Deduplication and population size estimation

Andrea Tancredi

Sapienza University of Rome

joint work with R. Steorts and B. Liseo

ITACOSM, Florence
June 7, 2019

Syrian data

- Two data sets reporting different lists of victims killed in the recent Syrian conflict
- Data provided by the
  - Violations Documentation Center in Syria (VDC) ▸ Link
  - Syrian Center for Statistics and Research (CSR) ▸ Link
- Both the list comprise first name, family name, date of death and death location
- We want to estimate the total number of victims

Record linkage, duplications and population entities

Assume L lists and N latent individuals, $1 \leq N < \infty$

- N is unknown
- The L files share a set of p categorical key variables
- $v_{ij} = (v_{ij1}, \ldots, v_{ijp})$ denotes the record j in file i $(j = 1, \ldots, n_i)$ $n = n_1 + \cdots + n_L$
- $\tilde{v}_{j'} = (\tilde{v}_{j'1} \ldots \tilde{v}_{j'p})$ is the *true record* for the latent individual j', $j' = 1, \ldots N$
- $\lambda_{ij} \in \{1 \ldots, N\}$ denotes the latent individual generating v_{ij}

$$\begin{aligned} \lambda_{ij_1} = \lambda_{ij_2} \quad &\Rightarrow \quad \text{a duplication in the same list} \\ \lambda_{i_1 j_1} = \lambda_{i_2 j_2} \quad &\Rightarrow \quad \text{a match between two lists} \end{aligned}$$

- $\tilde{V}_{j'\ell} \overset{D}{=} V_\ell$ independently for $j' = 1, \ldots N$ and $\ell = 1 \ldots p$ where $V_\ell \sim \{v_{\ell\,1} \ldots, v_{\ell\,M_l}; \theta_{\ell\,1} \ldots, \theta_{\ell\,M_\ell}\}$ $\ell = 1, \ldots p$

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

- The hit-miss model (Copas and Hilton (1990) JRSSA)

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

is the *conditional* generating processes of the key variables:

- Conditional independence among all the observed records given their respective unobserved true records

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

Partitions, record linkage and population size

- We assume $p(\lambda|N) = \prod_{ij} p(\lambda_{ij}) = \prod_{ij} \frac{1}{N} = \left(\frac{1}{N}\right)^n$
  - L independent SRSWR **of size $\mathbf{n_j}$** from N labels
  - $p(\lambda|N)$ induces a prior on the partition space,
  - Let k be the number of blocks of the partition Z with labels U

    $$p(Z|N) = \left(\frac{1}{N}\right)^n \frac{N!}{(N-k)!} \quad p(U|Z,N) = \frac{(N-k)!}{N!}$$

    and

    $$p(k|N) = \frac{S(n,k)}{N^n} \frac{N!}{(N-k)!}$$

    where $S(n,k)$ is the Stirling number of second kind

- As a prior for N we take $p(N) \propto 1/N^g$

Suppose that each population unit j' undergoes $T_{jj'}$ capture attempts $X_{jj't}$ for $t = 1, \ldots T_{jj'}$ (for list j) where

- $T_{jj'} \overset{ind.}{\sim} \mathit{Poisson}(\delta_j)$ for $j' = 1, \ldots N$, $j = 1, \ldots, L$
- $X_{jj't} \overset{ind.}{\sim} \mathit{Bernoulli}(p_j)$ for $j' = 1, \ldots N$, $j = 1, \ldots, L$, $t = 1, \ldots, T_{jj'}$

Let $X_{jj'} = \sum_{t=1}^{T_{jj'}} X_{jj't}$ be the captures for unit j' in list j. Then $n_j = \sum_{j'=1}^{N} X_{jj'}$ and

$$p(\lambda_1, \ldots, \lambda_L | n_1, \ldots, n_L) = \prod_{j=1}^{L} \left(\frac{1}{N}\right)^{n_j} = \left(\frac{1}{N}\right)^{n}$$

where $\lambda_j = (\lambda_{1j}, \ldots, \lambda_{n_j j})$ is the label sequence for list j

Condtioning on the list sizes eliminates the capture probabilities and the duplication rates p_j, δ_j $j = 1, \ldots, L$

Hit-miss model and clustering

Let $z \in Z$ be a partition block, $v_z = (v_{ij} : ij \in z)$ be the cluster of records, u_z be the label in U of the block z and $\tilde{v}_U = (\tilde{v}_{u_z}, z \in Z)$ be the relative sets of population records

$$\begin{aligned} p(v|\tilde{v}, \lambda, N, \alpha, \theta) &= \prod_{z \in Z} p(v_z|\tilde{v}_{u_z}, Z, U, \alpha, \theta) \\ p(\tilde{v}_U|\lambda, N, \alpha, \theta) &= \prod_{z \in Z} p(\tilde{v}_{u_z}|Z, U, \theta) \end{aligned}$$

Integrating out the unknown population values $\tilde{v}$ we obtain

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

$p(v_z|Z, U, \alpha, \theta)$ can be obtained analytically for small clusters (size ≤ 3) or via a recursive formula.

Computation and other prior assumptions

- Independent hierarchical normal priors for the logit trasformations β_ℓ of the measurement error parameters $\alpha_\ell = (\alpha_{\ell 1} \dots \alpha_{\ell N})$ for $\ell = 1, \dots p$
- Dirichlet priors for $\theta_\ell = (\theta_{\ell\, 1} \dots \theta_{\ell\, M_\ell})$ for $\ell = 1, \dots, p$
- We simulate the posterior

$$p(\lambda, N, \beta, \theta | v) \propto p(v | Z, U, \beta, \theta) p(U | Z, N) p(Z | N) p(\beta) p(\theta)$$

via a Metropolis within Gibbs algorithm

RL500 data set

- $n = 500$ synthetic records: complete names and date of birth
- 50 records duplicated and distorted. Single list with $k = 450$ entities.

	fname_c1	fname_c2	lname_c1	lname_c2	by	bm	bd
1	CARSTEN		MEIER		1949	7	22
2	GERD		BAUER		1968	7	27
3	ROBERT		HARTMANN		1930	4	30
4	STEFAN		WOLFF		1957	9	2
5	RALF		KRUEGER		1966	1	13
⋮							
43	GERD		BAUERH		1968	7	27
⋮							

- To apply our model we transform name and surname via the SOUNDEX algorithm and we divide the year of birth in 4 fields.

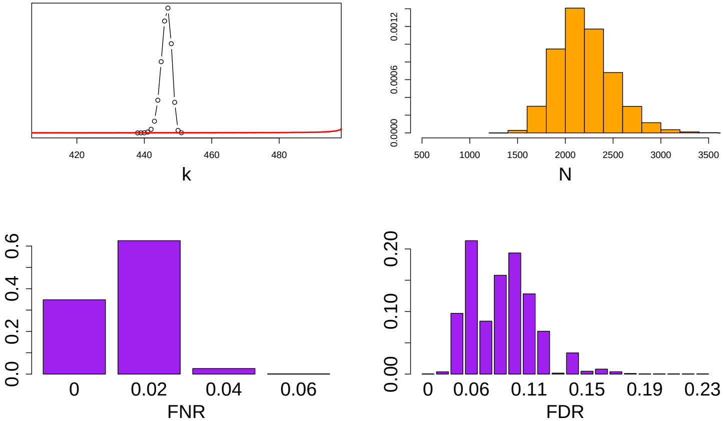
	name fields				*surname fields*				*day of birth fields*					
									year				*month*	*day*
1	C	6	2	3	M	6	0	0	1	9	4	9	7	22
2	G	6	3	0	B	6	0	0	1	9	6	8	7	27
3	R	1	6	3	H	6	3	5	1	9	3	0	4	30
4	S	3	1	5	W	4	1	0	1	9	5	7	9	2
5	R	4	1	0	K	6	2	6	1	9	6	6	1	13
6	J	6	2	5	F	6	5	2	1	9	2	9	7	4

- $p(N) \propto 1/N^g$ with g=1.02. θ_ℓ are uniform on the simplex.

  Prior mean and 0.99 quantile for $\alpha_{j'l}$ approx. equal to 0.01 and 0.06. → strong believe towards low block distortion probabilities.

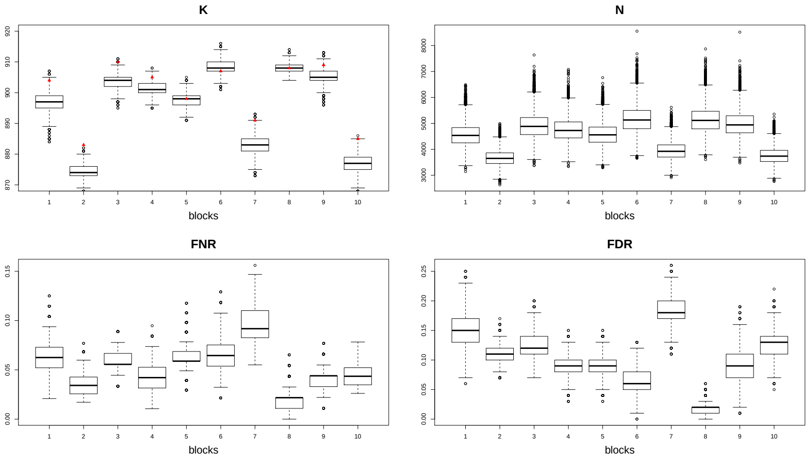
- Prior for **microclusters**: larger distortion probabilities would allow to gather more records into the same cluster even if they do not refer to the same entity. Instead, with low values of $\alpha_{j'l}$ we force all the clusters to have a reduced within-cluster variability and a greater between-cluster separation.

- Jhondrow et al. (BKA 2018): entity resolution via micro-clusters identification requires that the measurements errors go to zero as the number of entities increases, i.e. infeasibility of cluster based approaches for high dimensional record linkage problems without introducing further information



Posterior for K, N, $FNR = \frac{\sum_{j_1<j_2}(1-\Delta_{j_1j_2})\Delta^{true}_{j_1j_2}}{\sum_{j_1<j_2}\Delta^{true}_{j_1j_2}}$ $FDR = \frac{\sum_{j_1<j_2}\Delta_{j_1j_2}(1-\Delta^{true}_{j_1j_2})}{\sum_{j_1<j_2}\Delta_{j_1j_2}}$.

RL10000 data set



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.

Syrian data

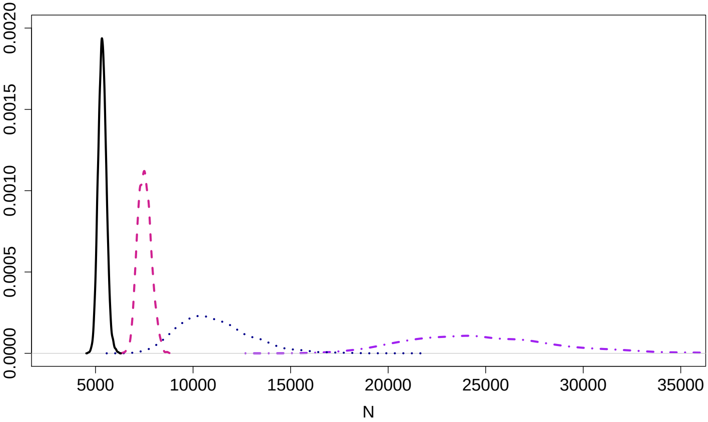
- We consider only victims in the province of Raqqa
- The VDC data set provides directly the English equivalents of the Arabic names. For the CSR list, the English equivalents have been obtained by software transliteration (additional noise!)
- After a preliminary data cleaning the resulting VDC data 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.

Three different analyses

- Separated lists. We fit our model to the single lists one by one (only within list duplications allowed)
- Joined lists. We fit the model jointly to both the lists.
- Record linkage (only beween lists duplications allowed)

		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

Distribution of the cluster sizes averaged across MCMC simulations.



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 line) and VDC (dotted line) single list analyses

Discussion

- General comments:
  - Priors used in the Bayesian nonparametric inference (species sampling) may be used in RL and duplications problems: (species $\leftrightarrow$ entity).
  - The hit-miss model provides computationally tractable marginal distributions for the cluster observations
- Specific comments on the population size estimation
  - Different sampling designs lead to different distributions $p(Z|N)$, we investigated only the simplest situation

Some references

- B. Liseo & A. Tancredi (2011). Bayesian estimation of population size via linkage of multivariate normal data sets. *Journal of Official Statistics*, 27 491–505.
- A. Tancredi & B. Liseo (2011). A hierarchical Bayesian approach to record linkage and population size estimation. *Annals of Applied Statistics* , 1553–1585.
- A. Tancredi, A., Liseo, B. (2015). Regression Analysis with linked data: Problems and possible solutions. Statistica, 75,119–35.
- R. C. Steorts. (2015) Entity resolution with empirically motivated priors. *Bayesian Analysis*, 10 879–875.
- R.C. Steorts, R. Hall & S. Fienberg (2016). A Bayesian approach to graphical record linkage and de-duplication. *Journal of the American Statistical Association* 111 1648–1659
- Johndrow, J. E., Lum, K., and Dunson, D. B. (2018). Theoretical limits of record linkage and microclustering. Biometrika, 431–446.
- A. Tancredi, R.C. Steorts, B. Liseo (2019). A unified framework for de-duplication and population size estimation *Bayesian Analysis*

THANK YOU!!!