Bayesian Analysis (2020) **15**, Number 2, pp. 633–682

A Unified Framework for De-Duplication and Population Size Estimation (with Discussion)

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

© 2020 International Society for Bayesian Analysis 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

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

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

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

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 $k = N\left(1 - e^{-n/N}\right)$ 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.

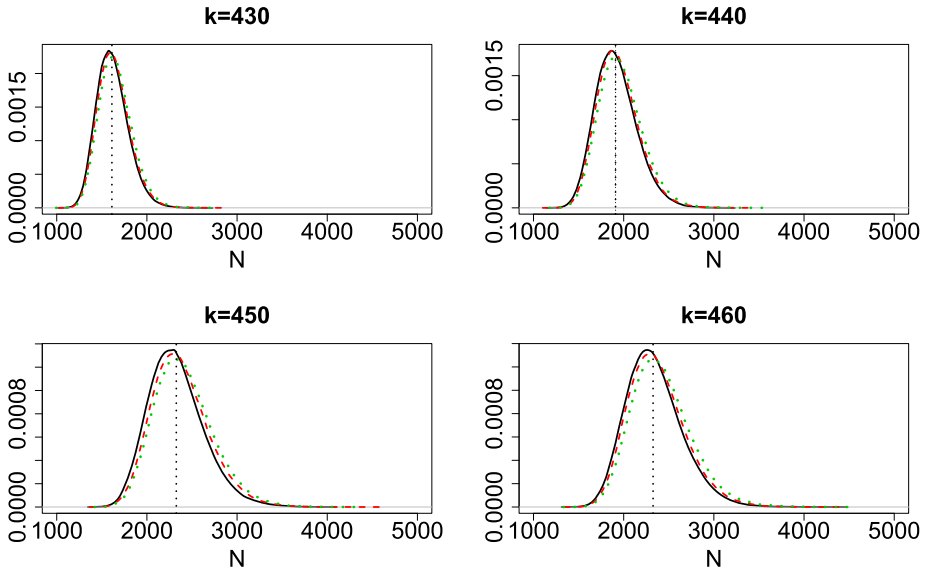

Figure 1: Posterior distributions for N when $n = 500$, k is known and equal to 430,440,450,460 and $p(N) \propto 1/N^g$ with $g = 0$ (dotted line), $g = 1$ dashed line and $g = 2$ solid line.

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 $j = 1, \ldots, L$, 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

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

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

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

Assuming duplication and capture independence across the lists, we also have that $p(\lambda|n_1, \ldots n_L) = \frac{1}{N^n}$. 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

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

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,

$$\begin{aligned} p(v_z|Z,U,N,\alpha,\theta) &= \prod_{\ell=1}^{p}\big[\delta(v_{i_1j_1\ell},v_{i_2j_2\ell})\theta_{\ell\, v_{i_1j_1\ell}}(1-\alpha_{u_z\ell})^2 + \\ &\quad (2\alpha_{u_z\ell}-\alpha_{u_z\ell}^2)\theta_{\ell\, v_{i_1j_1l}}\theta_{\ell\, v_{i_2j_2\ell}}\big]. \end{aligned}$$

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, $z = \{(i_1\,j_1),\ldots,(i_n\,j_n)\}$:

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

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

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

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

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

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

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

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 $j' = 1, \ldots, n$ 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

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

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 $\{1, \ldots, n\}$ and no longer in $\{1, \ldots, N\}$ 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

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

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

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

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

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

for $i=1,\dots,L, j=1,\dots,n_i$, 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

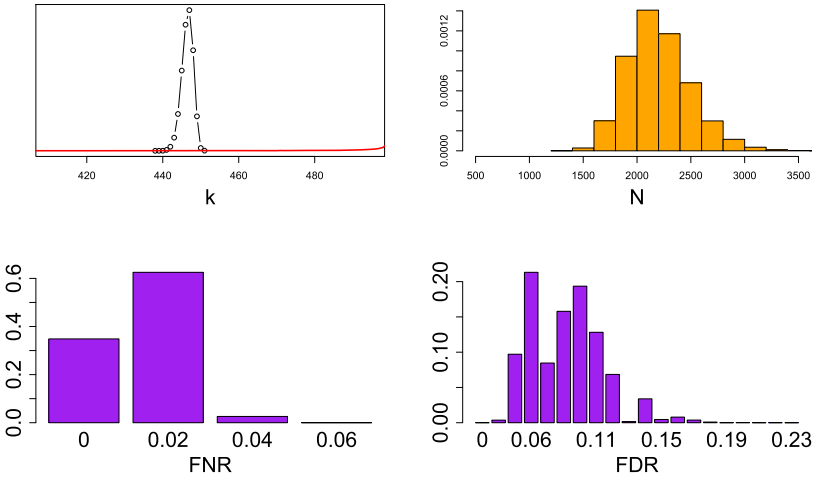

Figure 2: `RLdata500` data set. Prior and posterior distributions of K and posterior distribution of N, FNR and FDR when $s^2 = 0.5$, $m_0 = logit(0.01)$ and $s^2 = 0.1$ and $g = 1.02$.

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

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

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.

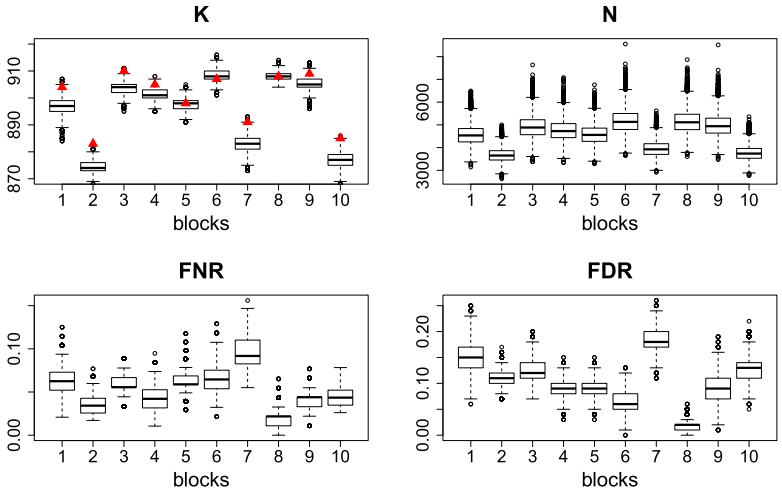

Figure 3: `Rldata10000` data set. Boxplots of the posterior distributions of K, N, FNR and FDR for ten blocks of size 1000. The triangles represent the true values of K. The hyperparameters values are $s^2 = 0.5$, $s_0^2 = 0.1$ $logit^{-1}(m_0) = 0.001$ and $g = 1.02$.

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 $j = 1, \ldots, n_1$ 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

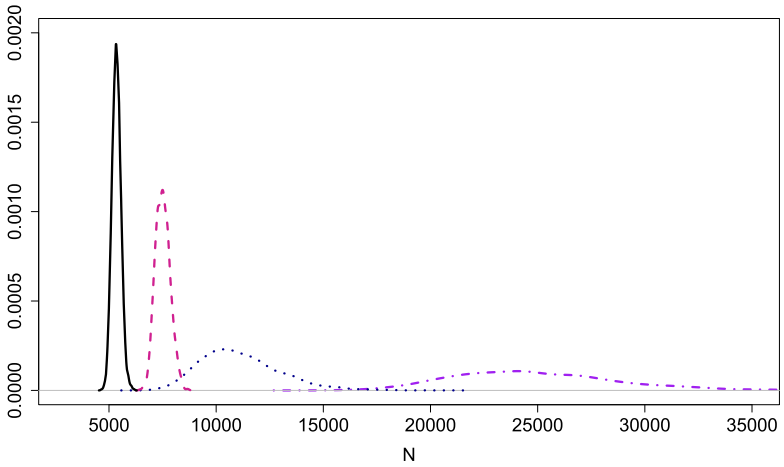

Figure 4: Syrian data. Posterior distribution for N obtained joining the CSR and VDC lists into a single data set (solid line), via a record linkage analysis without within list duplications (dashed line) and the CSR (dot-dashed line) and VDC (dotted line) single list analyses.

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 theoret-

ical 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. 634

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. 642

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. 634

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. 652

Copas, J. and Hilton, F. (1990). "Record linkage: statistical models for matching computer records." *Journal of the Royal Statistical Society, A*, 153: 287–320. 634, 636, 655

Devroye (1986). *Non-Uniform Random Variate Generation*. Springer-Verlag. MR0836973. doi: https://doi.org/10.1007/978-1-4613-8643-8. 639

Fellegi, I. and Sunter, A. (1969). "A theory of record linkage." *Journal of the American Statistical Association*, 64: 1183–1210. 634

Fortini, M., Liseo, B., Nuccitelli, A., and Scanu, M. (2001). "On Bayesian record linkage." *Research in Official Statistics*, 4: 185–198. 634

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. 639

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. 634

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. 639, 648, 649

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. 634

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. 634

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. 634

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. 646

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. 640

McCullagh, P. and Yang, J. (2008). "How many clusters?" *Bayesian Analysis*, 3(1): 101–120. MR2383253. doi: https://doi.org/10.1214/08-BA304. 642

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. 646

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. 646

Pitman, J. (2006). *Combinatiorial Stochastic Processes*. Ecole d'Eté de Probabilités de Saint-Flour XXXII, Lecture Notes in Mathematics, vol. 1875, Berlin, Springer. MR2245368. 639

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. 634

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. 634

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. 634

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. 634

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. 634, 636, 648

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. 634, 636

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. 634, 635, 636, 637

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. 634, 635, 637, 654

Tancredi, A. and Liseo, B.(2015). "Regression Analysis with linked data: Problems and possible solutions." *Statistica*, 75(1): 19–35. 634

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. 639, 643

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. 639

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*. 639

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.

Invited Discussion

Mauricio Sadinle*

Population size estimation techniques, such as multiple-systems or capture-recapture estimation, typically require multiple samples from the study population, in addition to the information on which individuals are included in which samples. In many contexts, these samples come from existing data sources that contain certain information on the individuals but no unique identifiers. The goal of record linkage and duplicate detection techniques is to identify unique individuals across and within samples based on the information collected on them, which might correspond to basic partial identifiers, such as given and family name, and other demographic information. Therefore, record linkage and duplicate detection are often needed to generate the input for a *given* population size estimation technique that a researcher might want to use. Linkage decisions, however, are subject to uncertainty when partial identifiers are limited or contain errors and missingness, and therefore, intuitively, uncertainty in the linkage and deduplication process should somehow be taken into account in the stage of population size estimation.

The contributions of the discussed article build on a framework initially proposed by Hall et al. (2012) for linking multiple datafiles, later extended by Steorts et al. (2014, 2016) to also handle duplication within datafiles. The framework of Hall et al. (2012) and Steorts et al. (2014, 2016) also partially coincides with the work of Tancredi and Liseo (2011) in the case of two duplicate-free datafiles. As presented in the discussed article, this framework can be summarized in terms of a generative process where a finite population is generated from a super-population, and in turn the records are generated from the finite population, as follows:[1]

- The super-population:
  - The random vector $\tilde{\boldsymbol{V}} = (\tilde{V}_1, \ldots, \tilde{V}_p)$ represents p categorical variables which follow a product multinomial distribution $\prod_{\ell=1}^{p} \text{MN}(1, \boldsymbol{\theta}_\ell)$, meaning that the entries of $\tilde{\boldsymbol{V}}$ are mutually independent, with $\tilde{V}_\ell \in \{1, \ldots, M_\ell\}$ being multinomial $\text{MN}(1, \boldsymbol{\theta}_\ell)$, with $\boldsymbol{\theta}_\ell = (\theta_{\ell 1}, \ldots, \theta_{\ell M_\ell})$.
- The finite population:
  - A finite population of N individuals is generated as a random sample from the super-population, that is, $\tilde{\boldsymbol{V}}_1, \ldots, \tilde{\boldsymbol{V}}_N \overset{iid}{\sim} \prod_{\ell=1}^{p} \text{MN}(1, \boldsymbol{\theta}_\ell)$. The realization $\tilde{\boldsymbol{v}}_{j'}$ of $\tilde{\boldsymbol{V}}_{j'}$ represents the true underlying values of an individual $j' = 1, \ldots, N$.
- The records:

*Department of Biostatistics, University of Washington, Seattle, msadinle@uw.edu

[1] Here I have slightly modified the original notation of the authors.

- Record j in datafile i, henceforth indexed as ij, is a realization of a random vector $\boldsymbol{V}_{ij} = (V_{ij1}, \ldots, V_{ijp})$, where entry ℓ, $V_{ij\ell}$, is a potentially erroneous measurement of characteristic ℓ of an individual in the finite population. There are $i = 1, \ldots, L$ datafiles, and $j = 1, \ldots, n_i$ records in datafile i.
- Denoting $\lambda_{ij} \in \{1, \ldots, N\}$ as an index that indicates the individual in the population to which record ij refers, we can denote $\tilde{\boldsymbol{v}}_{\lambda_{ij}} = (\tilde{v}_{\lambda_{ij}1}, \ldots, \tilde{v}_{\lambda_{ij}p})$ as the true underlying vector of characteristics of the individual to which record ij refers.
- The different records $\{\boldsymbol{V}_{ij}\}_{ij}$ are assumed to arise independently of each other across and within datafiles and across individuals, conditioning on the realized values $\tilde{\boldsymbol{v}}_1, \ldots, \tilde{\boldsymbol{v}}_N$ of $\tilde{\boldsymbol{V}}_1, \ldots, \tilde{\boldsymbol{V}}_N$ in the finite population.
- The entries of each record $\boldsymbol{V}_{ij}$ are generated independently of each other, conditioning on the true underlying values of individual λ_{ij} as follows: with probability $\alpha_{\lambda_{ij}\ell}$ take $V_{ij\ell}$ to be the true value $\tilde{v}_{\lambda_{ij}\ell}$, otherwise, draw a random value $V_{ij\ell} \in \{1, \ldots, M_\ell\}$ with the same probabilities $(\theta_{\ell 1}, \ldots, \theta_{\ell M_\ell})$ as in the super-population. This is known as the *hit-miss* model of Copas and Hilton (1990). Originally in the work of Hall et al. (2012) and Steorts et al. (2014, 2016), the probability of correct measurement $\alpha_{\lambda_{ij}\ell}$ was taken to be a single value α_ℓ for all individuals.

The model structure above requires imposing prior distributions on $\boldsymbol{\theta}_\ell$ for $\ell = 1, \ldots, p$; on $\alpha_{j'\ell}$ for $j' = 1, \ldots, N$ and $\ell = 1, \ldots, p$; and more crucially on $\boldsymbol{\lambda} = (\lambda_{11}, \ldots, \lambda_{1n_1}, \ldots, \lambda_{L1}, \ldots, \lambda_{Ln_L})$. We shall focus our discussion on $\boldsymbol{\lambda}$, as treatment of the other parameters is somewhat standard.

The parameter $\boldsymbol{\lambda}$ gives us the information that we need to link records: if $\lambda_{ij} = \lambda_{i'j'}$ then records ij and $i'j'$ refer to the same underlying individual; furthermore, if $i = i'$ then these records are duplicates of each other within datafile i, and if $i \neq i'$ then they represent the same individual in datafiles i and i'. Crucially, notice that the labels that we use to represent $\boldsymbol{\lambda}$ are not themselves relevant for linking records, as all we end up using is whether $\lambda_{ij} = \lambda_{i'j'}$ or $\lambda_{ij} \neq \lambda_{i'j'}$. In other words, the information that we need to extract from $\boldsymbol{\lambda}$ is the partition of the records that it induces; records that receive the same label represent records that refer to the same underlying individual. This means that $\boldsymbol{\lambda}$ is actually a labeling of the partition of the records that we are interested in recovering in record linkage and duplicate detection problems.

The prior on $\boldsymbol{\lambda}$ used by Hall et al. (2012) and Steorts et al. (2014, 2016) is uniform across all the possible values of $\boldsymbol{\lambda}$, which depends on the number of labels used to represent this vector. In Hall et al. (2012) and Steorts et al. (2014), this number of labels was taken to be $n = \sum_{i=1}^{L} n_i$, the number of records in all datafiles, which is the number of labels needed to represent the partition if all the records across all datafiles correspond to different individuals. In Section 6.2 of Steorts et al. (2016) this number of labels was allowed to be $M \geq n$, and it was used as a hyper-parameter to be selected based on the prior that it induces for the number of clusters of the partition of the records. In the data-generative process above, the population size N provides the

number of labels available for labeling the partition, and so the role of N is identical to the role of M in Steorts et al. (2016).

The main contribution of the discussed article is that now we see the number of labels N (or M in Steorts et al. 2016) as a population size that also needs to be estimated, and therefore the authors use a prior distribution for it, $p(N) \propto N^{-g}$, $g > 1$. The authors provided two ways of motivating this approach to population size estimation. First, if the partition of the datafiles is known, then this approach corresponds to a capture-recapture model where there is one capture occasion for each record in the datasets. Second, the uniform prior on the possible values of λ can be derived from a data-generating process where each individual in the finite population is subject to a Poisson-distributed datafile-specific random number of capture attempts, each of them being successful with a certain datafile-specific probability.

While it is interesting to see that the authors' approach can be justified from such an idealized data collection process, it is important to keep in mind that developments in this area of research should accommodate commonly used capture-recapture models. Unfortunately, the capture-recapture component of the approach proposed by the authors does not correspond to any commonly used model in this area, and its assumptions are too stringent to be practically useful, as acknowledged by the authors. The authors also argue that this can be seen as a starting point for more complicated approaches, but as we all know, *the devil is in the details*, and so it is not clear whether the contribution of the discussed article will straightforwardly enable the creation of joint modeling approaches that combine the record linkage approach of Hall et al. (2012) and Steorts et al. (2014, 2016) with more commonly used population size estimation models.

In the future it would be interesting to see joint modeling approaches that combine existing record linkage frameworks with commonly used capture-recapture models. To judge the relevance of such future contributions, an important requirement should be that the new joint models simplify to more flexible or commonly used models for population size estimation when the partition of the datafiles is known. For example, the approaches of Tancredi and Liseo (2011) and Liseo and Tancredi (2011) do satisfy this criterion, as they incorporate the commonly used hypergeometric capture-recapture model for two samples. Along these lines, it would have been good, for example, that the capture-recapture component of the proposed approach in the discussed article at least simplified to the hypergeometric capture-recapture model in the approach of Tancredi and Liseo (2011) in the case of two duplicate-free datafiles.

Extensions of the proposed modeling approach that accommodate more realistic capture-recapture models will lead to new posteriors $p(N, \boldsymbol{Z} \mid \tilde{\boldsymbol{v}})$ on the population size N and partition of the datafiles $\boldsymbol{Z}$ for each different capture-recapture model being considered, which will require approximation via new sampling algorithms (e.g., Markov chain Monte Carlo). While we could imagine having a plethora of articles where each possible model for record linkage is combined with each possible model for population size estimation, more meaningful contributions to this literature should develop very general approaches that allow users to obtain specific commonly-used capture-recapture models as particular cases.

Finally, we note that instead of developing joint models, an alternative endeavor is to develop approaches that allow users to re-use record linkage results for different capture-recapture models. An example of such an approach is *linkage-averaging*, proposed by Sadinle (2018). Linkage-averaging requires a Bayesian record linkage methodology that provides a posterior distribution on partitions of the datafiles, such as the approach of Hall et al. (2012) and Steorts et al. (2014, 2016) or the approach of Sadinle (2014, 2017), and a capture-recapture model with sufficient statistics that depend only on the overlap of the datafiles in terms of their individuals, such as the approaches of Fienberg (1972); Bishop et al. (1975); Castledine (1981); George and Robert (1992); Madigan and York (1997); Fienberg et al. (1999); Manrique-Vallier (2016). Linkage-averaging is a simple two-stage approach in which a capture-recapture model is used to obtain a posterior distribution of the population size for each partition of the datafile obtained from a posterior sample of partitions from a Bayesian record linkage model; the partition-specific population size posteriors are then averaged. Under some conditions the results of that approach can be shown to correspond to those of a proper Bayesian joint model for both record linkage and population size estimation. One of the advantages of a two-stage approach is that the results from the linkage stage can be re-used for different capture-recapture models, and therefore it facilitates model exploration and avoids having to derive new posterior sampling algorithms for each combination of record linkage and capture-recapture model. Nevertheless, linkage-averaging has certain limitations, as, for example, its performance depends on the quality of both the linkage and the capture-recapture models being used, and it does not support capture-recapture models with covariates.

Regardless of whether new developments come in the form of joint models or two-stage approaches, I believe that the applicability of the new approaches to realistic scenarios should be an important consideration in their overall evaluation.

References

Bishop, Y. M., Fienberg, S. E., and Holland, P. W. (1975). *Discrete Multivariate Analysis: Theory and Practice*. The MIT Press. Reprinted in 2007 by Springer, New York. MR0381130. 662

Castledine, B. J. (1981). "A Bayesian Analysis of Multiple-Recapture Sampling for a Closed Population." *Biometrika*, 68(1): 197–210. MR0614956. doi: https://doi.org/10.1093/biomet/68.1.197. 662

Copas, J. B. and Hilton, F. J. (1990). "Record Linkage: Statistical Models for Matching Computer Records." *Journal of the Royal Statistical Society. Series A (Statistics in Society)*, 153(3): 287–320. 660

Fienberg, S. E. (1972). "The Multiple Recapture Census for Closed Populations and Incomplete 2^k Contingency Tables." *Biometrika*, 59(3): 591–603. MR0383619. doi: https://doi.org/10.1093/biomet/59.3.591. 662

Fienberg, S. E., Johnson, M. S., and Junker, B. W. (1999). "Classical Multilevel and

Bayesian Approaches to Population Size Estimation Using Multiple Lists." *Journal of the Royal Statistical Society. Series A (Statistics in Society)*, 162(3): 383–405. 662

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. 662

Hall, R., Steorts, R. C., and Fienberg, S. E. (2012). "Bayesian Parametric and Nonparametric Inference for Multiple Record Linkage." In *Modern Nonparametric Methods in Machine Learning Workshop*. Neural Information Processing Systems. 659, 660, 661, 662

Liseo, B. and Tancredi, A. (2011). "Bayesian Estimation of Population Size via Linkage of Multivariate Normal Data Sets." *Journal of Official Statistics*, 27(3): 491–505. 661

Madigan, D. and York, J. C. (1997). "Bayesian Methods for Estimation of the Size of a Closed Population." *Biometrika*, 1(84): 19–31. MR1450189. doi: https://doi.org/10.1093/biomet/84.1.19. 662

Manrique-Vallier, D. (2016). "Bayesian Population Size Estimation Using Dirichlet Process Mixtures." *Biometrics*, 72(4): 1246–1254. MR3591609. doi: https://doi.org/10.1111/biom.12502. 662

Sadinle, M. (2014). "Detecting Duplicates in a Homicide Registry Using a Bayesian Partitioning Approach." *Annals of Applied Statistics*, 8(4): 2404–2434. MR3292503. doi: https://doi.org/10.1214/14-AOAS779. 662

Sadinle, M. (2017). "Bayesian Estimation of Bipartite Matchings for Record Linkage." *Journal of the American Statistical Association*, 112(518): 600–612. MR3671755. doi: https://doi.org/10.1080/01621459.2016.1148612. 662

Sadinle, M. (2018). "Bayesian Propagation of Record Linkage Uncertainty into Population Size Estimation of Human Rights Violations." *Annals of Applied Statistics*, 12(2): 1013–1038. MR3834293. doi: https://doi.org/10.1214/18-AOAS1178. 662

Steorts, R. C., Hall, R., and Fienberg, S. E. (2014). "SMERED: A Bayesian Approach to Graphical Record Linkage and De-Duplication." In *Proceedings of the 17th International Conference on Artificial Intelligence and Statistics (AISTATS)*, 922–930. MR3601725. doi: https://doi.org/10.1080/01621459.2015.1105807. 659, 660, 661, 662

Steorts, R. C., Hall, R., and Fienberg, S. E. (2016). "A Bayesian Approach to Graphical Record Linkage and Deduplication." *Journal of the American Statistical Association*, 111(516): 1660–1672. MR3601725. doi: https://doi.org/10.1080/01621459.2015.1105807. 659, 660, 661, 662

Tancredi, A. and Liseo, B. (2011). "A Hierarchical Bayesian Approach to Record Linkage and Population Size Problems." *Annals of Applied Statistics*, 5(2B): 1553–1585. MR2849786. doi: https://doi.org/10.1214/10-AOAS447. 659, 661

Invited Discussion

Jared S. Murray[*,†]

I would like to congratulate the authors on a stimulating contribution to the literature on record linkage/de-duplication and population size estimation. Tancredi and Liseo (2011) was one of the papers that first piqued my interest in record linkage, so I am pleased to see more work along these lines (with an author population size of N+1!) My discussion below focuses on two main themes: Providing a more nuanced picture of the costs and benefits of joint models for record linkage and the "downstream task" (i.e. whatever we might want to do with the linked and de-duplicated files), and how we should measure performance.

1 The promise and peril of joint modeling: A partial defense of disunity

The promise of a joint model for record linkage, de-duplication, and population size estimation is likely obvious to the readership of Bayesian Analysis: We immediately obtain valid posterior inference over the population size that accounts for uncertainty about duplicates and links across files – provided that we specify an adequate joint model. Which leads us predictably to the peril of joint modeling, the fact that specifying a model for any of these three tasks alone is nontrivial. Addressing them simultaneously in a single model requires specifying a joint model sufficiently rich to do well on all three tasks (linkage, de-duplication, and population size estimation) while being tractable enough to understand its properties and perform posterior inference.

The model presented here necessarily makes some compromises in service of joint modeling, and I wonder about their impact. For example, assumptions about the sampling process generating the lists are essential to modeling the unknown population size and therefore must appear in any unified model. This will consequently restrict the prior distribution over the overlap between files in the record linkage/de-duplication portion of the model, despite the fact that the assumption of simple random sampling from the population – or any sort of random sampling at all – is otherwise irrelevant to record linkage and de-duplication. The assumptions made by the authors imply a very particular, informative prior distribution on Z, the partition of records into co-referent sets, and therefore on K, the number of distinct units captured across all lists (as reported in Table 1).

This choice is consequential. Indeed, immediately prior to Section 3.1 the authors note that the induced prior distribution on K is probably *not* well-suited to record

[*]Department of Information, Risk, and Operations Management and Department of Statistical Science. University of Texas at Austin, jared.s.murray@mccombs.utexas.edu

[†]The author gratefully acknowledges support from the National Science Foundation under grant number SES-1824555. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the funding agencies.

linkage tasks in general, which makes me wonder why we should expect it to work well when doing record linkage and population size estimation simultaneously. I have to assume that either 1) we actually don't expect it to work particularly well but the joint model at hand demands it or 2) the assumptions about the sampling process are actually warranted here, at least approximately, while they may not be in general applications of record linkage. If the former, this seems to beg the question and ignore options beyond joint modeling. If the latter, things are more interesting.

If the assumptions are in fact correct, we would expect to obtain more accurate and efficient inferences by inducing the "true" prior over Z and K using the joint model. But what happens when the sampling assumptions are violated? It is difficult to say, and it must depend on a host of factors (such as the degree and frequency of errors among co-referent records). However, it is not hard to imagine a case where relatively minor deviations from the sampling assumptions are more or less innocuous in the context of a population size model with known partition Z but become influential when Z is unknown and jointly modeled, due to the influence of the "misspecified" informative prior over Z. It would be interesting to try and draw this out via a simulation exercise (particularly in light of how influential Steorts et al. (2016) found a similar prior to be in a solely record linkage/de-duplication context).

If posterior inference is not robust to deviations from the sampling assumptions, what could we do instead? The desire to mitigate this undesirable "feedback" from a misspecified sub-model appears in many different settings, from Bayesian causal inference with propensity score models (McCandless et al., 2010; Zigler et al., 2013) to astrophysics (Yu et al., 2018) and beyond (see Jacob et al. (2017) for additional examples). This is a difficult problem and an active area of research. The proposed solutions often take the form of (possibly incoherent) multistage inference, in this case inferring the linkage structure in stage 1 and the population size in stage 2, propagating uncertainty from stage 1 to stage 2 without allowing any information from stage 2 to flow to stage 1. Jacob et al. (2017) give examples of settings where these "posteriors" are better than the posterior under a misspecified joint model in a decision-theoretic sense.

In the context of de-duplication and population size modeling, Sadinle (2018) proposes a related two-stage alternative to joint modeling termed "linkage averaging". If (in the notation of the current paper) $h(\lambda)$ is the estimate of population size we would compute given complete data (i.e., a de-duplicated and linked set of files) then under certain conditions the posterior for $h(\lambda)$ under a record linkage/de-duplication model alone will give the same inferences as a proper Bayesian joint model for linkage, de-duplication, and population size estimation. With a single set of posterior samples one can perform inference over multiple models for the population size, again provided that they all satisfy some relatively mild conditions.

These conditions do necessarily demand a degree of compatibility between the prior on λ and the population size model. They bear a striking similarity to the conditions under which multiple imputation delivers (asymptotically) valid Bayesian inference ("congeniality", (Meng, 1994; Xie and Meng, 2017; Murray, 2018)). This raises the interesting question of whether the compatibility conditions might be relaxed while still yielding conservative inferences, similar to the way one can obtain conservative inferences using

imputations under an uncongenial imputation model, provided it is uncongenial in the "right" way (roughly, by making fewer assumptions during imputation than analysis).

2 Measuring and improving performance

Various sub-specialties of statistics have spawned their own de-facto benchmark datasets – think of the iris data for clustering or the galaxy dataset for density estimation. Likewise, `RLdata500` and `RLdata10000` have arguably become something of a benchmark in record linkage problems due in large part to their accessibility via the popular `RecordLinkage` R package. I have used them in publications myself (Murray, 2015). Benchmark datasets form a sort of lingua franca that is useful for teaching, exposition, and as a sort of sanity check (when our brilliant new method finds six distinct clusters in the iris data, it's back to the drawing board).

However, we have to be careful extrapolating from these datasets to more complex settings. In the provocatively titled "Leave the Pima Indians Alone", Chopin and Ridgway (2017) make the case that an excessive focus on relatively simple binary regression problems like the Pima Indians diabetes dataset has had a distortive impact on the Bayesian computation literature. I worry a little that repeatedly going back to the `RLdata` datasets might lead the record linkage literature up the same path. In particular, the errors in these synthetic datasets are rather minimal, and the duplicate record pairs are quite well-separated from the non-duplicates. In my experience this not representative of the datasets we see in the wild, at least not those that demand sophisticated statistical modeling. Like Britney and the Pima Indians, I think it may be time to leave `RLdata` alone.

However, the primary evidence that the authors provide in favor of their model is its performance on `RLdata` datasets. Even setting aside whether this is a representative testbed, I wonder if this is much evidence at all since no alternative approaches are presented. Several are available, at least for the record linkage and de-duplication tasks, including some developed by the authors themselves (e.g. Steorts et al. (2015) reports false negative and false discovery rates of 0.02 and 0.04 on `RLdata500`, versus 0.015 and 0.08 using the model in the current paper). How well do existing Bayesian models perform on the linkage/de-duplication task? What about even simpler methods, like the point estimates generated by Fellegi-Sunter methods (Fellegi and Sunter, 1969) or their generalizations (Sadinle and Fienberg, 2013; Murray, 2015)? This is important context; while the model proposed here offers richer inference, should we trust those inferences if the model does not perform relatively well on the linkage/de-duplication task?

The authors actually seem to go a step further and use results on `RLdata` to inform parameter selection when modeling the Syrian casualty data. This frankly seems like a bad idea; in my own experience with similar files (Dalmasso et al., 2019), including expert hand-linked datasets, we observed very different patterns of distortion among coreferent records than the simple patterns one would find in `RLdata`. Given how variable performance is across parameter settings in Section 4, I would suggest that at least some sensitivity analysis might be in order for the Syria application.

Rather than rely on unrepresentative benchmark datasets to measure performance and select parameters, what could we do instead? The longer I work on record linkage problems the more I am convinced of the need to include a hand-labeling exercise alongside every serious application. The synthetic datasets at our disposal are limited in the range of errors they include and are often poor representations of the problem at hand. Model-based estimates of error rates are only as good as the model, and if we're not sure about the model... However, provided that the true error rates are low, precise estimates of false match rates (false discovery rates) can be obtained via random sampling from matched record pairs. False match rates aren't everything, but they aren't nothing either. Sadly the authors missed an opportunity to do even a little inspection here; after finding a small number of duplicates in the Syria application, they note only that "visual inspection of these pairs may eventually confirm their matching status".

Ideally a labeling exercise to evaluate a record linkage/de-duplication method should include matches generated by other methods (to remove potential bias toward declaring estimated matches correct), blinding (to the method(s) that declared the link), multiple review, an "indeterminate" or "unsure" option for the labelers, and should present labelers with neighboring "near-match" record pairs. Stellar examples of hand-labeling study designs include Bailey et al. (2017); Frisoli and Nugent (2018). In McVeigh et al. (2019) we hand-labeled a relatively small number of links to compare two competing methods, including one Bayesian model. For the Bayesian model we also used these labels to obtain the posterior distribution of false match rate adjusted estimators by computing them on each posterior sample of the linkage structure (similar to Sadinle (2018)'s linkage averaging). For our estimands, we only found it necessary to adjust for the false match rate and we did not grapple with simultaneous de-duplication or multiple files. But we did find that variation due to assumptions about bias from linkage error tended to swamp variation due to uncertainty about the linkage structure.

Reducing or otherwise accounting for linkage error seems important in the context of the current paper as well. Observe that in Figure 3, the estimates of K are worse in the blocks with higher error rates (blocks 7, 1, 10, 3) and in each case the estimate for K is biased down with a rather concentrated posterior distribution. If the model cannot be improved further, perhaps we would be better off looking at the posterior distribution of linkage error adjusted estimates of the population size. Linkage error adjusted estimators for the population size do exist, at least for relatively simple settings (e.g. Ding and Fienberg (1994); Di Consiglio and Tuoto (2018); Heijden (2019)) and perhaps could be cast in Sadinle (2018)'s framework of linkage averaging (although I have not checked the compatibility conditions myself). These estimators depend on false non-match rates, which are more difficult to obtain through hand labeling but often can be reasoned about based on plausible levels of duplication and overlap. This reasoning could form the basis of a computationally efficient sensitivity analysis. This seems like a promising avenue for future research, alongside further improvements in model and prior specification to minimize error rates.

References

Bailey, M., Cole, C., Henderson, M., and Massey, C. (2017). "How well do automated methods perform in historical samples? Evidence from new ground truth." Technical report, National Bureau of Economic Research, Inc. 667

Chopin, N. and Ridgway, J. (2017). "Leave Pima Indians alone: binary regression as a benchmark for Bayesian computation." *Statistical Science*, 32(1): 64–87. MR3634307. doi: https://doi.org/10.1214/16-STS581. 666

Dalmasso, N., Mejia, R., Rodu, J., Price, M., and Murray, J. (2019). "Feature Engineering for Entity Resolution with Arabic Names: Improving Estimates of Observed Casualties in the Syrian Civil War." *Artificial Intelligence for Humanitarian Assistance and Disaster Response Workshop Workshop, NIPS*. 666

Di Consiglio, L. and Tuoto, T. (2018). "Population size estimation and linkage errors: The multiple lists case." *Journal of Official Statistics*, 34(4): 889–908. 667

Ding, Y. and Fienberg, S. (1994). "Dual system estimation of Census undercount in the presence of matching error." *Survey Methodology*, 20(2): 149–158. 667

Fellegi, I. P. and Sunter, A. B. (1969). "A Theory for Record Linkage." *Journal of the American Statistical Association*, 64(328): 1183–1210. URL http://www.tandfonline.com/doi/abs/10.1080/01621459.1969.10501049. 666

Frisoli, K. and Nugent, R. (2018). "Exploring the Effect of Household Structure in Historical Record Linkage of Early 1900s Ireland Census Records." In *2018 IEEE International Conference on Data Mining Workshops (ICDMW)*, 502–509. 667

Heijden, P. V. D. (2019). "A linkage error correction model for population size estimation with multiple sources." In *62nd World Statistics Congress (19/08/19–23/08/19)*. URL https://eprints.soton.ac.uk/436665/. 667

Jacob, P. E., Murray, L. M., Holmes, C. C., and Robert, C. P. (2017). "Better together? Statistical learning in models made of modules." *arXiv preprint* arXiv:1708.08719. 665

McCandless, L. C., Douglas, I. J., Evans, S. J., and Smeeth, L. (2010). "Cutting feedback in Bayesian regression adjustment for the propensity score." *The international journal of biostatistics*, 6(2). MR2602559. doi: https://doi.org/10.2202/1557-4679.1205. 665

McVeigh, B. S., Spahn, B. T., and Murray, J. S. (2019). "Scaling Bayesian Probabilistic Record Linkage with Post-Hoc Blocking: An Application to the California Great Registers." URL https://arxiv.org/abs/1905.05337 667

Meng, X.-L. (1994). "Multiple-imputation inferences with uncongenial sources of input." *Statistical Science*, 538–558. 665

Murray, J. S. (2015). "Probabilistic Record Linkage and Deduplication after Indexing, Blocking, and Filtering." *Journal of Privacy and Confidentiality*, 7(1). URL https://journalprivacyconfidentiality.org/index.php/jpc/article/view/643 666

Murray, J. S. (2018). "Multiple imputation: A review of practical and theoretical findings." *Statistical Science*, 33(2): 142–159. MR3797707. doi: https://doi.org/10.1214/18-STS644. 665

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. 665, 667

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. 666

Steorts, R. C., Hall, R., and Fienberg, S. E. (2016). "A Bayesian approach to graphical record linkage and deduplication." *Journal of the American Statistical Association*, 111(516): 1660–1672. MR3601725. doi: https://doi.org/10.1080/01621459.2015.1105807. 665

Steorts, R. C. et al. (2015). "Entity resolution with empirically motivated priors." *Bayesian Analysis*, 10(4): 849–875. MR3432242. doi: https://doi.org/10.1214/15-BA965SI. 666

Tancredi, A. and Liseo, B. (2011). "A hierarchical Bayesian approach to record linkage and population size problems." *The Annals of Applied Statistics*, 5(2B): 1553–1585. MR2849786. doi: https://doi.org/10.1214/10-AOAS447. 664

Xie, X. and Meng, X.-L. (2017). "Dissecting multiple imputation from a multi-phase inference perspective: what happens when God's, imputer's and analyst's models are uncongenial?" *Statistica Sinica*, 1485–1545. MR3701490. 665

Yu, X., Del Zanna, G., Stenning, D. C., Cisewski-Kehe, J., Kashyap, V. L., Stein, N., van Dyk, D. A., Warren, H. P., and Weber, M. A. (2018). "Incorporating Uncertainties in Atomic Data Into the Analysis of Solar and Stellar Observations: A Case Study in Fe XIII." *The Astrophysical Journal*, 866(2): 146. 665

Zigler, C. M., Watts, K., Yeh, R. W., Wang, Y., Coull, B. A., and Dominici, F. (2013). "Model feedback in Bayesian propensity score estimation." *Biometrics*, 69(1): 263–273. MR3058073. doi: https://doi.org/10.1111/j.1541-0420.2012.01830.x. 665

Contributed Discussion

Nianqiao Ju[*,§], Niloy Biswas[*], Pierre E. Jacob [*], Gonzalo Mena [*,†], John O'Leary[*], and Emilia Pompe[‡]

We congratulate the authors on their important contribution to the record linkage literature, performing data de-duplication in the presence of population size uncertainty. The authors' method involves a mix of discrete parameters, such as a latent population size N and an unobserved partition over $\{1, \ldots, n\}$, and continuous ones like the vectors β_0, β' and θ. Estimating this model entails significant computational challenges, resembling those associated with mixture models and Bayesian non-parametric methods. These arise as much in the design of a sampling algorithm as in the assessment of its convergence.

We would like to highlight a few tools that can be used to build confidence in MCMC results under such conditions. Many of these involve Markov chain couplings, as in Johnson (1996, 1998), and more recently in Glynn and Rhee (2014); Nikooienejad et al. (2016); Jacob et al. (2020); Biswas et al. (2019). These methods allow for the assessment and removal of the impact of the starting distribution. They apply when it is possible to run multiple chains that evolve marginally according to the proposed algorithm and jointly so that they meet after a random number of iterations.

In the following, we describe how to generate chains $X_t^{(1)} = (\eta^{(1)}, \beta_0^{(1)}\beta'^{(1)}, \theta^{(1)}, N^{(1)})$ and $X_t^{(2)} = (\eta^{(2)}, \beta_0^{(2)}, \beta'^{(2)}, \theta^{(2)}, N^{(2)})$ which follow the Gibbs sampler of Section 5 of this article and which meet exactly at a random time. A basic strategy for coupling Gibbs methods involves coupling each conditional update. The full conditional distribution of label indicators η is a Multinomial distribution on $\{1, \ldots, n\}$ with the vector of probabilities computed as in (5.5) of the article. To couple these updates, we compute such probabilities for both chains and implement a maximal coupling to obtain two labels which will be identical with the maximal probability. For the continuous variables β_0, β' and θ, which are updated with Metropolis–Hasting steps, we can employ maximal couplings of the Normal or Dirichlet proposal distributions, and use common Uniform draws to accept or reject them. Finally we can update N with an exact Gibbs step, truncating N to a very large integer, and implement a maximal coupling of this step.

However, an interesting difficulty arises, reminiscent of the infamous label switching issue (Stephens, 2000). Suppose that a first chain has $\eta^{(1)} = (1, 4, 3, 4, 2)$ and the second $\eta^{(2)} = (4, 3, 2, 3, 1)$, in a simple example with $n = 5$. These labels correspond to the same partition, and yet the η-components of the chains are different and thus the chains cannot coincide. Judging from our toy experiments, the associated meeting time would be long. This can be alleviated by an additional relabeling step, to be performed after the update of the components of η. A simple strategy, for example, is to relabel

[*]Department of Statistics, Harvard University
[†]Harvard Data Science Initiative
[‡]Department of Statistics, University of Oxford
[§]nju@g.harvard.edu

η according to the order of the occurrences of new blocks, from component 1 to n. That is, both the labels $\eta^{(1)} = (1, 4, 3, 4, 2)$ and $\eta^{(2)} = (4, 3, 2, 3, 1)$ would be relabelled $(1, 2, 3, 2, 4)$. This relabeling needs to be accompanied by an adequate reshuffling of the associated parameters, namely the β'-components in the notation of Section 5. Other, more sophisticated relabeling strategies could be devised, perhaps inspired by the literature on label-switching issues in mixture models (Stephens, 2000; Marin et al., 2005; Frühwirth-Schnatter, 2011).

We will make some R scripts implementing a coupling of the proposed Gibbs sampler available at https://github.com/EmiliaPompe/discussion_unified_framework, along with some simple numerical experiments on the synthetic dataset `RLdata500` from the `R` package `RecordLinkage` analysed in Section 6 of the paper.

References

Biswas, N., Jacob, P. E., and Vanetti, P. (2019). "Estimating convergence of Markov chains with L-lag couplings." In *Advances in Neural Information Processing Systems*, 7389–7399. 670

Frühwirth-Schnatter, S. (2011). "Label switching under model uncertainty." *Mixtures: Estimation and Application*, 213–239. MR2883354. doi: https://doi.org/10.1002/9781119995678.ch10. 671

Glynn, P. W. and Rhee, C.-h. (2014). "Exact estimation for Markov chain equilibrium expectations." *Journal of Applied Probability*, 51(A): 377–389. MR3317370. doi: https://doi.org/10.1239/jap/1417528487. 670

Jacob, P. E., O'Leary, J., and Atchadé, Y. F. (2020). "Unbiased Markov chain Monte Carlo with couplings." *Journal of the Royal Statistical Society: Series B (Statistical Methodology) (with discussion) (to appear)*. MR3949304. doi: https://doi.org/10.1093/biomet/asy074. 670

Johnson, V. E. (1996). "Studying Convergence of Markov Chain Monte Carlo Algorithms Using Coupled Sample Paths." *Journal of the American Statistical Association*, 91(433): 154–166. MR1394069. doi: https://doi.org/10.2307/2291391. 670

Johnson, V. E. (1998). "A coupling-regeneration scheme for diagnosing convergence in Markov chain Monte Carlo algorithms." *Journal of the American Statistical Association*, 93(441): 238–248. MR1614640. doi: https://doi.org/10.2307/2669620. 670

Marin, J.-M., Mengersen, K., and Robert, C. P. (2005). "Bayesian modelling and inference on mixtures of distributions." *Handbook of statistics*, 25: 459–507. MR2490536. doi: https://doi.org/10.1016/S0169-7161(05)25016-2. 671

Nikooienejad, A., Wang, W., and Johnson, V. E. (2016). "Bayesian variable selection for binary outcomes in high-dimensional genomic studies using non-local priors." *Bioinformatics*, 32(9): 1338–1345.
URL https://doi.org/10.1093/bioinformatics/btv764 670

Stephens, M. (2000). "Dealing with label switching in mixture models." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 62(4): 795–809. MR1796293. doi: https://doi.org/10.1111/1467-9868.00265. 670, 671

Contributed Discussion

Christian P. Robert*

Congratulations to the authors, for this paper that expand the modelling of populations investigated by faulty surveys, a poor quality feature that applies to extreme cases like Syria casualties. And possibly COVID-19 victims.

The model considered in this paper, as given by (2.1), is a latent variable model which appears as hyper-parameterised in the sense it involves a large number of parameters and latent variables. First, this means it is essentially intractable outside a Bayesian resolution. Second, within the Bayesian perspective, it calls for identifiability and consistency questions, namely which fraction of the unknown entities is identifiable and which fraction can be consistently estimated, eventually severing the dependence on the prior modelling. Personal experiences with capture-recapture models on social data like drug addict populations showed me that prior choices often significantly drive posterior inference on the population size. Here, it seems that the generative distortion mechanism between registry of individuals and actual records is paramount.

> We now investigate an alternative aspect of the uniform prior distribution of λ given N.

Since the practical application stressed in the title, namely some of civil casualties in Syria, interrogations take a more topical flavour as one wonders at the connection between the model and the actual data, between the prior modelling and the available prior information. It is however not the strategy adopted in the paper, which instead proposes a generic prior modelling that could be deemed to be non-informative. I find the property that conditioning on the list sizes eliminates the capture probabilities and the duplication rates quite amazing, reminding me indeed of similar properties for conjugate mixtures, although we found the property hard to exploit from a computational viewpoint. And that the hit-miss model provides computationally tractable marginal distributions for the cluster observations.

> 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.

This non-informative choice is however quite informative in the misreporting mechanism and does not address the issue that it presumably is misspecified. It indeed makes the assumption that individual label and type of record are jointly enough to explain the probability of misreporting the exact record. In practical cases, it seems more realistic that the probability to appear in a list depends on the characteristics of an individual, hence far from being uniform as well as independent from one list to the next. The same applies to the probability of being misreported. The alternative to the uniform allocation of individuals to lists found in (3.3) remains neutral to the reasons why (some)

*CEREMADE, Université Paris Dauphine, University of Warwick, and CREST, GENES, xian@ceremade.dauphine.fr

individuals are missing from (some) lists. No informative input is indeed made here on how duplicates could appear or on how errors are made in registering individuals. Furthermore, given the high variability observed in inferring the number of actual deaths covered by the collection of the two lists, it would have been of interest to include a model comparison assessment, especially when contemplating the clash between the four posteriors in Figure 4.

The implementation of a manageable Gibbs sampler in such a convoluted model is quite impressive and one would welcome further comments from the authors on its convergence properties, since it is facing a large dimensional space. Are there theoretical or numerical irreducibility issues for instance, created by the discrete nature of some latent variables as in mixture models?

Rejoinder

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

Introduction

We thank the editorial team for organizing this discussion; and we thank all the discussants for providing stimulating and thought-provoking comments, which we have summarized into the following four categories:

- prior modeling and identifiability
- joint modeling and inference with linked data;
- computational issues; and
- model and performance evaluation.[1]

Before turning to these four points, we first review our proposed model and some simplifying assumptions, without loss of generality. We assume one single list and assume that the error probabilities depend only on the key variables (and not on the population entities), i.e. $\alpha_{j'\ell} = \alpha_\ell$. In addition, we assume the probabilities θ_ℓ are known. Then conditionally on the true values $\tilde{v}_{j'}$, for $j' = 1, \dots N$, the statistical model on the observed values is

$$p(v_i) = \sum_{j'=1}^{N} \frac{1}{N} k(v_i; \tilde{v}_{j'}) \tag{1}$$

independently for $i = 1, \dots, n$, where

$$k(v_i; \tilde{v}_{j'}) = \prod_{\ell=1}^{p} k_\ell(v_{i\ell}; \tilde{v}_{j'\ell}) = \prod_{\ell=1}^{p} \left[(1-\alpha_\ell)\delta(v_{i\ell}; \tilde{v}_{i\ell}) + \alpha_\ell \theta_{v_{i\ell},\ell}\right],$$

which emphasizes the analogies with a mixture model as noted by the discussants.

Prior Modeling and Identifiability

In this section, we comment on the choice of the prior on the linkage structure as well as model identifiability. **Professor Murray** states

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

[†]Department of Statistical Science and Computer Science, Duke University, Durham, North Carolina, beka@stat.duke.edu

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

[1]There may be connections within each of the four sections given the subject matter, which become apparent in our rejoinder.

> "... This choice is consequential. Indeed, immediately prior to Section 3.1 the authors note that the induced prior distribution on K is probably not well-suited to record linkage tasks in general, which makes me wonder why we should expect it to work well when doing record linkage and population size estimation simultaneously."

We thank **Professor Murray** for this observation as we believe that the uniform prior on the linkage structure (or label space) may not be adequate in some record linkage problems where N is fixed and given, since prior uncertainty may be too low. Nevertheless, it has been shown to work quite well in a number of situations, where N is not known, namely, in the work of Marchant et al. (2019) on an application to the United States Census Bureau.

On the other hand, we are confident that introducing an extra prior on N makes the model more flexible and allows one to state uncertainty in a more intuitive way. For instance, Table 1 illustrates that a uniform prior on the linkage structure (record labels) Λ induces a prior that is too concentrated on $K \mid N$. However, by allowing N to vary, as in our paper, the resulting prior on the partition Z is dramatically more diffuse and it can be considered as a weakly informative prior, which is more sensible than alternative proposals. As pointed out by **Professor Sadinle**, N was used as a hyper-parameter in Steorts et al. (2016), where the authors explored the prior sensitivity issue with respect to N and showed that the choice of N was particularly influential. In fact, one motivation of our paper was to account for the uncertainty of this quantity in the de-duplication process. Moreover, according to our model $1/N$ is the prior probability that two records are co-referent hence by fixing N to a given value we strongly control the de-duplication process. We stress that our goal is to propose a simple, yet more flexible prior on the linkage structure than has been proposed in the literature.

Under our proposed model there is a partition Z, which is induced by the mixture model (1) with an unknown N. We stress that we have not stated that the distribution of Z nor the expression after elimination of N represents the "correct" model in a real de-duplication application. The perils assumed by **Professor Murray** are shared by other proposed de-duplication models in the literature, which even separate the linkage step from the population size parameter N. For example, Sadinle (2014) has proposed to use a uniform prior on the partition space, after observing n records, $p(Z) = B_n^{-1}$ where B_n is the Bell number. Under this assumption, the corresponding distribution of distinct entities K is given by $p(K) = S(n,k)/B_n$ where $S(n,k)$ is the Stirling number of second kind. Using Pitman (2006), one obtains $E(K) = B_{n+1}/B_n - 1$ and $Var(K) = B_{n+2}/B_n - (B_{n+1}/B_n)^2 - 1$. Moreover, the prior probability that two records are co-referent can be shown to be B_{n-1}/B_n. In fact, if we turn to the recent literature on Bayesian record linkage that take a clustering approach, most assume a uniform prior on the linkage structure or co-reference matrix (Sadinle, 2014; Steorts, 2015; Steorts et al., 2016; Sadinle, 2017, 2018; Marchant et al., 2019; McVeigh et al., 2020). Therefore, we ask why these particular assumptions should be considered less dangerous than ours? The only two papers that take a completely non-parametric approach regarding the prior on the linkage structure are that of Zanella et al. (2016) and Betancourt et al. (2020), however, putting such priors into a joint modeling framework that scales would be extremely challenging.

Our approach, with an additional latent quantity, can be criticized in terms of identifiability as noticed by **Professor Robert**. In fact, the value of N has a direct impact on the variability of the data. For example, when $N = 1$, we observe n duplications of the same entity. For larger values of N, the observed variability of the data will increase. In a de-duplication model, observe that N plays a role similar to that of α. While N controls the variability of the *duplication between entities*, the vector α controls the variability for the *duplication within entities*. Can we safely estimate both the parameters without specific assumptions? We suspect that one cannot, and we agree with **Professor Robert** that this problem is most likely intractable outside of Bayesian entity resolution.

However, it is true that the analysis of the effects of what we assume on Z on the joint modeling is of paramount importance. In particular, we agree that there could be an undesirable feed-back effect on the downstream task if we take a badly calibrated prior on Z. On the other hand, there is no guarantee to avoid this problem if we separate the record linkage step from the downstream one. The prior structure on Z would remain the same even when we do not need to estimate N. Furthermore, it remains the same in more complex situations based on de-duplicated data. This point is related to the second classes of issues.

Joint Modeling and Inference with Linked Data

In this section, at the request of **Professor Murray**, we discuss joint modeling of the record linkage problem and the post-linkage inference with respect to possible consequence due to the lack of robustness with respect to the sampling assumptions.

In a recent companion paper (Steorts et al., 2018), we have investigated how to make inference on the regression coefficients of a linear relation estimated with a response Y and covariates $(X_1, \ldots, X_p)$ coming from different files and calibrated on linked records. We have shown how the linkage uncertainty can be introduced in the inferential process. On the other hand, we have investigated how the information coming from the regression analysis can modify our posterior on the cluster structure. This feed-back effect may or may not be welcome, according to the application. Moreover, when the regression model is not a good choice, the feed-back effect may increase the noise and worsen the linkage process.

The goal of our paper, here, is, admittedly different. We argue that, in the presence of a record linkage problem, our proposed model is able to provide valid statistical information on N. That is, we are able to provide an integrated likelihood function for N, which combined with a suitable prior, $\pi(N)$ will produce a posterior which summarizes what the data can tell about the population size, without explicitly designing a statistical model for N. Taking this argument further, one could use our posterior on N as the prior for a better designed capture-recapture experiment!

This leads into a point raised **Professor Sadinle**, who noted:

> ... Unfortunately, the capture-recapture component of the approach proposed by the authors does not correspond to any commonly used model in this area, and its assumptions are too stringent to

be practically useful, as acknowledged by the authors.

In this respect, we observe that when conditioning on the list sizes, we reduce to a conditional likelihood framework, which is similar to the traditional model of Darroch (1958). Note that fixing the sample size eliminates the capture probabilities, which leaves as the unique parameter, the population size N. In addition, by conditioning on the absence of within-list duplications, the lists become independent simple random sample without replacement. Therefore, as hoped by **Professor Sadinle**, the number of matches between two lists follows a hypergeometric distribution as in Tancredi and Liseo (2011). Finally, the request of a standard capture-recapture model in a duplication framework with two lists could be accomplished by assuming a hypergeometric distribution for the number of clusters belonging to both the lists, conditioned on the number of clusters for each list and then modeling the number of clusters within each list with a distribution that does not depend on N. We did not explore this possibility, however, this could be of exploration in future work.

Computational Considerations

In this section, we speak to comments regarding computational considerations.

We agree with **Professor Robert** that the proposed model is hyper-parameterized. We also agree with comments and suggestion made by **Ju and Collegagues**. Both situations require care, and are typical of high-dimensional (in the parameter space) mixture models. After receiving the written discussions, we have tried to transform this potential criticism into an effort towards a simpler strategy. If we assume the simplified model (1), where the error probability $\alpha_{j'\ell} = \alpha \ \forall j'$ and $\forall \ell$ and the probabilities vector θ_ℓ are assumed known, we obtain a simple two-parameter problem (α and N, or a more suitable their re-parametrization, such as $\psi(\alpha, N) = E(K|N)$). This can be written as

$$p(v_1, \ldots, v_n|\alpha, N) = \sum_{Z \in \mathcal{Z}} p(v_1 \ldots, v_n|Z)p(Z|N). \tag{2}$$

However the evaluation of the likelihood in equation (2) is still prohibitive, even for small values of the total number of records n. In fact, it would require one to sum B_n terms! On the other hand, if we consider the pairwise likelihood (Varin et al., 2011), associated to equation (2), the computational burden is dramatically reduced to

$$L_{pw}(\alpha, N) \quad = \quad \sum_{r=1}^{n-1} \sum_{s=r+1}^{n} p(v_r, v_s|\alpha, N),$$

where

$$p(v_r, v_s|\alpha, N) = \frac{1}{N} \prod_{\ell} \left(\delta(v_{r\ell}, v_{s\ell})\theta_{v_{r\ell}}(1-\alpha)^2 + (2\alpha - \alpha^2)\theta_{v_{r\ell}}\theta_{v_{s\ell}}\right) + \left(1 - \frac{1}{N}\right) \prod_{\ell} \theta_{v_{r\ell}}\theta_{v_{s\ell}}.$$

Making this over-simplification, we find that each pair (v_r, v_s) can be ascribed to the same cluster with probability $1/N$ and, conditional on being on the same cluster, their joint probability is (as reported in the paper) $\delta(v_{r\ell}, v_{s\ell})\theta_{v_{r\ell}}(1-\alpha)^2 + (2\alpha - \alpha^2)\theta_{v_{r\ell}}\theta_{v_{s\ell}}$, while conditionally on being on different clusters, v_r and v_s are simply independent. In future work, we plan to explore this alternative version in order to quantify the loss in information compared to the use of (2).

We appreciate that **Ju and Colleagues** have highlighted up the infamous label-switching issue, which is prevalent in Bayesian latent variable model representations of record linkage as pointed out by (Steorts et al., 2016). They and others have dealt with this particular issue by using point estimates of the linkage structure or coreference matrix. This approach has been utilized by Steorts (2015) and Marchant et al. (2019) and similar approaches were adopted and extended by Sadinle (2014).

Convergence issues proved to be problematic in our experimental studies with the `RLdata500` data set, which we believe is due to the fact that it has a small number of duplicate records. Thus, it seems to be a nice candidate data set regarding understanding the mixing (or failure to mix) of samplers. For our proposed method, we found that some combinations of the hyper-parameters used in simulations (see Table 3 of our paper) only reached convergence after more than 10,000 iterations of the sampler. This is consistent with the literature on this particular data set, where in the previous literature, one cannot achieve convergence of the `RLdata500` data set after 11 hours due the use of using a standard Gibbs sampler (Marchant et al., 2019). Thus, it seems that one opportunity to overcome the convergence issues would be the proposed approach of **Ju and Colleagues**, which could prove to be a valuable tool in record linkage moving forward.

Along with the interesting and promising results of **Ju and Colleagues**, recent computational advancements have been obtained by Marchant et al. (2019) using a partially collapsed Gibbs sampler within the framework of Steorts (2015); Steorts et al. (2016). Both directions would be interesting explorations for future endeavors.

Model and Performance Evaluation

In this section, we discuss model performance and evaluation.

We understand the concern regarding **Professor Murray** to leave the RLdata data behind, however, there is merit in considering such datasets, especially because of the difficulty finding valuable and other alternatives. We believe that the importance to share real data sets is crucial for the record linkage community. In fact, we were delighted to see that the quite small but *real* data exemplification of matching a census Italian block with the subsequent post enumeration survey proposed in Tancredi and Liseo (2011) has been used as a benchmark in both McVeigh et al. (2020) and Marchant et al. (2019). The Syrian casualties dataset offered the rare opportunity to use a partially public source and we embarked in this analysis not withstanding all the difficulties of what would require an extensive case study. In the future, we hope that other researchers contribute to the diffusion of their real data sets or at least part of them as has been done in the open source work of Marchant et al. (2019) regarding synthetic and real

data sets, which can be found at https://github.com/cleanzr/dblink-experiments and https://github.com/cleanzr/RLdata.

Our aim in our proposed work was to demonstrate the applicability of the proposed methodology even in an *observational data* framework, it was not our goal to propose an applied paper. This cannot impede us to completely agree with all the suggestions made by **Professor Robert** about the need of heterogeneous or "covariate depending" capture probabilities, duplication, and misreporting rates. This would be a different paper altogether.

We agree with **Professor Murray** that a hand-labeling exercise might increase the quality of the matching process. We observe that in the original proposal of the hit-miss model, Copas and Hilton (1990) required that the error probabilities α were estimated on a set of true matches and that those estimates were subsequently used to infer the matching status of a new pair of records. With the complexity of data we use today, a hand-labeling exercise would surely contribute to improve the prior modeling of the α parameters so producing a more reliable inference. However, considering the specific case of the Syrian data we note that for the full data set, containing all the data sources from the Syrian conflict, there has already been a very intense, rich hand-matched analysis as described in Price et al. (2013). Unfortunately, these labels do not easily map to the publicly available data set, and the private data set is not easily shareable with others given the sensitivity of the conflict. In our opinion, does not make sense to muddy the waters regarding the outstanding work that the Human Rights Data Analysis Group (HRDAG) has done on the front regarding hand-matching for many applications in human rights, where they have had many experts to analyze the Syrian data set and resolve it. In our opinion, this is a project in its own right given the amount of biases that may result from performing hand-matching. In our experience, when this has been done in previous studies by experts or used in case studies in the literature as in applications to the human rights conflicts (Sadinle, 2014; Price et al., 2013; Chen and Shrivastava, 2018), both the proposed models and handmatching is extremely difficult.

We stress that producing hand-matches labels is difficult by experts. This is even harder by those that are not well versed or trained in producing hand-matched data sets. This begs the question regarding what can go wrong with producing a hand-matched data set, and why do we recommend caution? For example, a hand-matched data set can be constructed such that it favors the proposed model, and disfavors other models. Thus, any hand-matched data set can be constructed to bias any model and teasing out such biases is difficult, especially a reviewer. In addition, there are many real applications where hand-matching is difficult to resolve in the sense that there are many ambiguous matches. We often see this in the case of human rights applications. In this sense, would it not be better to let the models be fully unsupervised to inform inference in the spirit of being fully Bayesian?

Finally, turning to comparison with other methods, as suggested by **Professor Murray**. We would like to stress that the suggested methods in his discussion differ completely in scope and are not comparable as they require different assumptions. For example, the work of Steorts (2015) assumes both textual data and categorical data, whereas we only consider categorical data. Furthermore, other models suggested by the

discussant, such as that of Sadinle (2014) would not be directly fair as textual data is assumed and variations of the Fellegi–Sunter approach in a Bayesian de-duplication framework require careful tuning. It is paramount to remember that a comparison with other methods that focus on variations or completely different methods altogether may lead to unfair comparisons. Rather than focusing on this, we emphasize the importance of the sensitivity analysis with respect to the choice of the hyper-parameters as reported in Table 3 of our paper. In fact, this kind of comparisons could help in having a more general insight on the kind of information one needs to extract from the proposed model in order to improve the matching process.

References

Betancourt, B., Zanella, G., and Steorts, R. C. (2020). "Random Partition Models for Microclustering Tasks." *arXiv preprint* arXiv:2004.02008. 676

Chen, B. and Shrivastava, R. C., A.and Steorts (2018). "Flexible Models for Microclustering with Application to Entity Resolution." *The Annals of Applied Statistics*, 12(2): 1039–1067. MR3834294. doi: https://doi.org/10.1214/18-AOAS1163. 680

Copas, J. and Hilton, F. (1990). "Record linkage: statistical models for matching computer records." *Journal of the Royal Statistical Society, A*, 153: 287–320. 680

Darroch, J. (1958). "The multiple capture census I. Estimation of a closed population." *Biometrika*, 45: 343–358. MR0119360. doi: https://doi.org/10.2307/2333183. 678

Marchant, N., Steorts, R., Kaplan, A., Rubinstein, B., and Elazar, D. (2019). "Distributed End-to-End Bayesian Entity Resolution." *arXiv pre-print* arXiv:1909.06039. 676, 679

McVeigh, B., Spahn, B., and Murray, J. (2020). "Scaling Bayesian Probabilistic Record Linkage with Post-Hoc Blocking: An application to the California Great Register." *Technical Report*. 676, 679

Pitman, J. (2006). *Combinatiorial Stochastic Processes*. Ecole d'Eté de Probabilités de Saint-Flour XXXII, Lecture Notes in Mathematics, vol. 1875, Berlin, Springer. MR2245368. 676

Price, M., Kingner, J., and Ball, P. (2013). "Preliminary statistical analysis of documentation of killings in the Syrian Arab Republic." *Technical Report*. 680

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. 676, 679, 680, 681

Sadinle, M. (2017). "Bayesian Estimation of Bipartite Matchings for Record Linkage." *Journal of the American Statistical Association*, 112(518): 600–612. MR3671755. doi: https://doi.org/10.1080/01621459.2016.1148612. 676

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. 676

Steorts, R., Tancredi, A., and Liseo, B. (2018). "Generalized Bayesian record Linkage and Regression with Exact Error Propagation." In *Proceedings of the International Conference on Privacy in Statistical Databases, PSD2018*, 297–313. 677

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. 676, 679, 680

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. 676, 679

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. 678, 679

Varin, C., Reid, N., and Firth, D. (2011). "An overview of composite likelihood methods." *Statistica Sinica*, 21: 5–42. MR2796852. 678

Zanella, G., Betancourt, B., Wallach, H., Miller, J., Zaidi, A., and Steorts, R. C. (2016). "Flexible Models for Microclustering with Application to Entity Resolution." In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, NIPS'16, 1425–1433. NY, USA: Curran Associates Inc. 676