Bayesian Analysis (0000) **00**, Number 0, pp. 1

Discussion of "A Unified Framework for De-Duplication and Population Size Estimation" by Tancredi, Steorts, and Liseo

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 $j' = 1, \ldots, N$.

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

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

© 0000 International Society for Bayesian Analysis DOI: 0000

- The records:
  - 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. 4

Castledine, B. J. (1981). "A Bayesian Analysis of Multiple-Recapture Sampling for a Closed Population." *Biometrika*, 68(1): 197–210. 4

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. 2

Fienberg, S. E. (1972). "The Multiple Recapture Census for Closed Populations and Incomplete 2^k Contingency Tables." *Biometrika*, 59(3): 591–603. 4

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. 4

George, E. I. and Robert, C. P. (1992). "Capture-Recapture Estimation Via Gibbs Sampling." *Biometrika*, 79(4): 677–683. 4

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. 1, 2, 3, 4

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. 3

Madigan, D. and York, J. C. (1997). "Bayesian Methods for Estimation of the Size of a Closed Population." *Biometrika*, 1(84): 19–31. 4

Manrique-Vallier, D. (2016). "Bayesian Population Size Estimation Using Dirichlet Process Mixtures." *Biometrics*, 72(4): 1246–1254. 4

Sadinle, M. (2014). "Detecting Duplicates in a Homicide Registry Using a Bayesian Partitioning Approach." *Annals of Applied Statistics*, 8(4): 2404–2434. 4

— (2017). "Bayesian Estimation of Bipartite Matchings for Record Linkage." *Journal of the American Statistical Association*, 112(518): 600–612. 4

— (2018). "Bayesian propagation of record linkage uncertainty into population size estimation of human rights violations." *Annals of Applied Statistics*, 12(2): 1013–1038. 4

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. 1, 2, 3, 4

— (2016). "A Bayesian Approach to Graphical Record Linkage and Deduplication." *Journal of the American Statistical Association*, 111(516): 1660–1672. 1, 2, 3, 4

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. 1, 3