Estimating population size from multisource data: a Bayesian perspective on latent class modelling.

Davide Di Cecco
Istat, Roma, Italy dicecco@istat.it

Marco Di Zio
Istat, Roma, Italy dizio@istat.it

Brunero Liseo
Sapienza University, Roma, Italy brunero.liseo@uniroma1.it

Abstract

In this work we aim at estimating the size of a population of interest based on a set of administrative sources to be used as capturing lists in a capture-recapture setting, that is, to estimate the number of units not captured by any list. In using administrative data in this context, the most relevant problems we encountered are the non-independence of the captures of the same unit in different lists/sources, and the presence of out-of-scope units in the lists. In order to deal with these problems, we propose a Bayesian latent class model.

Keywords: Capture–Recapture; Administrative Data; Latent Variable; Bayesian Analysis.

1. Introduction

In recent years, National Statistics Institutes are more and more investing in the possibility of producing statistics based solely on administrative data. In this work our goal is to estimate the size of a target population of interest, based on a set of administrative sources to be used as capturing lists in a capture–recapture setting.
In using administrative data in this context, the most relevant problems one encounters are:

- The non independence of the captures of the same unit in different lists/sources;
- The presence of out–of–scope units in the lists ("false captures").

As far as the first issue is concerned, the use of loglinear models is typical in situations where each list has a different capture probability and captures of the same unit in different lists are not independent. For the latter, we consider a situation where none of the sources can be considered as error free, so that we cannot utilize a supervised model to estimate the probability of a unit of being out–of–scope. For this reason, we deal with the problem of false captures in an unsupervised way, by including in our loglinear models a latent binary variable X identifying the units belonging to our target population (X=1), and the out–of–scope units (X=0). We are only interested in the total population count conditionally on X=1.
This setting has been studied in [Di Cecco et al. 2016] in a non–Bayesian approach. The choice of developing a Bayesian approach to that problem is essentially due to two reasons: firstly, as we will see in the next sections, if we restrict our loglinear models to be decomposable, a Bayesian approach is straightforward. Secondly, in the previous work we experienced a great sensitivity of the estimates to model selection, and in a Bayesian approach we have several ways to overcome this difficulty. For a start, we can easily obtain an interval estimation of the population size, whereas this task would require a complex use of Bootstrap sampling in a frequentist approach. Furthermore, (although not reported in the present paper), we could adopt a more thorough technique in which the uncertainty about the correct model is explicitly included, that is, we can consider model averaging where a prior is placed over a set of possible models (see [Madigan and York 97] for an application in capture–recapture).

This paper is structured as follows: in Section 2 the model is presented. In Section 3 we focus on decomposable loglinear models to introduce the prior distributions. In Section 4 the MCMC algorithm to sample from the posterior distribution of N, the total population size parameter, conditionally on X =1 is detailed. In Section 5 we present some simulation results.

2. The model

Let us introduce some notation. Consider k lists (or capture occasions). Let Y_i be the random variable indicating whether a unit has been captured by the i–th source (i.e., is listed on the i–th list):

$$Y_i = \begin{cases} 1 & \text{if a unit is captured in the } i\text{–th list;} \\ 0 & \text{otherwise.} \end{cases}$$

Let $\underline{\mathbf{Y}} = (Y_1, \ldots, Y_k)$, be the r.v. representing the capture history of a unit, and let $\{P(\underline{\mathbf{Y}} = \underline{\mathbf{y}}) = p_{\underline{\mathbf{y}}}\}_{\underline{\mathbf{y}} \in \{0,1\}^k}$ be the relative distribution. Let X be the latent variable identifying the units belonging to our target population:

$$X = \begin{cases} 1 & \text{if a unit belongs to the target population;} \\ 0 & \text{otherwise.} \end{cases}$$

We observed n_{obs} units, and $\underline{\mathbf{y}}_{obs}$ is the set $\{\underline{\mathbf{y}}_j\}_{j=1,\ldots,n_{obs}}$ of all the capture histories observed on the n_{obs} units. $n_{\underline{\mathbf{y}}}$ is the number of units having capture history $\underline{\mathbf{y}}$, $n_{x,\underline{\mathbf{y}}}$ of which belong to latent class x, such that $\sum_{x\in 0,1} n_{x,\underline{\mathbf{y}}} = n_{\underline{\mathbf{y}}}$ and $\sum_{\underline{\mathbf{y}}} n_{\underline{\mathbf{y}}} = n_{obs}$.
The units having capture history $\underline{\mathbf{y}} = \underline{\mathbf{0}} = (0, \ldots, 0)$ are unobserved and the relative count $n_{\underline{\mathbf{0}}}$ has to be estimated. The total population size, denoted as N, is consequently equal to $n_{obs} + n_{\underline{\mathbf{0}}}$.
We are interested in estimating the dimension of our target population, let us denote it as N_1, i.e., we want to estimate the number of units (both captured and uncaptured) for which $X = 1$. In fact, as a result, we will derive the posterior distribution of N_1, $\pi(N_1 \mid \underline{\mathbf{y}}_{obs})$.
We have $N_1 = \sum_{\underline{\mathbf{y}}} n_{1,\underline{\mathbf{y}}}$. That is, N_1 equals the sum of the number of in–scope units for each capture history, including the units not captured in any list. Let $N_0 = \sum_{\underline{\mathbf{y}}} n_{0,\underline{\mathbf{y}}}$ represent the overcoverage, so that $N_0 + N_1 = N$.
The likelihood function is:

$$L(p_{\underline{\mathbf{y}}}\,;\, n_{\underline{\mathbf{y}}}) \propto \prod_{\underline{\mathbf{y}}} p_{\underline{\mathbf{y}}}^{n_{\underline{\mathbf{y}}}} = \prod_{\underline{\mathbf{y}}} \left(\sum_x p_x p_{\underline{\mathbf{y}}|x} \right)^{n_{\underline{\mathbf{y}}}}, \tag{1}$$

where $p_{\underline{\mathbf{y}}}$ is the probability of observing capture history $\underline{\mathbf{y}}$ in a unit, and $p_{\underline{\mathbf{y}}|x}$ is the same probability conditional on $X = x$.
To illustrate our examples throughout the paper, we will refer to the decomposable model with four sources $\underline{\mathbf{Y}} = (A, B, C, D)$ of Figure 1 for which $p_{\underline{\mathbf{y}}} = p_{abcd} = \sum_x p_x\, p_{ab|x}\, p_{cd|x}$.

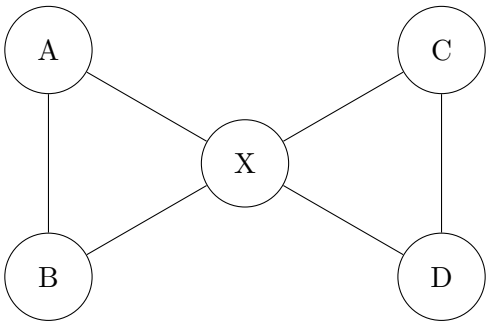

Figure 1: A simple decomposable graph with 5 nodes (4 sources and a latent variable).

3. Prior distributions

A model is said **graphical** if its dependency graph G exhaustively defines the structure of its joint distribution, i.e., if the joint distribution has no constraint other than those defined by G. Note that, in a graphical loglinear model, this is equivalent to say that the model includes all and only the interaction terms represented by the cliques of G, and it will be denoted as $\pi(G)$. A model is said **decomposable** iff its dependency graph G is chordal, i.e., if any cycle of G of four or more nodes has a chord.
Consider a decomposable graphical model and let $\{\mathcal{C}_1, \ldots, \mathcal{C}_g\}$ be the maximal cliques of the associated graph G, ($\bigcup_{i=1}^{g} \mathcal{C}_i = G$). Then, we have an ordering $\{\mathcal{C}_{\sigma(1)}, \ldots, \mathcal{C}_{\sigma(g)}\}$ of the set, such that, having defined the set of separators $(\mathcal{S}_2, \ldots, \mathcal{S}_g)$ as

$$\mathcal{S}_i = \mathcal{C}_{\sigma(i)} \cap \bigcup_{j=1}^{i-1} \mathcal{C}_{\sigma(j)} \qquad i = 2, \ldots, g,$$

we have that

- each separator $\mathcal{S}_i$ is a clique of G;
- each $\mathcal{S}_i$ is contained in one of the maximal cliques $\mathcal{C}_{\sigma(j)}$, $\sigma(j) < i$.

As a consequence, the joint distribution can be written in the following form:

$$p_G = \frac{\prod_{i=1}^{g} p_{\mathcal{C}_i}}{\prod_{j=2}^{g} p_{\mathcal{S}_j}}, \tag{2}$$

where p over a (sub)graph is to be intended as the (marginal) distribution over the variables included in the (sub)graph.
The joint probability can also be described as a product of conditional distributions in several ways (we will use the term "conditional representation" of the model). For example, denoting the conditional distribution $\frac{p_{\mathcal{C}_i}}{p_{\mathcal{S}_i}}$ as $p_{\mathcal{C}_i|\mathcal{S}_i}$ (with a slight abuse of notation), we can write the joint distribution as

$$p_{\mathcal{C}_1} \prod_{i=2}^{g} p_{\mathcal{C}_i|\mathcal{S}_i}.$$

In our models the latent variable X interacts with all other variables, so X is included in all maximal cliques and in all separators $\mathcal{S}_i$. Then, $p_{\mathcal{C}_1}$ can be written as $p_X p_{\mathcal{C}_1|X}$. As a consequence, we can always write a conditional representation of p_G as follows:

$$p_X p_{\mathcal{C}_1|X} \prod_{i=2}^{g} p_{\mathcal{C}_i|\mathcal{S}_i} \tag{3}$$

Madigan and York (1997) present a Bayesian approach to a class of capture–recapture models substantially equivalent to (2), but with no latent variable X. They propose the use of a class of prior distribution known as Hyper–Dirichlet which is conjugate to model (2) and has the property of being closed under marginalization (see also Dawid and Lauritzen, 1993). We preferred the conditional representation (3) as it is more natural in this context where the probability conditioned on X have a clear meaning in terms of true/false captures and under/overcoverage.
We denote as Θ the parameters of distribution (3):

$$\Theta = \left(P_X, P_{\mathcal{C}_1|X}, P_{\mathcal{C}_2|\mathcal{S}_2}, \ldots, P_{\mathcal{C}_g|\mathcal{S}_g}\right).$$

We define a prior distribution on the parameters Θ as follows: for each $P_{\mathcal{C}_i|\mathcal{S}_i}$ and for each value of $\mathcal{S}_i$ we set a Dirichlet distribution with no further restriction. These Dirichlet distributions are independent of each other, and it is not hard to see that this class of priors is conjugate our model. In fact, for each distribution

$P_{\mathcal{C}_i|\mathcal{S}_i}$ the corresponding Dirichlet is defined by the parameters $\alpha_{\underline{\mathbf{y}}_{\mathcal{C}}}$ defined for each possible combination of values $\underline{\mathbf{y}}_{\mathcal{C}} \in \{0,1\}^{|\mathcal{C}|}$ of the variables in $\mathcal{C}$. Hence, we have a prior

$$\pi(\Theta) \propto \prod_{i=1}^{g} \prod_{\underline{\mathbf{y}}} P_{\mathcal{C}_i|\mathcal{S}_i}^{\alpha_{\underline{\mathbf{y}}_{\mathcal{C}_i}}-1}$$

which is conjugate to model (3).
For example, in the model of Figure 1, we have $\mathcal{C}_1 = \{A, B, X\}$, $\mathcal{C}_2 = \{C, D, X\}$, $\mathcal{S}_2 = \{X\}$, and $p_G = p_x\, p_{\mathcal{C}_1|x}\, p_{\mathcal{C}_2|\mathcal{S}_2}$. We have $\Theta = (P_X, P_{AB|X}, P_{CD|X})$, and the following priors:

$$P_X \sim Beta(\alpha_0^X, \alpha_1^X) \qquad P_{AB|X} \sim Dir(\alpha_{ab|x}^{AB|X}), (x = 0, 1) \qquad P_{CD|X} \sim Dir(\alpha_{cd|x}^{CD|X}), (x = 0, 1)$$

$$\pi(\Theta) \propto \prod_{xabcd} P_x^{\alpha_x - 1} P_{ab|x}^{\alpha_{ab|x}-1} P_{cd|x}^{\alpha_{cd|x}-1}$$

It remains to set a prior distribution over $n_{\underline{\mathbf{0}}}$, or, equivalently, over N. In accordance with the literature on Bayesian capture–recapture, we will consider the following options:

- Jeffreys' prior, which in this case is: $\pi(N) \propto 1/N$;
- Poisson distribution, eventually together with a hyper prior over its parameter: $N \sim Poi(\lambda), \quad \lambda \sim Gamma(\alpha, \beta)$;
- Rissanen's distribution (Rissanen (1983)) which is always proper and is given by $\pi(N) \propto 2\log^*(N)$, where $\log^*(N)$ is the sum of the positive terms in the sequence $\{\log_2(N), \log_2(\log_2(N)), \ldots\}$.

We further hypothesize that the prior distributions over N and Θ are independent of each other.

4. MCMC algorithm

In this section we detail the steps of a Gibbs–based MCMC algorithm to sample from the posterior distribution of N_1.
At iteration $(t+1)$

1. sample all parameter $\Theta^{(t+1)}$ from their posterior conditional distributions which are Dirichlet distributions;

2. sample $n_{x,\underline{\mathbf{y}}}^{(t+1)}$ from $\pi(N_{x,\underline{\mathbf{y}}} \mid \Theta^{(t+1)}, \underline{\mathbf{y}}_{obs})$ for all $\underline{\mathbf{y}} \neq \underline{\mathbf{0}}$,
   $N_{x,\underline{\mathbf{y}}}^{(t+1)} \sim Bin\left(n_{\underline{\mathbf{y}}}, p_{x|\underline{\mathbf{y}}}^{(t+1)}\right) \quad$ where $\quad p_{x|\underline{\mathbf{y}}} = \dfrac{p_{x,\underline{\mathbf{y}}}}{\sum_x p_{x,\underline{\mathbf{y}}}}$;

3. sample from $\pi(N, N_{x,\underline{\mathbf{0}}} \mid \Theta, \underline{\mathbf{y}}_{obs}) = \pi(N \mid \Theta, \underline{\mathbf{y}}_{obs})\, \pi(N_{x,\underline{\mathbf{0}}} \mid N, \Theta, \underline{\mathbf{y}}_{obs})$
   To do this, first we notice that

$$\begin{aligned} \pi(N \mid \Theta, \underline{\mathbf{y}}_{obs}) = \pi(N \mid \Theta, n_{obs}) &= \frac{\pi(N)}{\pi(n_{obs} \mid \Theta)} \pi(n_{obs} \mid N, \Theta) \\ &\propto \pi(N) \binom{N}{n_{obs}} p_{\underline{\mathbf{0}}}^{N-n_{obs}} (1 - p_{\underline{\mathbf{0}}})^{n_{obs}} \end{aligned} \tag{4}$$

   where $p_{\underline{\mathbf{0}}} = \sum_x p_{x,\underline{\mathbf{0}}}$,
   and

$$\pi(N_{x,\underline{\mathbf{0}}} \mid N, \Theta, \underline{\mathbf{y}}_{obs}) \sim Bin\left(N - n_{obs}, p_{x|\underline{\mathbf{0}}}\right). \tag{5}$$

   Then, if we choose $\pi(N) \propto 1/N$, we simply have to

   - sample $N^{(t+1)}$ from $NegBin\left(n_{obs}, 1 - p_{\underline{\mathbf{0}}}^{(t+1)}\right)$;

- sample $n_{x,\underline{\mathbf{0}}}^{(t+1)}$ from $Bin\left((N^{(t+1)} - n_{obs}), p_{x|\underline{\mathbf{0}}}^{(t+1)}\right)$.

If we refer to the example of Figure 1, the first step consists of the following three steps:

- sample $p_x^{(t+1)}$ from $\pi\left(P_X \mid N_{x,\underline{\mathbf{y}}}, N_{x,\underline{\mathbf{0}}}, N, \underline{\mathbf{y}}_{obs}\right)$,
  $P_X^{(t+1)} \sim Beta\left(n_x^{(t)} + \alpha_x\right)$ where $n_x^{(t)} = \sum_{\underline{\mathbf{y}} \neq \underline{\mathbf{0}}} n_{x,\underline{\mathbf{y}}} + n_{x,\underline{\mathbf{0}}}$;
- sample $p_{ab|x}^{(t+1)}$ from $\pi\left(P_{AB|X} \mid N_{x,\underline{\mathbf{y}}}, N_{x,\underline{\mathbf{0}}}, N, \underline{\mathbf{y}}_{obs}\right)$,
  $P_{AB|X}^{(t+1)} \sim Dir\left(n_{abx}^{(t)} + \alpha_{ab|x}\right)$ where $n_{abx}^{(t)} = \sum_{\underline{\mathbf{y}}:(A=a,B=b)} n_{x,\underline{\mathbf{y}}}$;
- sample $p_{cd|x}^{(t+1)}$ from $\pi\left(P_{CD|X} \mid N_{x\mathbf{y}}, N_{x\mathbf{0}}, N, \underline{\mathbf{y}}_{obs}\right)$,
  $P_{CD|X}^{(t+1)} \sim Dir\left(n_{cdx}^{(t)} + \alpha_{cd|x}\right)$ where $n_{cdx}^{(t)} = \sum_{\underline{\mathbf{y}}:(C=c,D=d)} n_{x,\underline{\mathbf{y}}}$;

and, in the second step, we can explicitly write $p_{x|\underline{\mathbf{y}}}$ as

$$\frac{p_x^{(t+1)}\, p_{ab|x}^{(t+1)}\, p_{cd|x}^{(t+1)}}{\sum_x p_x^{(t+1)}\, p_{ab|x}^{(t+1)}\, p_{cd|x}^{(t+1)}}.$$

The above algorithm becomes slightly more involved if a Poisson or Rissanen prior instead of $1/N$ are adopted for N. In particular, we include in step 3 above a Metropolis-Hastings step (Metropolis-Hastings-within-Gibbs algorithm), to sample a value $N^{(t+1)}$ from $\pi(N \mid \Theta, \underline{\mathbf{y}}_{obs})$. In practice, this step presented no particular difficulty, as different choices for the proposal distribution (Gaussian, uniform,...) appear to lead to negligible differences in the results.

5. Simulation

In this section we report the results of a simulation to empirically assess the estimation algorithm.
Once again, we considered the example of Figure 1. We considered two scenarios: in the first one (reported in Figure 2), the model generating the data, characterised by the interaction cliques $[AX], [BX], [CDX]$, coincides with the estimating model. In the second scenario (reported in Figure 3) we test the robustness of the model to misspecification by generating the data from a model $[ABX], [CDX]$ and estimating it with the model $[AX], [BX], [CDX]$. For each scenario we generated two datasets with two different population size: one with $N = 500$, and one with $N = 1,000,000$. The parameters of the generating model are such that 40% of the total population size are false captures, and the rate of unobserved units (both in–scope and out–of–scope) is 23%. In the generating model of Scenario 2, A and B have a correlation of about 0.6 both under $X = 1$ and $X = 0$, while they are conditionally independent in the estimating model. We set non informative priors for all Dirichlet (all parameters α equal to 1), and test various prior distributions for N. We just reported the results for the informative Poisson distribution having parameter λ equal to the true value of N, and the non–informative prior $1/N$. All other priors give results in between these two.

References

Dawid, A. P., Lauritzen, S. L., (1993), Hyper Markov laws in the statistical analysis of decomposable graphical models, *The Annals of Statistics, 21:3, 1272–1317.*
Di Cecco D., Di Zio M., Filipponi D., Rocchetti I. (2016). Estimating Population Size from Multisource Data with Coverage and Unit Errors, *Proceedings ICES V* 20–23 June 2016, Geneva.
Madigan, D., York, J.C., (1997), Bayesian methods for estimation of the size of a closed population, *Biometrika, 84, 19–31.*
Rissanen, J. (1983), A universal prior for integers and estimation by minimum description length, *Ann. Statist. 11:2, 416–431* .

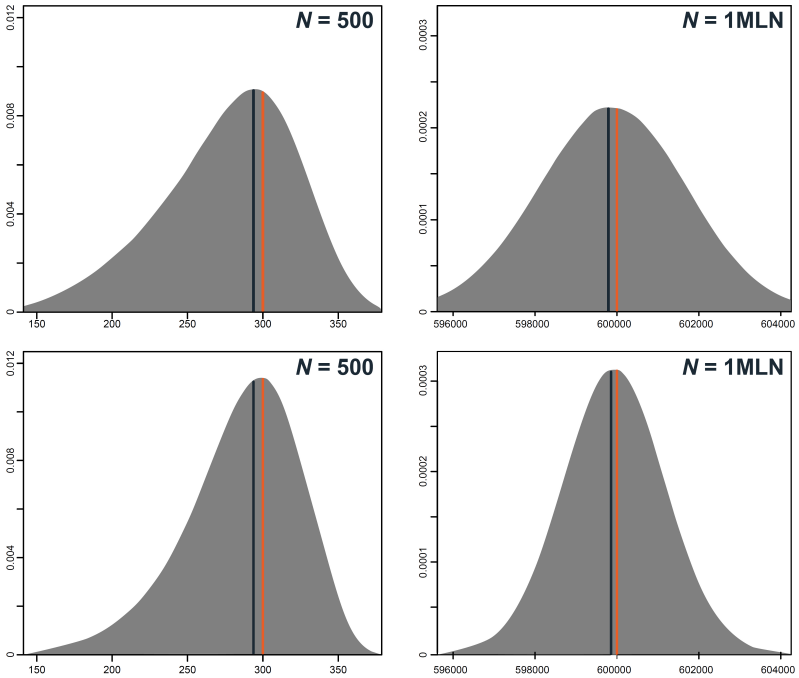

Figure 2: Posterior distribution of N_1 in Scenario 1. Results for Jeffreys'prior (top graphs) and Poisson prior with $\lambda = N$ (bottom graphs). The orange line indicates the true value of N_1; the blue line the Maximum Likelihood Estimate of N_1.

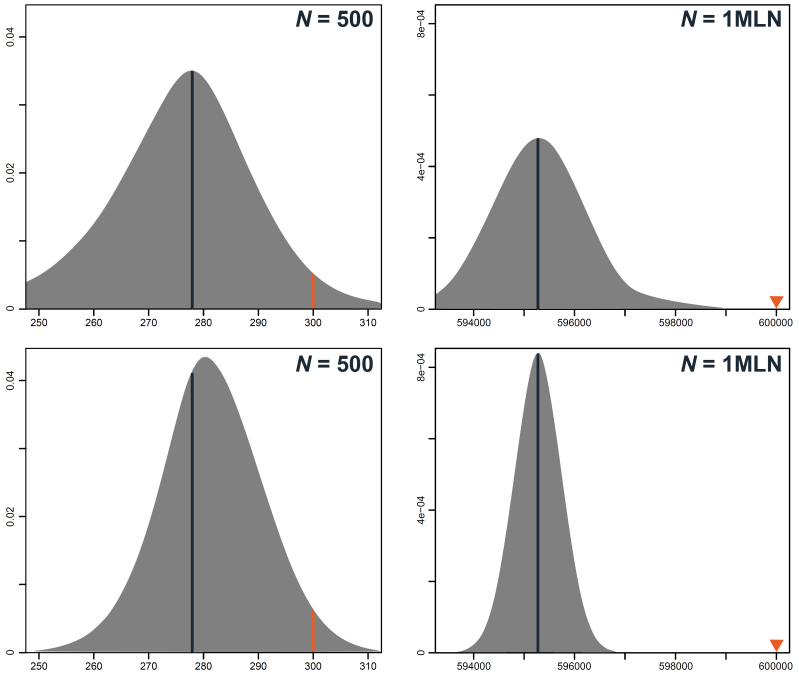

Figure 3: Posterior distribution of N_1 in Scenario 2. Results for Jeffreys'prior (top graphs) and Poisson prior with $\lambda = N$ (bottom graphs). The orange line indicates the true value of N_1; the blue line the Maximum Likelihood Estimate of N_1.