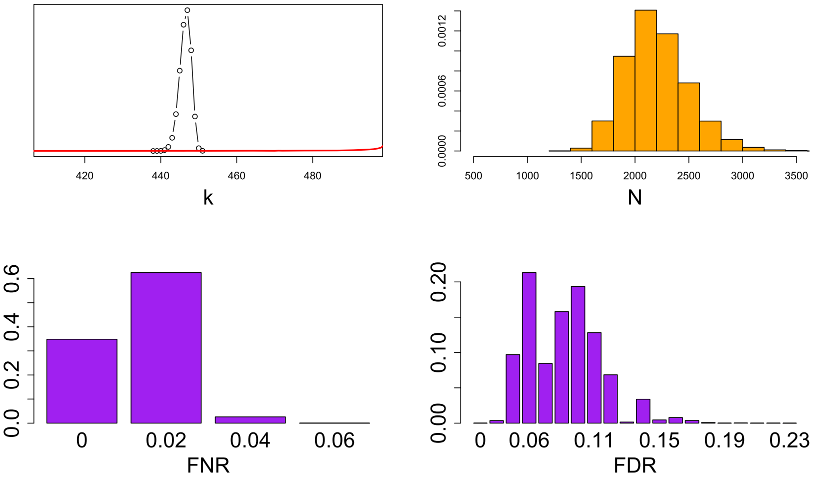

Figure 2: `RLdata500` data set. Prior and posterior distributions of K and posterior distribution of N, FNR and FDR when $s^2 = 0.5$, $m_0 = logit(0.01)$ and $s^2 = 0.1$ and $g = 1.02$.

probabilities would allow to gather more records into the same cluster even if they do not refer to the same entity. Instead, with low values of $\alpha_{j'l}$ we force all the clusters to have a reduced within-cluster variability and a greater between-cluster separation. At this regard, Johndrow et al. (2018) show from a more general and theoretical point of view that, in order to be effective, entity resolution via micro-clusters identification requires that the measurements errors go to zero as the number of entities increases. Such a condition practically states the infeasibility of cluster based approaches for high dimensional record linkage problems without introducing further information that may facilitate the correct aggregation into microclusters as our informative prior on α_{jl} tries to do.

The Metropolis within Gibbs algorithm described in Section 5, was run for 50000 iterations. Figure 2 reports the posterior distributions for K and N and the performance of the record linkage procedure measured in terms of the posterior distributions of the false negative rates (FNR) and the false discovery rates (FDR) (third and fourth rows). For a review of false negative and false discovery rates in the context of record linkage we refer to Steorts (2015). In single list framework, these rates are obtained by setting

$$\Delta_{j_1 j_2} = \begin{cases} 1 & \eta_{1j_1} = \eta_{1j_2} \\ 0 & \eta_{1j_1} \neq \eta_{1j_2} \end{cases}$$

and calculating

$$FNR = \frac{\sum_{j_1<j_2}(1-\Delta_{j_1j_2})\Delta^{true}_{j_1j_2}}{\sum_{j_1<j_2}\Delta^{true}_{j_1j_2}} \quad FDR = \frac{\sum_{j_1<j_2}\Delta_{j_1j_2}(1-\Delta^{true}_{j_1j_2})}{\sum_{j_1<j_2}\Delta_{j_1j_2}}$$

across the MCMC simulation.

Note that the posterior means for K and N are equal to 446.6 and 2209 while the 95% posterior intervals are respectively [443,449] and [1710,2854]. Hence we have a considerable uncertainty reduction with respect to the prior specification for these quantities. The low posterior mean for the FNR, equal to 0.015, indicates that almost all the true matches are correctly linked in the same cluster. In addition, the posterior mean for the FDR, equal to 0.080, suggests that the model produces a limited number of false links. Hence the performance of the de-duplication process is quite satisfactory considering also the information lost in the data set transformation via the soundex algorithm and the diffuse prior specification of N and K.

Table 3 shows the result of a sensitivity analysis with respect to the hyperparameters controlling the prior for the β_{jl}s, i.e. s^2, m_0 and s_0^2, and with respect to hyperparameter g regulating the fatness of the prior for N. In particular we show the posterior means for K, N, the FNR and the FDR obtained when $logit^{-1}(m_0) = 0.01, 0.1, 0.2$, $s^2 = 0.1, 0.5, 1$ $s_0^2 = 0.1, 0.5, 1$ and $g = 1.01, 1.02, 1.05, 1.1, 1.5, 2$. For each value of g, the results are ordered with respect to increasing values of $logit^{-1}(m_0)$, then by the variance of $\beta_{j'l}$ $s^2 + s_0^2$ and finally by the covariance between $\beta_{j'l}$ and $\beta_{j''l}$. As expected increasing a priori the mean and the variance of the distortion probabilities leads to increase the cluster sizes as we can see from the reduced values of $E(K|y)$. In fact the posterior mean of K switches dramatically from the corrected values by about 450 microclusters to inconsistent values of less than 200 clusters, confirming the theoretical findings of Johndrow et al. (2018) regarding the necessity to introduce external information to obtain micro-clusters via a mixture model based approach. The small FNR and the high FDR when the microcluster effect do not occur confirm that records of the same entity are gathered into the same cluster although together with the other records generated by entities without list duplications. Note also that with the same variance values, micro-clustering is more likely to occur with lower covariance between $\beta_{j'l}$ and $\beta_{j''l}$. Finally notice that the effect of g is practically negligible with higher values that slightly reduce the posterior mean of N.

Table 4 shows the posteriors means for K, FNR and FDR, obtained by conditioning on grid of known values for N varying from 250 to 10000 and the hyperparameters values s^2, s_0^2 and $logit^{-1}(m0)$ equal to 0.5, 0.1 and 0.01. Note that also by fixing the values of N we regulate the microclustering effect with larger values producing the desired effect. Anyway we observe a greater sensitivity of the results when we vary N with respect to g. In fact for $N \geq 1000$ we have the posterior means of K varying from 443 to 451, while when we vary g the posterior means of K are always 446.5 despite a wider range for the prior means in this setting.

To increase the difficulty of the deduplication problem in a situation where we know the exact matching configuration, we have also considered the `RLdata10000` data set. Figure 3 shows the box-plots of the posterior distributions of K, N, FNR and FDR for ten blocks of size 1000 with approximately 800 single clusters and 100 two-elements clusters. The hyperparameters values are $s^2 = 0.01$, $s_0^2 = 0.001$ $logit^{-1}(m_0) = 0.01$ and $g = 1.02$. Note that the true value of K (represented by a triangle) is always covered by the corresponding posterior drawings except for one block. Moreover, the posterior distributions of N partially overlap even when the related posterior for K are well

			$g = 1.01$				$g = 1.02$				$g = 1.05$			
$\frac{e^{m_0}}{1+e^{m_0}}$	$s^2 + s_0^2$	$\frac{s_0^2}{s^2+s_0^2}$	K	N	FNR	FDR	K	N	FNR	FDR	K	N	FNR	FDR
0.001	0.20	0.50	453.2	2550	0.065	0.001	453.2	2548	0.065	0.002	453.2	2554	0.065	0.002
0.001	0.60	0.17	452.8	2525	0.057	0.002	452.7	2524	0.057	0.003	452.7	2526	0.057	0.003
0.001	0.60	0.83	448.7	2307	0.024	0.052	448.7	2305	0.024	0.053	448.8	2308	0.024	0.050
0.001	1.00	0.50	448.4	2292	0.024	0.058	448.4	2296	0.024	0.058	448.4	2292	0.024	0.058
0.001	1.10	0.09	452.1	2487	0.048	0.005	452.1	2488	0.047	0.005	452.0	2485	0.046	0.006
0.001	1.10	0.91	152.2	159	0.040	0.940	151.2	158	0.039	0.941	151.6	159	0.039	0.940
0.001	1.50	0.33	448.0	2273	0.022	0.065	447.9	2274	0.022	0.066	447.9	2268	0.022	0.067
0.001	1.50	0.67	138.7	143	0.037	0.947	139.0	144	0.034	0.947	140.7	146	0.037	0.946
0.001	2.00	0.50	136.0	140	0.035	0.949	138.3	143	0.034	0.948	130.9	134	0.036	0.951
0.010	0.20	0.50	447.4	2244	0.016	0.068	447.5	2247	0.016	0.066	447.4	2245	0.016	0.067
0.010	0.60	0.17	446.5	2206	0.014	0.082	446.6	2209	0.015	0.080	446.5	2205	0.014	0.082
0.010	0.60	0.83	148.4	155	0.032	0.940	149.2	156	0.032	0.939	149.3	156	0.033	0.940
0.010	1.00	0.50	144.3	150	0.030	0.943	145.1	151	0.031	0.942	143.6	149	0.031	0.943
0.010	1.10	0.09	445.2	2145	0.011	0.104	445.1	2146	0.011	0.104	445.1	2139	0.010	0.105
0.010	1.10	0.91	129.2	132	0.042	0.950	131.2	135	0.042	0.949	127.2	130	0.043	0.951
0.010	1.50	0.33	140.6	145	0.029	0.947	141.2	146	0.031	0.946	141.9	147	0.028	0.945
0.010	1.50	0.67	122.5	125	0.042	0.954	120.9	123	0.041	0.956	117.4	119	0.042	0.957
0.010	2.00	0.50	119.1	121	0.042	0.956	120.7	123	0.045	0.956	110.9	112	0.044	0.960
0.020	0.20	0.50	442.2	2028	0.008	0.150	442.1	2023	0.008	0.154	442.1	2020	0.008	0.152
0.020	0.60	0.17	439.9	1944	0.008	0.189	439.9	1946	0.007	0.188	440.1	1948	0.007	0.185
0.020	0.60	0.83	139.3	144	0.035	0.945	140.9	146	0.035	0.944	142.0	147	0.033	0.943
0.020	1.00	0.50	134.8	139	0.034	0.948	132.3	136	0.034	0.949	135.0	139	0.033	0.948
0.020	1.10	0.09	436.4	1825	0.006	0.242	436.5	1831	0.007	0.241	436.3	1820	0.007	0.244
0.020	1.10	0.91	123.8	126	0.047	0.953	128.6	132	0.045	0.951	122.8	125	0.047	0.954
0.020	1.50	0.33	135.8	140	0.031	0.948	136.2	140	0.031	0.948	135.7	140	0.033	0.948
0.020	1.50	0.67	117.1	119	0.046	0.957	116.8	119	0.042	0.958	114.5	116	0.046	0.958
0.020	2.00	0.50	117.3	119	0.043	0.958	114.8	117	0.046	0.958	109.9	111	0.044	0.960

Table 3: `Rldata500` data set. Posterior means for K, N, the FNR and the FDR obtained when $logit^{-1}(m_0) = 0.01, 0.1, 0.2$, $s^2 = 0.1, 0.5, 1$ $s_0^2 = 0.1, 0.5, 1$ and $g = 1.01, 1.02, 1.05, 1.1, 1.5, 2$.

			$g = 1.1$				$g = 1.5$				$g = 2$			
$\frac{e^{m_0}}{1+e^{m_0}}$	$s^2 + s_0^2$	$\frac{s_0^2}{s^2+s_0^2}$	K	N	FNR	FDR	K	N	FNR	FDR	K	N	FNR	FDR
0.001	0.20	0.50	453.2	2551	0.066	0.001	453.2	2525	0.066	0.001	453.1	2505	0.064	0.002
0.001	0.60	0.17	452.7	2517	0.057	0.003	452.7	2503	0.058	0.003	452.7	2480	0.057	0.002
0.001	0.60	0.83	448.7	2306	0.025	0.052	448.7	2289	0.025	0.051	448.7	2269	0.025	0.052
0.001	1.00	0.50	448.4	2292	0.024	0.057	448.4	2273	0.023	0.058	448.4	2254	0.024	0.058
0.001	1.10	0.09	452.1	2486	0.047	0.005	452.2	2471	0.049	0.006	452.2	2448	0.048	0.005
0.001	1.10	0.91	141.4	146	0.039	0.946	147.3	153	0.038	0.942	150.7	158	0.040	0.941
0.001	1.50	0.33	448.0	2271	0.022	0.065	447.9	2252	0.022	0.067	447.9	2238	0.022	0.066
0.001	1.50	0.67	138.7	143	0.035	0.947	134.8	139	0.037	0.949	139.4	144	0.036	0.947
0.001	2.00	0.50	133.3	137	0.034	0.950	139.0	144	0.036	0.948	129.7	133	0.038	0.952
0.010	0.20	0.50	447.4	2246	0.016	0.067	447.5	2229	0.016	0.067	447.5	2214	0.016	0.066
0.010	0.60	0.17	446.6	2207	0.014	0.081	446.6	2192	0.014	0.080	446.5	2171	0.013	0.082
0.010	0.60	0.83	148.3	155	0.033	0.940	145.9	152	0.033	0.941	150.7	157	0.032	0.939
0.010	1.00	0.50	145.7	151	0.030	0.942	142.6	148	0.032	0.943	138.1	143	0.033	0.946
0.010	1.10	0.09	445.2	2141	0.011	0.106	445.0	2122	0.011	0.107	445.0	2106	0.011	0.108
0.010	1.10	0.91	130.9	134	0.040	0.949	126.4	129	0.043	0.952	126.7	130	0.044	0.951
0.010	1.50	0.33	139.2	144	0.031	0.947	141.7	147	0.030	0.947	143.1	148	0.030	0.945
0.010	1.50	0.67	120.8	123	0.043	0.955	119.0	121	0.044	0.956	114.7	116	0.045	0.957
0.010	2.00	0.50	117.3	119	0.041	0.958	114.1	116	0.042	0.959	114.7	116	0.042	0.959
0.020	0.20	0.50	442.3	2024	0.007	0.150	442.1	2009	0.008	0.152	442.1	1991	0.008	0.153
0.020	0.60	0.17	440.0	1940	0.007	0.188	440.0	1929	0.007	0.187	440.0	1916	0.007	0.188
0.020	0.60	0.83	140.8	146	0.035	0.944	140.5	145	0.035	0.944	142.3	147	0.034	0.943
0.020	1.00	0.50	135.7	140	0.035	0.947	135.2	139	0.036	0.948	136.3	140	0.032	0.946
0.020	1.10	0.09	436.5	1825	0.007	0.242	436.5	1816	0.007	0.242	435.8	1787	0.007	0.252
0.020	1.10	0.91	123.2	126	0.044	0.953	126.5	129	0.043	0.951	120.7	123	0.046	0.955
0.020	1.50	0.33	131.0	134	0.030	0.951	134.7	139	0.031	0.948	130.7	134	0.033	0.951
0.020	1.50	0.67	115.5	117	0.048	0.958	112.4	114	0.046	0.958	120.4	123	0.044	0.954
0.020	2.00	0.50	107.2	108	0.045	0.962	111.8	113	0.046	0.960	112.1	114	0.043	0.959

Table 3: Continued.

N	K	FNR	FDR
250	227	0.02	0.89
500	389	0.01	0.63
1000	443	0.01	0.14
2500	447	0.01	0.08
5000	448	0.02	0.06
10000	449	0.02	0.05
100000	451	0.05	0.02

Table 4: `Rldata500` data set. Posterior means for K, FNR and the FDR conditional on fixed values of N.

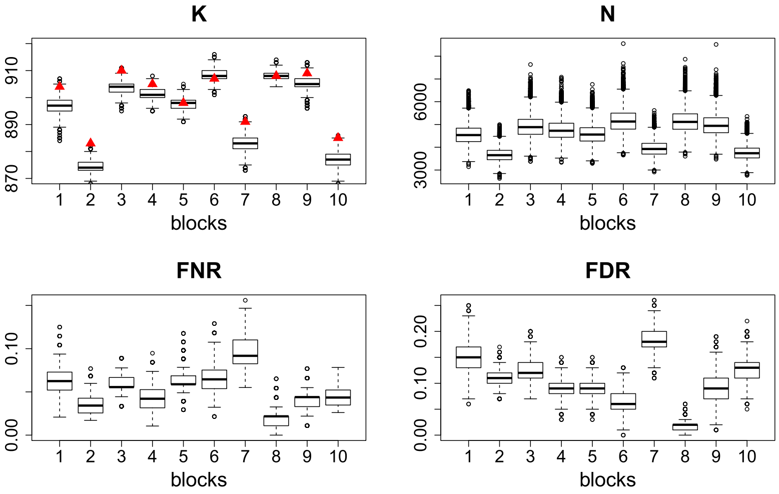

Figure 3: `Rldata10000` data set. Boxplots of the posterior distributions of K, N, FNR and FDR for ten blocks of size 1000. The triangles represent the true values of K. The hyperparameters values are $s^2 = 0.5$, $s_0^2 = 0.1$ $logit^{-1}(m_0) = 0.001$ and $g = 1.02$.

separated confirming the robustness of population size inference when we account for matching uncertainty. Finally record linkage performances are quite satisfactory with posterior medians for FNR and FDR respectively less than 0.07 and 0.15 except for one block

7 Application with Syrian data

As a real application we now face the problem of matching records from two public available data sets reporting different number of recorded victims killed in the recent Syrian conflict, along with available identifying information including first and family names, date of death, and death location. A more detailed application can be found in Chen et al. (2018). Here we consider the data provided by the Violations Documentation

		Cluster size			
Analysis	Data set	1	2	3	4
Separated lists	VDC	1582.35	49.66	3.97	0.10
	CSR	916.02	39.34	2.07	0.43
Joined lists	VDC and CSR	1588.88	482.78	43.06	1.60
	VDC	1519.14	77.01	5.91	0.63
	CSR	899.89	46.00	2.60	0.48
Record linkage	VDC and CSR	1833.25	431.52	0.00	0.00

Table 5: Syrian data. Distribution of the cluster sizes averaged across MCMC simulations.

Center in Syria (VDC) and the Syrian Center for Statistics and Research (CSR) and we focus on the killings in the province of Raqqa from the beginning of the conflict until March 2017, since the CSR data set does not report records after this date.

The VDC data set provides directly the English equivalents of the Arabic names while, for the CSR list, the English equivalents have been obtained by software transliteration of the reported Arabic names causing additional noise. Several records of the VDC data set represent unidentified victims and report only the date of death or do not have the first name and report only the relationship with the head of the family. All these records have been eliminated and the resulting VDC data sets comprises 1694 records. The CSR list presents only completely identified victims for a total size of 1003 records. As in the previous experiments first and family names have been transformed by the English version of the soundex algorithm and the resulting fields have been considered as key variables together with year, month and day of death for a total of 11 variables.

We show the results obtained with the same hyperparameters set for the `Rldata10000` data set and considering three different analyses. In the first case, that we call separated list analysis, we investigate only the within list deduplication problem. Hence we fit our model to the single lists one by one. Note that identification of true within list duplicates is a very challenging problem with these data since most attacks killed whole families causing records differing only in the first name that may easily confused as duplicates. Anyway the number of record pairs that exceed a 0.5 posterior probability of being duplicates, $p(\eta_{ij_1} = \eta_{ij_2}|v)$, is small. In fact we have 51 pairs in the first data set and 43 in the second one hence visual inspection of these pairs may eventually confirm their matching status. Table 5 reports the distribution of the of cluster sizes averaged across the MCMC simulations showing the microcluster effect for both the lists.

In the second analysis we consider both within and between lists de-duplication, that is the natural scenario for our model where the two lists are joined into a single data set. The total number of pairs with $p(\eta_{i_1j_1} = \eta_{i_2j_2}|v) > 0.5$ is 617 out of which 481 are between lists duplicates and 84 and 52 are respectively within the first and the second list. Hence about 78% of duplicates link the same victim across the two lists. Table 5 shows the distribution of the cluster sizes for the joined lists but also within the two lists separately. Note the cluster size distribution within the two lists are quite similar to the previous case where the lists are separated before fitting the model.

In the third analysis we exclude the within list duplications and we consider only the record linkage problem across the two lists. One way to adapt our proposed model to that particular case is to modify the prior distribution on the λ's such that $\eta_{ij_1} \neq \eta_{ij_2}$ $\forall j_1 \neq j_2$ and for $i = 1, 2$. Note that, in this case, clusters consist of at most two elements so that the distribution of the observed records v, conditional on η and α, can be calculated analytically without exploiting the recursive formula. Moreover, the above conditioning is equivalent to assuming that the two lists are two simple random samples without replacement from a population of N units.

This is the same situation described in Tancredi and Liseo (2011). From a computational perspective, this scenario does not imply substantial changes. In fact, we can arbitrarily fix the labels of the first file, for example by assuming that $\eta_{1j} = j$ for $j = 1, \dots, n_1$ and update only the labels of the second file. In particular, indicating with m_q the size of the cluster identified by the label q without the record (i, j) we can use the Gibbs step provided by equation (5.5) by setting

$$p(\eta_{2j} = q | \eta_{-(2j)}) \propto \begin{cases} 1 & \text{if } q \leq n_1 \text{ and } m_q = 1 \\ 0 & \text{if } q \leq n_1 \text{ and } m_q = 2 \\ 0 & \text{if } q > n_1 \text{ and } m_q = 1 \end{cases}$$

and

$$p(\eta_{2j} = new | \eta_{-(2j)}) \propto (N - k_{-(2j)})$$

The number of pairs with $p(\eta_{1j_1} = \eta_{2j_2} | v) > 0.5$ in the record linkage framework is 423 and the posterior mean of the match number, that is the frequency of the two elements clusters, is 431.52. The reduced number of matches between the lists with respect to the previous case is due to a larger estimate of the measurement error when within list duplications are taken into account with the consequent increase also of between lists estimated duplications.

Finally, Figure 4 shows the posterior distribution for N provided by the three different data analyses described above. Notice that, since we eliminated records with missing information from the first list, here N represents the size of a smaller population than all the victims killed in the province of Raqqa until March 2017. We may say that N represents the size of victims with *recordable* information about first and last name. The posterior mean for N, when accounting for duplications both within and between the lists is equal to 5350 while when accounting only for between lists duplications is equal to 7507. When considering the two lists separately the posterior mean of N increases considerably to 24832 with the VDC list and to 11116 with the CSR list. Anyway the former two estimates are more reliable for the additional informative content obtained by joining the two lists. Note also that our estimates depend on the information retrieved on the original records via the soundex algorithm and that adding other key variables or using the full Arabic names with suitable string distance may lead to different estimates. Moreover population size estimates are strongly dependent on the capture-recapture model specifications, hence introducing heterogeneous and/or

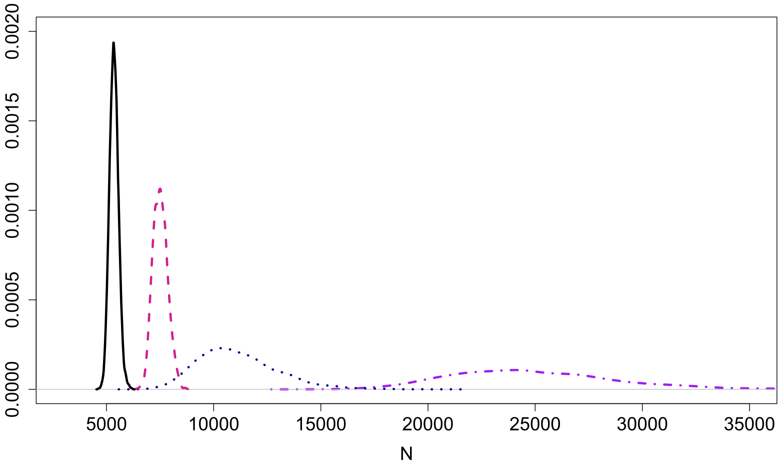

Figure 4: Syrian data. Posterior distribution for N obtained joining the CSR and VDC lists into a single data set (solid line), via a record linkage analysis without within list duplications (dashed line) and the CSR (dot-dashed line) and VDC (dotted line) single list analyses.

dependent captures may also produce different estimates. However our estimates can be seen as a starting point for future comparisons.

8 Discussion

In this paper we have shown how population size estimation can be performed when records related to population units have been sampled and duplicated across multiple files and the matching reconstruction within the same file and across different files is uncertain. In particular, through the prior specification of the matching process, we assumed that the observed lists are obtained as independent simple random sampling with replacement from a closed population of unknown size N. The hit-and-miss model (Copas and Hilton, 1990) has been used as a measurement error model in order to interpret differences among the sample records and the population records.

As a by-product of this approach, we obtained a more adequate prior distribution for the matching pattern, which can also be used when the population size estimation is not the primary task of the de-duplication process. However, more sophisticated prior distributions could be used to incorporate more realistic sampling design. For example, it would be important to extend our approach by introducing both heterogeneity and dependence in the sampling probability of the population units as in usual capture-recapture models. In particular the independence among the L lists is a very strong assumption which rarely occurs in real applications. Note also that, in the de-duplication framework, the problem is even more involved, because we may have different degrees of dependence among captures and duplications across the lists. Moreover, from a theoretical perspective, it would be also worthwhile to investigate the role that different prior distributions on the partition space, like that one induced by the Pitman-Yor process, may play in the facilitation of the microclustering effect.

Other specific assumptions that we made throughout the paper concern the independence of the key variables at the population level and the conditional independence of the measurement error mechanism. Also in this case, more sophisticated versions of the hit-and-miss model together with an appropriate model for the key variables should be used to take into account more realistic scenarios. Anyway, we are confident that our framework may provide a basis for all these kinds of extensions.

Supplementary Material

Supplementary Material for "A Unified Framework for De-Duplication and Population Size Estimation"
(DOI: 10.1214/19-BA1146SUPP; .pdf).

References

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. 2

Booth, J. G., Casella, G., and Hobert, J. P. (2008). "Clustering using objective functions and stochastic search." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70(1): 119–139. MR2412634. doi: https://doi.org/10.1111/j.1467-9868.2007.00629.x. 10

Briscolini, D., Di Consiglio, L., Liseo, B., Tancredi, A., and Tuoto, T. (2018). "New methods for Small Area Estimation with Linkage Uncertainty." *International Journal of Approximate Reasoning*, 94: 30–42. MR3760873. doi: https://doi.org/10.1016/j.ijar.2017.12.005. 2

Chen, B., Shrivastava, A., and Steorts, R. C. (2018). "Unique Entity Estimation with Application to the Syrian Conflict." *Annals of Applied Statistics*, 12: 1039–1067. MR3834294. doi: https://doi.org/10.1214/18-AOAS1163. 20

Copas, J. and Hilton, F. (1990). "Record linkage: statistical models for matching computer records." *Journal of the Royal Statistical Society, A*, 153: 287–320. 2, 4, 23

Devroye (1986). *Non-Uniform Random Variate Generation*. Springer-Verlag. MR0836973. doi: https://doi.org/10.1007/978-1-4613-8643-8. 7

Fellegi, I. and Sunter, A. (1969). "A theory of record linkage." *Journal of the American Statistical Association*, 64: 1183–1210. 2

Fortini, M., Liseo, B., Nuccitelli, A., and Scanu, M. (2001). "On Bayesian record linkage." *Research in Official Statistics*, 4: 185–198. 2

George, E. I. and Robert, C. P. (1992). "Capture recapture estimation via Gibbs sampling." *Biometrika*, 79(4): 677–683. MR1209469. doi: https://doi.org/10.2307/2337223. 7

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. 2

Johndrow, J. E., Lum, K., and Dunson, D. B. (2018). "Theoretical limits of record linkage and microclustering." *Biometrika*, 105: 431–446. MR3804412. doi: https://doi.org/10.1093/biomet/asy003. 7, 16, 17

Larsen, M. (2005). "Advances in Record Linkage Theory: Hierarchical Bayesian Record Linkage Theory." *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 3277–3283. 2

Larsen, M. D. and Rubin, D. (2001). "Iterative automated record linkage using mixture models." *Journal of the American Statistical Association*, 96: 32–41. MR1973781. doi: https://doi.org/10.1198/016214501750332956. 2

Liseo, B. and Tancredi, A. (2011). "Bayesian estimation of population size via linkage of multivariate normal data sets." *Journal of Official Statistics*, 27: 491–505. 2

MacEachern, S. N. (1994). "Estimating normal means with a conjugate style Dirichlet process prior." *Communications in Statistics-Simulation and Computation*, 23(3): 727–741. MR1293996. doi: https://doi.org/10.1080/03610919408813196. 14

Marin, J.-M. and Robert, C. P. (2014). *Bayesian essentials with R*. Springer. MR3136532. doi: https://doi.org/10.1007/978-1-4614-8687-9. 8

McCullagh, P. and Yang, J. (2008). "How many clusters?" *Bayesian Analysis*, 3(1): 101–120. MR2383253. doi: https://doi.org/10.1214/08-BA304. 10

Neal, R. M. (2000). "Markov chain sampling methods for Dirichlet process mixture models." *Journal of Computational and Graphical Statistics*, 9: 249–265. MR1823804. doi: https://doi.org/10.2307/1390653. 14

Papaspiliopoulos, O., Roberts, G., and Skold, M. (2003). "Non-centered parameterisations for hierarchical models and data augmentation." Bayesian Statistics 7: Proceedings of the Seventh Valencia International Meeting, vol. 307, Oxford University Press, USA MR2003180. 14

Pitman, J. (2006). *Combinatiorial Stochastic Processes*. Ecole d'Eté de Probabilités de Saint-Flour XXXII, Lecture Notes in Mathematics, vol. 1875, Berlin, Springer. MR2245368. 7

Sadinle, M. (2014). "Detecting duplicates in a homicide registry using a Bayesian partitioning approach." *The Annals of Applied Statistics*, 8(4): 2404–2434. MR3292503. doi: https://doi.org/10.1214/14-AOAS779. 2

Sadinle, M. (2017). "Bayesian Estimation of Bipartite Matchings for Record Linkage." *Journal of the American Statistical Association*, 112: 600–612. MR3671755. doi: https://doi.org/10.1080/01621459.2016.1148612. 2

Sadinle, M. (2018). “Bayesian propagation of record linkage uncertainty into population size estimation of human rights violations.” *The Annals of Applied Statistics*, 12(2): 1013–1038. MR3834293. doi: https://doi.org/10.1214/18-AOAS1178. 2

Sadinle, M. and Fienberg, S. E. (2013). “A generalized Fellegi–Sunter framework for multiple record linkage with application to homicide record systems.” *Journal of the American Statistical Association*, 108(502): 385–397. MR3174628. doi: https://doi.org/10.1080/01621459.2012.757231. 2

Steorts, R. C. (2015). “Entity Resolution with Empirically Motivated Priors.” *Bayesian Analysis*, 10(4): 849–875. MR3432242. doi: https://doi.org/10.1214/15-BA965SI. 3, 4, 16

Steorts, R. C., Hall, R., and Fienberg, S. E. (2014). “SMERED: A Bayesian Approach to Graphical Record Linkage and De-duplication.” *Journal of Machine Learning Research*, 33: 922–930. 2, 4

Steorts, R. C., Hall, R., and Fienberg, S. E. (2016). “A Bayesian approach to graphical record linkage and de-duplication.” *Journal of the American Statistical Association: Theory and Methods*, 111(516): 1660–1672. MR3601725. doi: https://doi.org/10.1080/01621459.2015.1105807. 2, 3, 4, 5

Tancredi, A. and Liseo, B. (2011). “A hierarchical Bayesian approach to record linkage and population size problems.” *Annals of Applied Statistics*, 5: 1553–1585. MR2849786. doi: https://doi.org/10.1214/10-AOAS447. 2, 3, 5, 22

Tancredi, A. and Liseo, B.(2015). “Regression Analysis with linked data: Problems and possible solutions.” *Statistica*, 75(1): 19–35. 2

Tancredi, A., Steorts, R., and Liseo, B. (2019). “Supplementary Material for “A Unified Framework for De-Duplication and Population Size Estimation”.” *Bayesian Analysis*. doi: https://doi.org/10.1214/19-BA1146. 7, 11

Wang, X., He, C. Z., and Sun, D. (2007). “Bayesian population estimation for small sample capture-recapture data using noninformative priors.” *Journal of Statistical Planning and Inference*, 137(4): 1099–1118. MR2301466. doi: https://doi.org/10.1016/j.jspi.2006.03.004. 7

Zanella, G., Betancourt, B., Wallach, H., Miller, J., Zaidi, A., and Steorts, R. C. (2016). “Flexible Models for Microclustering with Application to Entity Resolution.” *Neural Information Processing Systems*. 7

Acknowledgments

The authors are grateful to an associate editor and two referees for helpful comments and suggestions. A. Tancredi and B. Liseo are supported by MIUR, PRIN project 2015EASZFS-PE1.